The SCAPE preservation lifecycle

Kresimir Duretec, Artur Kulmukhametov,
Michael Kraxner, Markus Plangg,
Christoph Becker
Vienna University of Technology
Vienna, Austria
{firstname.lastname}@tuwien.ac.at

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

ABSTRACT

Continuous activities such as preservation monitoring, planning and operations, including the provisioning of access mechanisms or the creation of derivatives through migration, are needed to enable continuous access to content across evolving technological contexts without affecting the authenticity of digital objects. This article describes the SCAPE preservation suite, a loosely coupled set of systems and open APIs that facilitate scalable content profiling, monitoring, planning and workflow execution.

Categories and Subject Descriptors

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

Keywords

Repositories; Digital Preservation; Digital Curation; Scalability

1. INTRODUCTION

The main goal of digital repositories is to provide continuous access to the stored information. The changes in the socio-technical environment can pose challenges to providing access to the information in an understandable form connected to the contemporary computing ecosystem. Continuous activities such as monitoring, planning and operations, including the provisioning of access mechanisms or the creation of derivatives through migration, are needed to enable continuous access across evolving technological contexts without affecting the authenticity of the digital objects.

In recent years, the focus in digital repositories has turned to tackling challenges in scalability, in terms of the content size and technological variety.

A fully scalable system will be achieved only with the integration of preservation monitoring and planning components with digital repositories and their execution environment. While a number of components covering different aspects do exist, deeper integration between them is missing. For example, monitoring content starts with running tools such as JHove and FITS and extracting metadata describing each file. However, a more detailed analysis of content and its evolution over time is needed, and the analysis should lead seamlessly to providing an overview of the content, assessing the risks on various aspects, and devising treatment plans for these risks to be applied to the very specific set of objects to which these risks apply. This requires several distinct processes to collaborate and reference a common set of concepts and objects.

This demonstration presents a tool suite of several loosely coupled components that bring capabilities such as content analysis, preservation monitoring and preservation planning to a digital repository. The tools are integrated among themselves and with the repository through a set of openly specified APIs and provide substantial capabilities for scalable preservation ecosystems.

2. THE SCAPE PRESERVATION SUITE

The main components of the presented system are described in the following.

Content profiling has to face the combined challenges of vast numbers of objects, considerable diversity of formats and properties, and considerable gaps in characterization tool features, consistency, and correctness. The content profiling tool **C3PO**[1] aims to address this by providing a generic, modular architecture based on MapReduce, which enables it to provide highly scalable and highly customizable analytic capabilities together with an interactive web interface. The result is an aggregated, clearly described profile of a selected set of objects, which is generally a subset of the entire collection at hand as narrowed down by a curator for preservation purposes.

Preservation watch is supported by **Scout**[2], an extensible monitoring component designed to address the diversity of risks and opportunities through an architecture that allows arbitrary information sources to be plugged in dynamically [1]. These cover external information of relevance as well as drivers internal to the repository. Of these sources, two key elements are constituted by specific organizational policies, which are encoded semantically, and the content profiles generated by C3PO. **Preservation planning** is supported by an evolved version of the planning tool Plato[3], which provides a trustworthy and well documented

[1] https://github.com/openplanets/c3po
[2] https://github.com/openplanets/scout
[3] https://github.com/openplanets/plato

978-1-4799-5569-5/14/$31.00

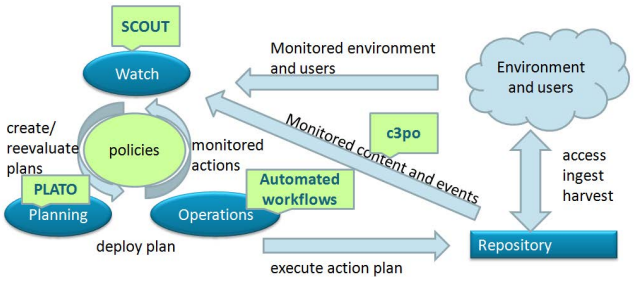

Figure 1: SCAPE preservation lifecycle [4]

decision support system [2]. Plans can be created to specifically treat issues identified through Scout, leveraging the policies, content profiles and criteria identified in these activities.

Without a clear translation into executable processes, the execution of preservation plans is difficult, and the risks of unnoticed errors in the operational execution arises. Additionally, these operations need to be monitored over time. Besides, much of the difficulty in preservation planning arises from lack of information, lack of tool support, and technical difficulties in experimentation [3]. Therefore **operations** are increasingly designed and tested as executable workflows using the Taverna workflow engine and published on myExperiment[4]. A seamless experience of evaluation and decision making is challenging to provide, but the integration with a workflow environment provides considerable support and facilitates the reuse of workflows across organizations.

Two connecting elements are paramount to the success of this design: (1) A small, extensible **controlled vocabulary** of quality elements, attributes and measures, that can be shared across all components has been created and is accessible at `purl.org/DP/quality`. The components increasingly leverage these annotations to provide resource discovery and clear labelling of measures, workflow descriptions and decision criteria to be collected and monitored. (2) A set of **open APIs** facilitates the integration of any of these components, or multiple components, with arbitrary repository or asset management systems[5].

Figure1 shows the SCAPE **preservation lifecycle model**, the main components, and the interactions between these components. The challenges, goals and designs are described in detail in [5].

The lifecycle starts in the digital repository. This demonstration focuses on the RODA digital repository, the first to support the full set of APIs. The repository holds a certain amount of content. To analyse this content RODA uses C3PO, which performs an in-depth analysis of the content by aggregating provided FITS metadata and analysing the aggregated data. This process creates a collection profile summarizing the key characteristics of a certain collection. The repository is also connected to an execution environment, which in this case is a Taverna server. The operations in this environment, and constant access and ingest events from the repository, are fed into the monitoring component. Scout monitors the repository and its environment, combines the information and detects risks and opportunities. Scout connects to C3PO and gathers information about collections stored in the repository. This provides a continuous monitoring of changes in collections. The repository events are monitored by connecting to the Report API implemented by the repository. Furthermore the repository environment is monitored by collecting information from different sources such as format catalogues. Scout can match collected information about the content with the organizational objectives and detects violations such as the existence of a format with compression scheme. This is enabled by a model of control policies expressed in a machine understandable language.

Once Scout detects states which might require some action it will notify the decision maker, who can start the preservation planning process to address the detected issue with the optimal operation available. Plato 4 downloads sample records from the repository over the Data Connector API, understands control policies, and automatically extracts decision criteria from them. Once published in myExperiment, different components are discoverable by Plato.

A newly created or revised preservation plan can be deployed to the repository via the Plan management API. The plan contains a Taverna workflow which combines the selected action with quality assurance components. As the repository has an execution environment which understands such workflows, the plan can be executed immediately without human intervention. The plan explicitly identifies the files to which it applies and provides a set of conditions that need to be satisfied. These conditions are simultaneously deployed to Scout, which monitors the execution of a plan and its conformance to the specified levels of quality.

All components are published under a free license. It is envisioned that they will be integrated with further repository environments in the future to provide some of the key capabilities needed to preserve the data deluge.

Acknowledgements

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

3. REFERENCES

[1] L. Faria, P. Petrov, K. Duretec, C. Becker, M. Ferreira, and J. Ramalho, "Design and architecture of a novel preservation watch system," in *Proc. ICADL*, ser. LNCS. Springer, 2012, vol. 7634, pp. 168–178.

[2] C. Becker, H. Kulovits, A. Rauber, and H. Hofman, "Plato: A service oriented decision support system for preservation planning," in *Proc. JCDL*, Pittsburgh, PA, 2008, pp. 367–370.

[3] C. Becker and A. Rauber, "Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning," in *Proc. JCDL*, Ottawa, Canada, 2011, pp. 67–76.

[4] M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria, "The scape planning and watch suite," in *Proc. 10th Int. Conf. Preservation Digital Objects*, Lisbon, Portugal, 2013, pp. 262–265.

[5] C. Becker, L. Faria, and K. Duretec, "Scalable preservation intelligence for information longevity," *OCLC Systems & Services*, vol. to be published, 2014.

[4] `http://www.myexperiment.org/`

[5] `https://github.com/openplanets/scape-apis`