The DELOS Testbed for Choosing a Digital Preservation Strategy

Stephan Strodl[1], Andreas Rauber[1], Carl Rauch[1],
Hans Hofman[2], Franca Debole[3] and Giuseppe Amato[3]

[1] Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at
[2] Nationaal Archief, Den Haag, The Netherlands
www.nationaalarchief.nl
[3] Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy
www.isti.cnr.it

Abstract. With the rapid technological changes, digital preservation, i.e. the endeavor to provide long-term access to digital objects, is turning into one of the most pressing challenges to ensure the survival of our digital artefacts. A set of strategies has been proposed, with a range of tools supporting parts of digital preservation actions. Yet, with requirements on which strategy to follow and which tools to employ being different for each setting, depending e.g. on object characteristics or institutional requirements, deciding which solution to implement has turned into a crucial decision. This paper presents the DELOS Digital Preservation Testbed. 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. It is based on Utility Analysis to evaluate the performance of various solutions against well-defined objectives, and facilitates repeatable experiments in a standardized laboratory setting.

1 Introduction

Digital Preservation (DP) is turning into one of the most pressing challenges, for any setting handling and relying on digital objects, be it e-commerce, e-government, or private photo collections, requiring immediate action on an international level. With the rapid change in technology, both hardware and software, current objects will turn into uninterpretable bit-streams in relatively short periods of time, when the original environment to interpret them correctly becomes unavailable. Research in DP tries to mitigate this risk by devising a set of preservation strategies in order to ensure long-term access to digital objects. A number of strategies habe been devised over the last years, the most prominent ones being (1) migration, i.e. the repeated conversion of files into different, more current or more easily preservable, file formats (such as, e.g. to the recently adopted PDF/A standard [4], implementing a subset of PDF optimized for long-term preservation); or (2) the emulation of either a certain hardware infrastructure, operating system, or software functionality. All of the proposed

strategies have their advantages and disadvantages, and may be suitable in different settings [13]. When implementing a digital preservation strategy, the choice of the most suitable preservation solution is the most difficult part. The decision which strategy to follow and which tools and system to use is usually taken by groups of experts in the individual institution, who select the solution that best seems to satisfy their requirements. While with the profound expertise of the record managers these decisions are usually correct, it is hard to document them, to be able to later on re-establish the reasons why a certain tool was preferred over another, and why a certain parameter setting was chosen. With less expertise, or imprecise definitions of the requirements of different user groups, even the selection of a certain strategy may cause considerable difficulties. To be able to make profound, accountable decisions an evaluation process is needed, which allows a structured and documented evaluation of available DP solutions against well-defined requirements.

The DELOS DP Testbed presented in this paper allows the selection of the most suitable preservation solution for individual requirements. It enforces the explicit definition of preservation requirements, supports the appropriate documentation and evaluation by assisting in the process of running preservation experiments. This provides a means to perform structured and repeatable evaluations of various solutions, tools and systems for a given challenge, providing a means to make informed and accountable decisions on which solution to adopt.

In this paper we describe the workflow for evaluating and selecting DP solutions following the principles of the DELOS DP Testbed. We present a tool to support the automatic acquisition and documentation of the various requirements. Additionally, it provides a guidance for institutions having less expertise in the subtleties of DP challenges to identify core requirements that any solution should fulfill in a given setting. A set of initial case studies demonstrates the feasibility of the proposed approach.

The remainder of this paper is organized as follows: Section 2 provides some pointers to related initiatives. Following an overview of the principles of the DELOS DP Testbed in Section 3, a detailed description of the workflow is presented in Section 4. We report on a set of initial case studies in Section 5, followed by conclusions, lessons learned as well as an outlook on future work in Section 6.

2 Related work

The increasing amount of cultural and scientific information in digital form and the heterogeneity and complexity of the digital formats make it difficult to keep the heritage accessible and usable. While libraries, archives and cultural institutions may be the primary stakeholders, other institutions such as government agencies and increasingly also large industries as well as SME's and private persons, who have increasing amounts of legally or personally important data, are facing this challenge. Thus, a number of large scale initiatives are created, that integrate digital preservation capabilities into digital repository systems [11].

During the last couple of years, a lot of effort was spent to define, improve and evaluate preservation strategies. A good overview of preservation strategies is provided by the companion document to the UNESCO charter for the preservation of the digital heritage [13]. Research on technical preservation issues is focused on two dominant strategies, namely Migration and Emulation. Scientific results on Migration, which is at the current time the most common preservation strategy, were published for example by the Council of Library and Information Resources (CLIR) [6], where different kinds of risks for a migration project are presented. Migration requires the repeated conversion of a digital object into more stable or current file formats.

Work on the second important preservation strategy, Emulation, was advocated by Jeff Rothenberg [10], envisioning a framework of an ideal preservation surrounding. In order to make Emulation usable in practice, several projects developed it further. One of them is the CAMILEON project [2], trying to implement first solutions and to compare Emulation to Migration. More recently, the Universal Virtual Computer (UVC) has been proposed as a promising solution [3]. Emulation aims at providing programs that mimic a certain environment, e.g. emulating a certain processor or the features of a certain operating system, allowing users, for example, to run Microsoft WORD on a Linux operating system using the WINE emulator.

Similar to the Utility Analysis based approach for identifying and documenting the objectives for a preservation endeavor [9], the Arts and Humanities Data Service (AHDS) and University of London Computer Centre started the DAAT Project (Digital Asset Assessment Tool) [12]. The aim is to develop a tool to identify the preservation needs of various digital holdings.

The approach presented in this paper basically focuses on the elicitation and documentation of the requirements (objectives), as well as running and evaluating experiments in a structured way. In order to automate the evaluation, a number of tools like JHove [1] may be employed to analyze the resulting files after applying a preservation action. PANIC [5] addresses the challenges of integrating and leveraging existing tools and services and assisting organizations to dynamically discover the optimum preservation strategy. File format repositories, such as PRONOM [8] may be used to identify specific characteristics of the digital objects at hand.

3 DELOS DP Testbed

During the last couple of years two frameworks were created for supporting the establishment of DP solutions, namely the Utility Analysis approach [9] and the Dutch testbed designed by the Dutch National Archive. The advantages of these two were integrated and form the basis for the DELOS Digital Preservation Testbed. The strengths of the Utility Analysis are the clear hierarchical structuring of the preservation objectives, which document the requirements and the goals for a optimal preservation solution. The strength of the Dutch testbed is the detailed definition of the environment and the experiment basis.

3.1 Testbed principles

Figure 1 provides an overview of the workflow of the DELOS DP Testbed. The 3-phase process, consisting of 14 steps, starts with defining the scenario, setting the boundaries, defining and describing the requirements, which are to be fulfilled by the possible alternatives. After the definition of the requirements the second part of the process is to identify and evaluate potential alternatives. Therefore, first 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 stableness of the final ranking is analyzed with respect to minor changes in the weighting and performance of the individual objectives using Sensitivity Analysis. The results are finally considered by taking non-measurable influences on the decision into account. After this consideration a clear and well argumented accountable recommendation for one of the alternatives can be made.

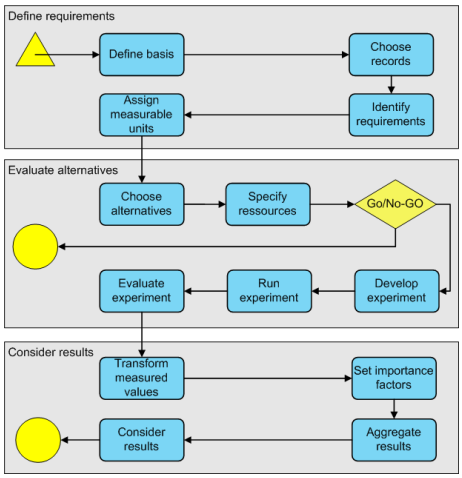

Fig. 1. Overview of DELOS Digital Preservation Testbed's workflow

To simplify the process, to guide users and to automate the structured documentation a software tool is introduced[4]. 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.

[4] `http://ifs.tuwien.ac.at/dp`

4 Testbed Workflow

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

Step 1 - Define basis

The testbed process starts with defining the basis. This is a semi-structured description including (i) the required types of records, which are considered (e.g. E-Mail correspondence or immigration records), (ii) a description of the environment, in which the testbed process takes place (e.g. governmental archive, a university library), and (iii) information on the amount of files or records, which are expected to be preserved with the chosen alternative.

Step 2 - Choose records

In order to be able to evaluate the DP solutions, sample records are needed to run the experiments. In this step sample records are chosen. Some file repositories are available or under construction, where well described files in many different formats and types can be downloaded, e.g. [7]. Alternatively, representative files from the collection to be preserved can be chosen with respect to the variety of document characteristics. The result of this stage is a sample of characteristic records, between 5 and 20 files, which are later used for evaluating the alternatives.

Step 3 - Define requirements

In order to decide which preservation solution is most suitable for a given setting, detailed requirements have to be specified in a structured and well documented way. Requirements definition is thus a decisive step and usually the most time-consuming. The goal of this step is to clearly define the requirements and goals, which should be fulfilled by the preservation solution. In the so-called objective tree, different goals and requirements, high-level and detailed ones, are collected and organized in a tree structure. Generally, there are two ways to define the objectives, the bottom-up and the top-down approach. The former is to collect a list of basic attributes (such as character encoding, font color representation or hardware costs), and to aggregate them on a higher level (such as the preservation of the look and the accountability of costs). The opposite is done in the top-down approach where general aims (such as record characteristics or cost structure) are defined and gradually broken down into increasingly fine-granular objectives.

A synthesis of these two approaches is probably the best solution, combining high-level aims with basic requirements. While the resulting objective trees usually differ from preservation setting to preservation setting, some general principles can be observed. At top level, the objectives usually can be organized into four main categories, namely:

- *File characteristics*: In this part of the tree all objectives are mentioned, that describe the visual and contextual experience, a user has by dealing with a digital record. Subdivisions may be “Appearance”, “Content”, “Structure” and “Behavior”, with lowest level objectives being e.g. color depth, image resolution, forms of interactivity, macro support, embedded metadata.

– *Record characteristics*: Here the technical foundations of a digital record are described, the context, the storage medium, interrelationships and metadata.
– *Process characteristics*: The third group are those objectives that describe the preservation process. These include usability, complexity or scalability.
– *Costs*: The last group of objectives, which have a significant influence on the choice of a preservation solution, are costs. Usually, costs 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 elicitation of requirements. These trees document the individual preservation requirements of institution and for a given partially homogeneous collection of objects, for example, scientific papers and dissertations in PDF format; historic audio recordings; video holdings from ethnographic studies. Typical trees may contain between 50 to several hundred objectives, organized in usually 4-6 levels of hierarchy.

Step 4 - Assign measurable units

In order to be able to objectively measure and compare the performance of the various preservation solutions with the set of requirements, units of measurement need to be defined for each objective, i.e. leaf of the tree. Wherever possible, these objectives should be objectively (and, preferably, automatically) measurable (e.g. seconds per object, Euro per object, dots-per-inch resolution, bits of color depth). In some cases, (semi-)subjective scales will need to be employed (e.g. degrees of openness and stability, support of a standard, diffusion of a file format, number of access tools available for a specific object type).

Step 5 - Choose alternatives

In order to find the most suitable preservation solution, different alternatives need to be identified, which subsequently are to be evaluated in the DELOS DP Testbed. Alternatives can be from all different preservation strategies, such as specific emulators, tools to migrate digital objects from one format to another (version of the same or a different) format, put data into a computer museum, etc. Descriptions of these alternatives should be detailed enough to allow later re-evaluation of the analysis, thus describing the specific tools used, including their release version, which operating system they are being run on, and which parameter settings are being used. An example is "Migration from MS Word to PDF" using Acrobat 7.0 Distiller running on WIN XP (SP2) with a documented list of the parameter settings.

Step 6 - Specify resources

In order to assess the resources that are need to run the evaluation, for each potential alternative the amount of work, time and money is estimated. The input for this step is the list of the alternatives. For every alternative a project and work description plan is developed, where the amount of work, time and money required for testing these alternative are estimated. The effort and workflow for building or installing such a process is determined, the knowledge of the required personnel set and the duration for the whole process estimated.

Step 7 - Go/No-Go decision

This stage considers the resources and requirements definition to determine if the proposed alternatives are feasible at all and whether one may proceed with the process as planned (Go), if revisions to the design or the strategy are needed before the process can go on (Provisional-Go), if the suggested strategy should be delayed for a specified period or until a specified event, such as the availability of additional research results, occurs (Deferred-Go), or if the strategy should not be considered any longer (No Go).

Step 8 - Develop experiments

In order to run repeatable tests a documented setting is necessary, which includes the workflow of the experiment, software and system of the experiment environment and the mechanism to capture the results. All of the items needed for the experiment will be developed and/or installed and tested, including copies of all the objects needed for the experiment, software packages and programs needed, and mechanisms for capturing the results and the evaluation.

Step 9 - Run experiments

An experiment will test one or more aspects of applying a specific preservation solution to the previously defined sample objects. Running an experiment will produce results, e.g. converted computer files, revised metadata, etc., that will be evaluated in the next step.

Step 10 - Evaluate experiments

The results of the experiment will be evaluated to determine how successfully the requirements were met. Therefore, the leaf objectives defined in the objective tree are evaluated with the defined unit of measurement.

Step 11 - Transform measured values

The measurements taken in the experiments all have different scales (such as time in seconds, costs in Euro, resolution in dots-per-inch). In order to make these comparable they are transformed to a uniform scale using transformation tables. The subjectively measured objectives on a uniform scale e.g. 0 to 5 can be used directly as comparable numbers. The objectively measured ones are transformed to the uniform scale, experience so far has shown that a performance scale of 1 to 5 is a reasonable approach. On several occasions the definition of a special performance level 0 (or "not acceptable", "n/a") turned out to be helpful. If the measures for a certain objective are below a certain threshold, this value will be assigned, serving as a drop-out criterion for that alternative no matter how well it performs in all other aspects. The threshold values cannot be generally defined, but have to be individually specified for every implementation. After applying the transformation functions we obtain a list of comparable values per alternative. These values form the input to the aggregation and final ranking.

Step 12 - Set importance factors

The objective tree consists of many objectives. Not all of them are equally important, and we may decide to accept different degrees of conformance of a solution in different objectives. Thus, importance factors, also referred to as weights, are assigned to each node in the tree to explicitly describe, which objectives play a major or minor role for the final decision. In a top-down manner, relative importance factors between 0 and 1 are assigned to all the children of

a given node. These weights depend largely on individual requirements. While there are different ways of assigning the weights, practice has shown that group decision processes result in stable evaluations of the relative importance of the various objectives. The weights of the single leaves can be obtained by multiplying their value times the weights of their parent nodes, summing up to one for the whole tree. The software implementation supports sets of weights from different users, which are further used for the Sensitive Analysis of the evaluation. For the normal evaluation of the alternatives an average value of the weights assigned by different users is used. The result of this stage is an objective tree with importance factors assigned to each objective, representing their relative relevance with respect to the overall goals.

Step 13 - Aggregate 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 as transformed by the transformation tables and multiplied with the weighting factor. These numbers are summarized to a single comparable number per alternative. We thus obtain aggregated performance values for each part of the objective tree for each alternative, including, of course, an overall performance value at the root level. A first ranking of the alternatives can be done based on the final values per alternative.

Step 14 - Perform Sensitivity Analysis

In the last step a ranking of the performance of the various alternatives is created based on the overall degrees of fulfillment of the objectives. 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, some Sensitivity Analysis may be automatically performed by analyzing, for example, the stableness 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. Additionally, some side effects can be consider, which are not included in the numerical evaluation. Such effects could be relationships with a supplier, expertise in a certain alternative, or individual assessment that one or the other solution might become the market leader within a couple of years. All of which, of course will need to be carefully documented if used to influence the final solution. The result of this analysis 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 standardized and repeatable experiment setting, it forms the basis for sound and accountable decisions on which solution to implement.

All the stages of the experiment will be considered to make recommendations for the refinement and enhancement of future experiments, to propose further experiments, and to provide input into the evaluation of the testbed.

5 Case Studies

To evaluate the potential of the presented approach, a set of case studies was performed with different partner institutions.

- Video Files of the Austrian Phonogrammarchiv
  The Austrian Phonogrammarchiv is re-considering its appraisal regulations for video files, specifically with respect to most suitable 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 revealed, that the preservation solutions differ in only few objectives, such as signal representation, color proofness and stereo quality.
- Document records of the Dutch National Archive
  The Dutch National Archive is responsible for storing all information, which is generated by the Dutch government. The case study tried to define the objectives for the preservation of different kinds of documents, such as video and audio document focusing particularly on the record characteristics. The resulting objective tree contains around 450 objectives.
- Migration of 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.

6 Conclusions

The proposed 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. It thus helps to establish and maintain a trusted preservation environment.

While many of the processing steps are automated, a significant amount of work is involved still in the evaluation of the results of applying a preservation action in order to acquire the measures for the various objectives. Integrating

tools for file analysis as well as adding further measurements during the experiment runs is needed in order to reduce this.

Furthermore, a significantly larger series of case studies will need to be performed in order to establish a solid basis of best practice models for different institutions and different types of digital objects. this may later-on even lead to a kind of recommender process, where – upon specifying e.g. the type of institution and the type of objects concerned – a pre-defined objective tree, or at least a set of building blocks, is proposed by the system.

Acknowledgements

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

References

1. Harvard University Library. Jhove - jstor/harvard object validation environment. Website, 2005. `http://hul.harvard.edu/jhove`.
2. M. Hedstrom and C Lampe. Emulation vs. migration. do users care?,. *RLG DigiNews*, Dec. 2001, Vol. 5, No. 6, 2001.
3. J.R. Hoeven, R.J. Van Der Diessen, and K. Van En Meer. Development of a universal virtual computer (uvc) for long-term preservation of digital objects. *Journal of Information Science,*, Vol. 31 (3), 2005.
4. *ISO/CD 19005-1, Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A)*, 2004.
5. Sharmin Choudhury Jane Hunter. PANIC - An integrated approach to the preservation of composite digital objects using semantic web services. In *Proc. 5th Int. Web Archiving Workshop (IWAW05)*, Vienna, Austria, September 2005.
6. G.W. Lawrence, W.R. Kehoe, O.Y. Rieger, W.H. Walters, and A.R. Kenney. Risk management of digital information: A file format investigation,. CLIR, 2000.
7. Virginia Ogle and Robert Wilensky. Testbed development for the Berkeley Digital Library Project. *D-LIB Magazine*, July/August 1996. URL http://www.dlib.org.
8. Jo Pettitt. *PRONOM - Field Descriptions*. The National Archives, Digital Preservation Department, 2003. URL http://www.records.pro.gov.uk/-pronom.
9. Carl Rauch and Andreas Rauber. Preserving digital media: Towards a preservation solution evaluation metric. In *Proceedings of the 7th International Conference on Asian Digital Libraries, ICADL 2004*, pages 203–212. Springer, December 2004.
10. J Rothenberg. Avoiding technological quicksand: Finding a viable technical foundation for digital preservation. CLIR, 1999.
11. Mackenzie Smith. Eternal bits: How can we preserve digital files and save our collective memory? *IEEE Spectrum*, 42(7), July 2005.
12. ULCC. DAAT: Digital asset assessment tool. Website, 2004. `http://ahds.ac.uk/about/projects/daat/`.
13. UNESCO, Information Society Division. *Guidelines for the preservation of digital heritage*, October 2003. URL http://www.unesco.org/webworld/mdm.