Plato: A Service Oriented Decision Support System for Preservation Planning

Christoph Becker, Hannes Kulovits,
Andreas Rauber
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/dp

Hans Hofman
Nationaal Archief
The Hague, The Netherlands
www.nationaalarchief.nl

ABSTRACT

The fast changes of technologies in today's information landscape have considerably shortened the lifespan of digital objects. Digital preservation has become a pressing challenge. Different strategies such as migration and emulation have been proposed; however, the decision for a specific tool e.g. for format migration or an emulator is very complex. The process of evaluating potential solutions against specific requirements and building a plan for preserving a given set of objects is called preservation planning. So far, it is a mainly manual, sometimes ad-hoc process with little or no tool support. This paper presents a service-oriented architecture and decision support tool that implements a solid preservation planning process and integrates services for content characterisation, preservation action and automatic object comparison to provide maximum support for preservation planning endeavours.

Categories and Subject Descriptors

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

General Terms

Design, Experimentation, Measurement, Standardization

Keywords

Digital Preservation, Preservation Planning, decision support system, service oriented architecture

1. INTRODUCTION

Digital preservation as the effort to preserve digital objects for a given purpose over long periods of time has become a highly recognised matter during the last years. Its urgency has recently been reemphasised by the results of a survey among archiving professionals [11].

The two most often considered strategies for preservation actions today are migration and emulation. Migration operates on the objects that are at risk and transforms them to representations that are considered to be better suited for long-term archiving in a given context. Emulation operates on the environment of the objects, trying to simulate the original environment that the objects need.

A number of tools performing preservation actions are available; most often, there is no optimal solution for preserving a given set of objects. Complex requirements need to be considered during the process of deciding upon a solution to adopt. Careful documentation and a well-defined procedure are necessary to ensure that the outcome of a preservation planning activity meets the instution's needs.

Preservation planning aids in the decision making process by evaluating available solutions against clearly defined and measurable criteria and arriving at concrete plans for action based on these evaluation results. During this process, the planner needs to be informed about possible actions that are applicable to the objects at question. On the other hand, a preferably automatic verification and comparison of documents and objects before and after migration (or during emulation) is needed to support the judgement of quality of the applied preservation actions in terms of defined requirements. Moreover, the planning activity needs to be repeatable and well-documented to ensure traceability of both decisions and the reasons underlying them.

This paper presents the Planets preservation planning component (*Plato*) which implements the planning methodology [8] developed within the DELOS and Planets projects and integrates services for content characterisation, preservation action and object comparison to provide a service-oriented decision support system for preservation planning.

The remainder of this paper is structured as follows. The next section outlines related work in the area of preservation planning, content characterisation and distributed preservation services. Section 3 describes the planning tool *Plato*, providing a walkthrough of the underlying workflow and highlighting the integration of services. Section 4 draws conclusions and points out directions for future work.

2. RELATED WORK

Digital preservation has become a highly active research area in the last decade, as many memory institutions realised that their content will cease to be accessible within years[12]. At the heart of a preservation endeavour lies preservation planning, which is a core entity in the Reference Model for an Open Archival Information System, OAIS [4].

The PLANETS preservation planning methodology[8] defines measurable requirements for preservation strategies in a hierarchical form and evaluates them in a standardised setting to arrive at a recommendation for a solution. The procedure is independent of the solutions considered; it can be applied for any class of strategy, be it migration, emulation or different approaches, and has been validated in a series of case studies [1, 9]. An OAIS-based analysis of the approach is shown in [10].

An important aspect of the evaluation process is the need for automatic validation and comparison of objects. A number of tools and services have been developed that perform content characterisation specifically for digital preservation. The National Library of New Zealand Metadata Extraction Tool[1] extracts preservation metadata for various input file formats. Harvard University Library's tool JHove[2] enables the identification and characterisation of digital objects. Collection profiling services build upon characterisation tools and registries such as PRONOM[3] to create profiles of repository collections [3].

The eXtensible Characterisation Languages presented in [2] support the automatic validation of document conversions and the evaluation of migration quality through a decomposition of digital objects into their elements, thus representing them in an abstract XML language.

Several approaches deal with distributed preservation architectures. Hunter [7] describes a distributed architecture for preserving composite digital objects using ontologies and web services. Ferreira [6] presents a system for performing format migrations based on pre-specified requirements.

The EU project 'Preservation and Long-Term Access via Networked Services' (PLANETS)[4] is creating a distributed service-oriented architecture as well as practical services and tools for digital preservation [5]. Based on a common conceptual framework, it is developing services for preservation action, characterisation, testing and planning, of which the system presented in this paper forms part of.

3. PLATO: PRESERVATION PLANNING

3.1 The preservation planning workflow

The Planets preservation planning workflow as described in [8] consists of three main stages:

1. **Requirements definition** is the natural first step in the planning procedure, collecting requirements from the wide range of stakeholders and influence factors that have to be considered for a given institutional setting. This includes the involvement of curators and domain experts as well as IT administrators and consumers. Requirements are specified in a quantifiable way, starting at high-level objectives and breaking them down into measurable criteria, thus creating an *objective tree* which forms the basis of the evaluation of alternative strategies. Furthermore, as this evaluation would be infeasible on the potentially very large collection of objects, the planner selects representative sample objects that should cover the range of essential characteristics present in the collection at hand.

---

[1] http://meta-extractor.sourceforge.net/

[2] http://hul.harvard.edu/jhove

[3] http://www.nationalarchives.gov.uk/pronom
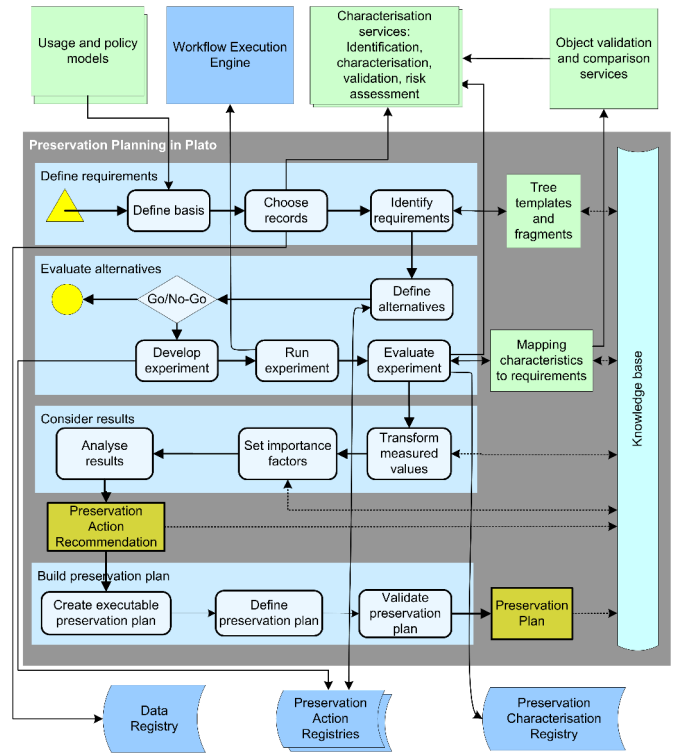
[4] http://www.planets-project.eu

Figure 1: Preservation planning environment

2. The **evaluation of potential strategies** is carried out empirically by applying selected tools to the defined sample content and evaluating the outcomes against the specified requirements.

3. **Analysis of the results** takes into account the different weighting of requirements and allows the planner to arrive at a well-informed recommendation for a solution to adopt.

3.2 The planning tool Plato

The planning tool presented here implements this three-stage workflow and includes additional external services to automate the process. It further extends it with a fourth phase in which an executable preservation plan is created, based on the well-documented recommendation. The software itself is a J2EE web application relying on open frameworks such as Java Server Faces and AJAX for the presentation layer and Enterprise Java Beans for the backend. It is integrated in an interoperability framework that guarantees loose coupling of services and registries through standard interfaces and provides common services such as user management, security, and a common workspace. Based on this technical foundation, the aim is to create an interactive and highly supportive software environment that advances the insight of preservation planners and enables proactive preservation planning.

Figure 1 illustrates the preservation planning environment, putting the described workflow in the working context of services and registries as they are currently being implemented. In principle, there are three aspects to consider: (1) Integrating registries for information discovery; (2) Integrating
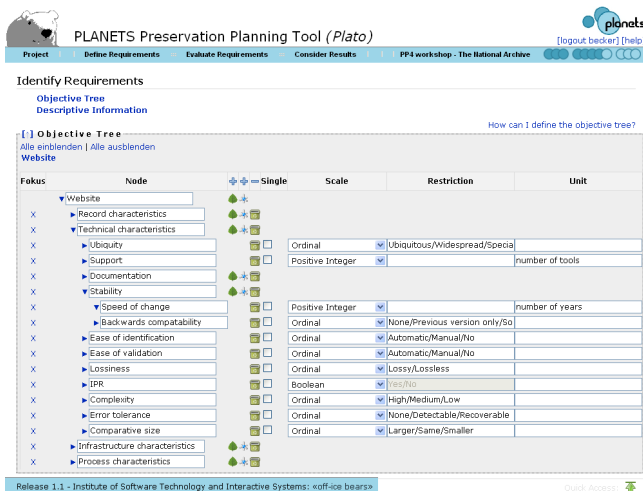
Figure 2: Requirements definition in Plato

Figure 3: Visualisation of results

services for preservation action and characterisation of objects; and (3) Proactively supporting the planning with a knowledge base that holds reusable patterns and templates for requirements recurring in different planning situations.

The right choice of samples that are representative for the collection under consideration is essential, as any skewed representation might lead to wrong results. **Collection profiling services** based on characterisation services and format registries inform the selection process and ensure the right stratification of samples. **Risk assessment services** further assist by quantifying both the inherent risks of object formats and the salient risks present in the objects which are of particular relevance to a specific file format, such as the number of pages for some document formats or the presence of transparency layers in images.

The specification of requirements in a tree structure is often done in a workshop setting. This is supported by both a flexible web interface as depicted in Figure 2 and a direct tree import from mind-mapping software[5]. The knowledge base provides recurring fragments and templates, such as process requirements for an archival institution or essential object characteristics for electronic documents in a library, to assist in the process of tree creation.

Service discovery is the prime issue during the next step of defining alternatives to consider for evaluation. Starting from the sample objects and their formats, the system queries available registries of preservation actions and looks up applicable tools such as emulators of the original environment or migration tools that can handle the provided input format. The Planets registry moreover holds information on benchmark evaluation results produced by experiments carried out in the Planets Testbed, which provides a controlled environment for preservation experiments[5].

Preservation action tools that are accessible through a web service are directly invoked during the execution of experiments on the sample objects; other tools such as emulators have to be executed externally.

The evaluation of experiments is probably the most complex and, so far, least automated step in preservation planning. Until now, most of the judgement, e.g. if a migration
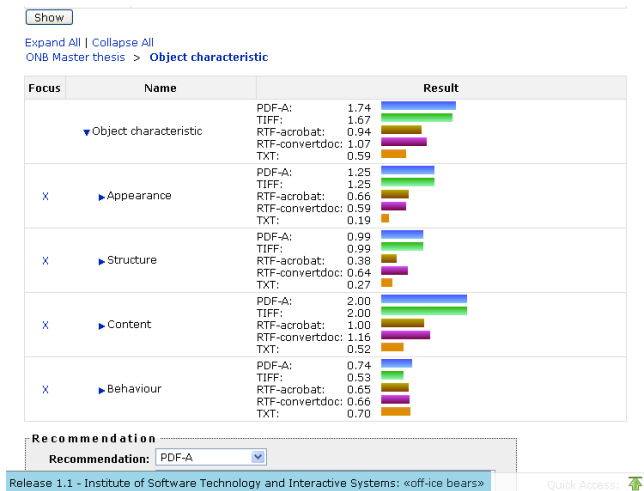
tool accurately preserves the colour model of an image or the line breaks in a document, has to be carried out manually by looking at the rendered objects. However, **characterisation services** are available that can measure some of the essential characteristics of objects such as the dimensions of images. In contrast to characterisation tools like JHove, the extensible characterisation languages (XCL) [2] do not attempt to extract a set of characteristics from a file, but instead are able to express the complete informational content of a file in a format independent model.

Comparison services specify measurable properties as well as property-specific metrics and their implementation as algorithms in order to identify degrees of equality between two objects. This is in principle independent of the applied strategy, i.e. migration or emulation. The compared objects can be both the original and a migrated object, or the original object in two different environments. To allow comparison and evaluation, a mapping is created between the requirements specified in the objective tree and the characteristics that can be measured and compared automatically by the available characterisation tools. This mapping partly stems from the knowledge base, but can be adapted by the user.

Both XCL and other characterisation tools such as JHove are integrated in the evaluation of experiments. This also includes risk assessment services which compare the risk scores of objects resulting from the application of preservation actions against the scores of the original samples.

The transformation of measured values to a uniform scale as needed for the aggregation of results and the importance weighting of requirements are supported by the knowledge base. Analysis of results is facilitated by a dynamic and flexible visualisation as depicted in Figure 3, where the planner can choose between different aggregation methods and dynamically configure the information content to analyse the strengths and weaknesses of the alternatives considered. Based on this analysis, a well-documented and solid recommendation for a solution can be made.

This recommendation forms the basis for **building a preservation plan** in the fourth stage. A preservation plan contains a description of the context and the decision taken, including the complete evidence base. This evidence base comprises a thorough description of the planning context

---

[5]http://freemind.sourceforge.net

and environment, ranging from the institution's mission statement via user group characteristics and policies to the collection at hand (documented in a collection profile). Moreover it contains the chosen sample objects, the requirements and additional documentation as well as considered solutions and the evaluation results. The plan furthermore contains cost indications and triggers for re-iterating the planning, and as a core part it entails a *preservation action plan*. If the applied strategy and its deployment support it, this is an executable workflow accessing distributed services.

During the fourth stage, the planner may select a subset of the criteria used for evaluating solutions to be applied automatically with each preservation action as a mechanism for quality assurance. The corresponding characterisation actions which are used for property extraction and validation are then included in the executable preservation plan.

The first version of Plato is publicly accessible[6]. It implements the workflow described in [8] and provides partial service integration such as file format identification. The next version will include a wider set of services for preservation actions and characterization, and result in a well-documented preservation plan. The final version will then include the creation of an executable preservation plan.

4. DISCUSSION AND OUTLOOK

Until now, preservation planning is largely a manual and tedious process where available solutions are evaluated against the specific requirements of a particular situation. This paper described the basic architecture and features of a decision support system for preservation planning based on a service oriented approach for distributed preservation solutions. The system implements a well-documented and validated preservation planning methodology and integrates registries and services for preservation action and characterisation. It furthermore provides a sophisticated web-based interface for guiding the planner through the process.

Preservation action services are discovered in registries and invoked through a BPEL-based workflow execution engine. The time-consuming and inherently subjective process of evaluating the results is being objectified and automated as far as possible by mapping identified requirements such as essential characteristics of objects to properties that can be automatically extracted and compared by characterisation tools. A knowledge base supports the preservation planner step by step in identifying requirements and mappings to characteristics as well as transformation of the results and importance weighting of the requirements.

Current and future work is aimed at the following aspects:

- Advanced collection profiling services that go beyond the currently available solutions and deliver detailed characteristics of the objects in a collection,

- Improving the automatic evaluation of preservation actions by integrating comparison services,

- Developing technology watch services that monitor the environment and proactively trigger a planning activity, and

- Integrating recommender systems in order to provide advanced decision support.

---

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

5. REFERENCES

[1] BECKER, C., KOLAR, G., KUENG, J., AND RAUBER, A. Preserving interactive multimedia art: A case study in preservation planning. In *Proc. Tenth Conf. on Asian Digital Libraries (ICADL'07)* (Hanoi, Vietnam, December 10-13 2007).

[2] BECKER, C., RAUBER, A., HEYDEGGER, V., SCHNASSE, J., AND THALLER, M. A generic XML language for characterising objects to support digital preservation. In *Proc. 23rd Annual ACM Symposium on Applied Computing (SAC'08)* (Fortaleza, Brazil, March 16-20 2008), vol. 1, ACM, pp. 402–406.

[3] BRODY, T., CARR, L., HEY, J. M., BROWN, A., AND HITCHCOCK, S. PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *Int. Journal of Digital Curation 2*, 2 (November 2007), 3–19.

[4] CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1, 2002.

[5] FARQUHAR, A., AND HOCKX-YU, H. Planets: Integrated services for digital preservation. *Int. Journal of Digital Curation 2*, 2 (November 2007), 88–99.

[6] FERREIRA, M., BAPTISTA, A. A., AND RAMALHO, J. C. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries 6*, 4 (July 2007), 295–304.

[7] HUNTER, J., AND CHOUDHURY, S. PANIC - an integrated approach to the preservation of complex digital objects using semantic web services. *Int. Journal on Digital Libraries: Special Issue on Complex Digital Objects 6*, 2 (April 2006), 174–183.

[8] STRODL, S., BECKER, C., NEUMAYER, R., AND RAUBER, A. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proc. 7th ACM IEEE Joint Conf. on Digital Libraries (JCDL'07)* (2007), pp. 29–38.

[9] STRODL, S., BECKER, C., NEUMAYER, R., RAUBER, A., BETTELLI, E. N., KAISER, M., HOFMAN, H., NEUROTH, H., STRATHMANN, S., DEBOLE, F., AND AMATO, G. Evaluating preservation strategies for electronic theses and dissertations. In *Revised Selected Papers of the 1st International DELOS Conf.* (Pisa, Italy, 2007), Springer, pp. 238–247.

[10] STRODL, S., AND RAUBER, A. Preservation planning in the OAIS model. In *Int. Conf. on Digital Preservation (IPRES'07)* (Beijing, China, 2007).

[11] THE 100 YEAR ARCHIVE TASK FORCE. The 100 year archive requirements survey. `http://www.snia.org/forums/dmf/programs/ltacsi/100_year/`, 2007.

[12] UNESCO. UNESCO charter on the preservation of digital heritage. Adopted at the 32nd session of the General Conference of UNESCO, October 17, 2003. `http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf`.