The SCAPE Planning and Watch suite

Supporting the preservation lifecycle in repositories

Michael Kraxner, Markus Plangg,
Kresimir Duretec, Christoph Becker
Vienna University of Technology
Vienna, Austria
{michael.kraxner, markus.plangg,
kresimir.duretec,
christoph.becker}@tuwien.ac.at

Luis Faria
KEEP Solutions
Braga, Portugal
lfaria@keep.pt

ABSTRACT

Increasingly, content owners are operating repositories with large, heterogeneous collections. The responsibility to provide access to these collections on the long term requires preservation processes such as planning, monitoring, and actual preservation operations such as migration and quality assurance, which have to be managed and integrated with the repositories. This article presents a suite of systems designed to support the preservation lifecycle in repositories. The SCAPE Planning and Watch suite provides the framework and toolset for controlling and monitoring scalable preservation operations. We present the main components for content profiling, preservation planning, and monitoring, and show how they can be combined to support scalable management of preservation over time.

Keywords

Digital Preservation, Preservation Planning, Preservation Watch, Content Profiling, Characterization, Scalability

1. MOTIVATION

The main focus of most digital repositories is to provide content access to its user community. To keep the content authentic and understandable to the user community on the long-term requires continuous monitoring, planning, and execution of corrective actions when needed to minimize risks and ensure continuous access. These processes need to be put together properly and integrated with repositories. The set of integrated digital preservation processes cover what we call the preservation lifecycle.

Scalability requires that automated tools support these processes, and a number of tools has emerged over the years to address parts of these processes. This ranges from aspects such as characterization, where tool such as JHove[1] and FITS[2] are commonly used, to preservation planning. Here, the preservation planning framework and tool *Plato* provides a trustworthy method and support for decision making. A key challenge here is to scale the decision making support to enable decision makers to manage large collections effectively and efficiently, and to integrate this support with repository systems. Additionally, continuous monitoring of the automated operations and the actual state of the repository and its content is needed to ensure that the repository's goals are met and risks and opportunities can be detected. This can only scale with automated tool support. Finally, a core aspect of this monitoring and in effect the starting point for the preservation lifecycle is an awareness of the the content itself and its various, potentially complex, and heterogeneous properties.

This demonstration presents an open, free, publicly available tool set that covers the crucial aspects of planning and monitoring outlined above and is integrated with a scalable environment for operations. We outline the key components and their conceptual interfaces, show how they address key issues of the preservation lifecycle, and demonstrate their interplay to address real-world preservation issues. All components are freely available on open licenses and can be accessed publicly on the web.

2. BACKGROUND

Digital preservation starts by knowing what content a repository has and what its key characteristics are. After digital objects have been characterized, this information can be aggregated and analysed and allows the content owner to get aware of the amount of content and the file format distribution. Moreover, this analysis process should support ways to drill down on important content characteristics to gain an in-depth understanding of preservation risks. Previous work has shown the value of format profiles across repositories [4], but was restricted to file formats only. Petrov showed that large-scale content profiling should not be restricted to the format label, but include more of the specific features that cause preservation issues, and that it is feasible to create and analyse large-scale in-depth profiles [7].

A preservation watch service has to cross-relate the results of this content characterization with institutional policies and external information about the technological, economic, so-

[1] `http://hul.harvard.edu/jhove/`

[2] `http://code.google.com/p/fits/`

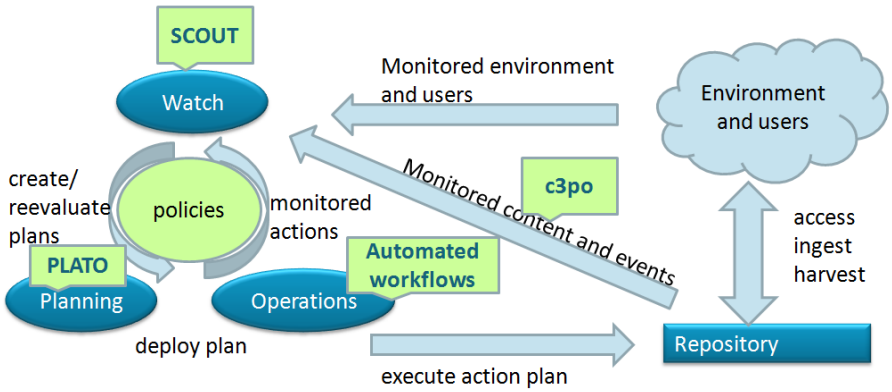

Figure 1: Preservation lifecycle

cial, and sometimes even political environment that provides the context of a repository. This allows for the identification of preservation risks and cost-reduction opportunities. Repository events such as ingest or download of content can also be useful for tracking producer and consumer trends and reveal preservation risks. The requirements on a preservation watch service have been described in [1].

These possible risks and opportunities should then be addressed by creating or revising a preservation plan. *Plato*, the preservation planning tool [2], guides the planner through a well defined and approved workflow. After the organisational setting is described, decision criteria are defined, and representative samples as well as possible actions are selected, these actions are applied to the sample objects in controlled experiments, and the outcome is measured. Based on this, the planner decides which operation should be implemented.

This is a very solid, trustworthy process, but creating a plan used to be rather effort intensive until recent improvements [3, 6]. Creating a plan was effort-intensive, and sharing experience was difficult: Plans could be made public and exported to XML, but collaborative planning was only possible on public plans. Integrating such planning with the organisational context, the strategies and operations of an organisation was difficult, and monitoring changes was a manual process.

The SCAPE Planning and Watch suite has been developed to provide an open, scalable environment for preservation control and monitoring. It builds on the conceptual foundation of *Plato*, supports step-wise integration of systems through open interfaces, and enables organisations to practically apply Planning and Watch in a scalable, semi-automated way.

The next section will describe the key elements of the suite and their key features and relationships, while Section 4 gives an overview on upcoming improvements.

3. SCAPE PLANNING AND WATCH

Figure 1 illustrates the preservation lifecycle and how the the preservation components interact to support it.

The lifecycle starts with a repository containing content that is preserved for a designated community over time. Apart from this community, there is a wide variety of factors of interest to be monitored, including other repositories, technical solutions, format risks and other aspects [1]. The SCAPE Planning and Watch (PW) suite is designed so that in principle, any repository can be connected. All interfaces are open, and a reference implementation for each API is being produced. This demonstration focuses on the current integration with the RODA repository[3].

In order to create a content profile, RODA includes a plugin that characterizes the files its holding using FITS [4]. The results of characterization are fed into the scalable content profiling tool *c3po* [5]. *c3po* is using a scale-out NoSQL approach based on MongoDB[6] to provide highly scalable profiling of millions of objects, and creates a content profile that is exported to XML. This content profile is then linked with other aspects in the Watch component, *Scout* [5]. *Scout* can match these content profiles against organisational objectives and detect mismatches such as format profiles that pose specific risks, the presence of risk factors such as compression, or other conditions that are of interest and should lead to a mitigation.

A core enabler here is a semantic model for organisational policies and objectives that represents the drivers and constraints for preservation processes using an extensible ontology[7].

Upon discovering a condition that requires intervention, *Scout* notifies the responsible decision maker, who can use the visual analysis features of *c3po* to get an in-depth understanding of the issue at hand.

The result of preservation planning is a preservation plan, which contains an executable workflow. Upon completion and approval of the plan, it can directly be deployed to the configured endpoint of the repository. RODA contains a plugin to activate execution of such a plan on the original content set for which the content profile was created, thus closing the first circle.

[3] http://roda.scape.keep.pt/
[4] http://code.google.com/p/fits/
[5] http://www.ifs.tuwien.ac.at/imp/c3po
[6] http://www.mongodb.org/
[7] http://www.purl.org/DP/control-policy

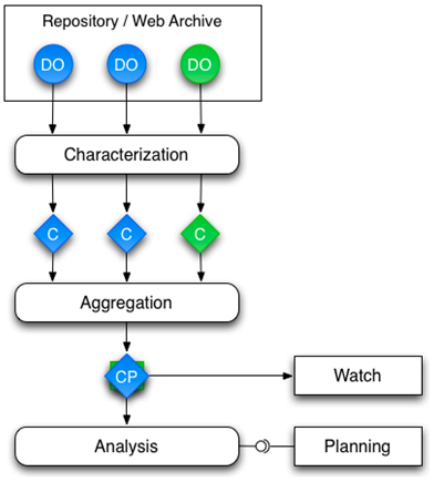

Figure 2: Content profiling

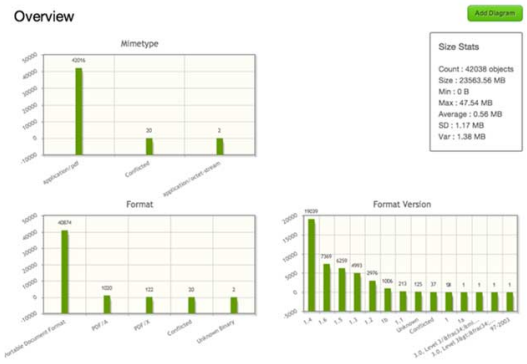

Figure 3: c3po showing a profile

Executing such operations on millions of objects will in the future be parallelized, for example on the SCAPE platform [8]. These operations will be monitored for compliance to the service level agreement, which will be done again using the monitoring component *Scout*.

3.1 Content profiling

c3po[8] enables in-depth analysis of the content of a repository. Figure 2 outlines the key steps of profiling. Content profiling covers Aggregation and Analysis of characterization data and distills it into a form suitable for Planning and Watch. The aggregated data can then be used by other services like preservation watch, moreover it is visualised via a web front-end, as shown in Figure 3.

The distribution of digital object characteristics can be analysed and used to create sub-sets with certain properties, using interactive filtering on the charts themselves. Different algorithms are available to calculate representative samples.

This aggregated information can then be exported as content profile and used for preservation planning and monitoring.

3.2 Preservation Watch

[8] https://github.com/peshkira/c3po

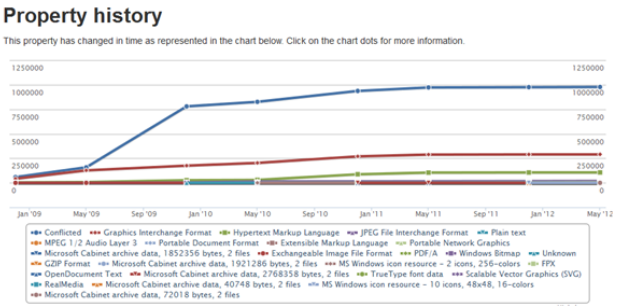

Figure 4: Property values over time in Scout

Scout[9], a preservation monitoring service, supports the scalable preservation planning process by implementing an automated service for collecting and analysing information on the preservation environment.

The information is collected by implementing different source adaptors. *Scout* has no restrictions on type of data that can be collected and it is planned to collect a variety of data from different sources like format and tool registries to repositories and policies. It already implements source adaptors for the PRONOM registry, content profiles from *c3po*, and policies. Repository adaptor, which is planned to monitor different events (ingest, access, migration) in a repository is currently being developed and with combination of content profiles from *c3po* will provide a complete overview of the current content in a repository and trends that are related to that content.

Once information is collected it is saved in a unified way to the knowledge base [5]. Build upon linked data principles the knowledge base supports reasoning on and analysis of the collected data. By providing different queries different information can be found. By now queries are used to provide a mechanism for automatic change detection. To do so a user simply deploys a trigger (watch condition) which will be executed periodically. When the condition is met a notification to the user is sent. Upon receiving the notification the user can initiate additional actions like preservation planning.

Scout has a simple web interface which allows operations like management, adding new adaptors and triggers and browsing the collected data. By operating over a longer period *Scout* is expected to have valuable collection of historic data. In Figure 4 evolution of file formats through time is shown. The resulting graph is created by analysing approximately 1.5 million of files gathered in the period from December 2008 to December 2012.

3.3 Preservation Planning

Creation of a plan is supported by *Plato* 4[10], which is integrated with a number of aspects that provide substantial improvements in planning efficiency by reducing previously manual steps:

- The core understanding of the organisational drivers

[9] https://github.com/openplanets/scout
[10] https://github.com/openplanets/plato

and constraints is provided by the semantic policy model which can be shared across members of the same organisation. This removes much of the contextual clarification that previously made starting a planning process difficult [3, 6]. When control policies have been defined, a preservation case can then be selected, and its information is applied to plan. Decision criteria are derived from the objectives and mapped to the corresponding measures. Later quality assurance components will be looked up based on these measures, and the results can be applied automatically.

- Content profiles created by *c3po* are directly integrated with the planning workflow, so that both the analytical step of analysing content sets and the technical processes of characterization and selection of sample data for experimentation are fully automated.
- Figure 5 shows how discovery of applicable preservation actions can rely on the discovery of preservation workflows shared and published using myExperiment [11], a social workflow sharing platform increasingly used by preservation practitioners. There are already a number of workflows for migration and quality assurance of image and audio files, some of them based on well-known tools like *FFmpeg*, others use new tools specifically developed for quality assurance, like jpylyzer[12].
- Experiment execution is highly eased by integrating the Taverna workflow engine [13] which is used to run the candidate preservation actions. The content profiles reference the permanent identifiers of the original objects, so that the byte-streams of sample objects can be directly used to test applicable actions on examples of the actual dataset that should be preserved.
- After the outcome of the preservation actions have been measured, the results have been evaluated, and the best alternative has been identified, a Preservation Action Plan can be created based on this workflow and the content profile, and once the plan is approved, it can be deployed to the configured repository endpoint, where it can be executed.

This is a major step upwards from previous iterations, where policies were implicit, content profiles manual, requirements specification effort-intensive, action discovery limited and plan deployment manual.

4. SUMMARY AND OUTLOOK

This demonstration presents a suite of systems that enable scalable monitoring and control of preservation in real-world environments. While each tool can be (and is) used independently, they are designed to be highly interoperable, so that the compound value contribution is larger than the sum of its parts. Using the SCAPE Planning and Watch tool suite, we can manage and streamline the continuous execution of digital preservation processes (the preservation lifecycle) on a systematic and semi-automatic way, mitigating some of the problems of large-scale digital preservation in an effective way.

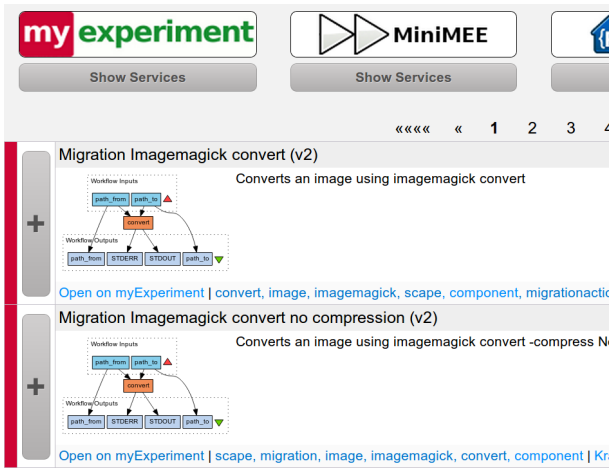

Figure 5: PLATO 4 querying myExperiment

As a next step the introduced APIs will be published, so they can be adopted by parties outside of SCAPE. Annotated SCAPE components will improve lookup, and ease composition, which enables to improve automation of quality assurance as well as generation of Preservation Action Plans.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

5. REFERENCES

[1] C. Becker, K. Duretec, P. Petrov, L. Faria, M. Ferreira, and J. C. Ramalho. Preservation watch: What to monitor and how. In *Proc. IPRES*, 2012.

[2] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proc. JCDL*, 2008.

[3] C. Becker and A. Rauber. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proc. JCDL 2011*, June 2011.

[4] T. Brody, L. Carr, J. Hey, A. Brown, and S. Hitchcock. Pronom-roar: Adding format profiles to a repository registry to inform preservation services. *IJDC*, 2(2), November 2007.

[5] L. Faria, C. Becker, P. Petrov, K. Duretec, M. Ferreira, and J. Ramalho. Design and architecture of a novel preservation watch system. In *Proc. ICADL*, 2012.

[6] H. Kulovits, C. Becker, and B. Andersen. Scalable preservation decisions: A controlled case study. *Archiving 2013*, 2013.

[7] P. Petrov and C. Becker. Large-scale content profiling for preservation analysis. In *Proc. IPRES*, 2012.

[8] R. Schmidt. An architectural overview of the scape preservation platform. In *Proc. IPRES*, 2012.

[11] `http://myexperiment.org`
[12] `https://github.com/openplanets/jpylyzer`
[13] `http://www.taverna.org`