Supporting the preservation lifecycle in repositories

Luís Faria[1], Christoph Becker[2], Kresimir Duretec[2], Miguel Ferreira[1], José Carlos Ramalho[3]

[1] KEEP SOLUTIONS, [2] Technical University of Vienna, [3] University of Minho

To accomplish effective digital preservation, repositories need to be able to incorporate processes such as planning, monitoring and preservation operations. These processes feed into each other and create a continuous cycle that allows a repository to detect opportunities and risks and act accordingly.

Each of these digital preservation processes have already been extensively studied (Antunes et al., 2011; CCSDS, 2002; Hunter & Choudhury, 2006) and tools to support each process have already been developed (Asseg et al., 2013; Becker et al., 2009; Faria et al., 2012), but many repository implementations still lack complete and continuous digital preservation features. This paper presents a global view on digital preservation processes and how they fit together in a digital preservation cycle. Furthermore, it describes tools that support these processes and explains how to incrementally integrate them into digital repositories providing a complete systematic and semi-automatic digital preservation system.

Digital preservation in current repository implementations

The main focus of most digital repositories is to provide content access to its user community. However, keeping the content authentic and understandable by the user community on the long-term requires continuous monitoring, planning and execution of corrective actions when needed. These processes need to be put together properly so they can be integrated with repositories.

Many implementations of digital repositories assume these digital preservation processes are manual or completely detached from native the repository workflows and the digital object lifecycle. This limits the ability to scale processes that achieve digital preservation and with the escalating growth of volume and heterogeneity of data it may become unfeasible for the repository to provide authentic access to digital content. What is needed is a clearly defined and interoperable set of processes that work together to produce a continuous, well-managed preservation lifecycle, continuously adapting to the changing environment.

The purpose of this article is to present such a suite of processes, all of which are already supported by open and free software. We present the key elements that are required, explain how they are supported by tools, and point to openly available API specifications that can be used to integrate them with virtually any repository system. We are further pointing to existing reference implementations that showcase the benefit of this integrated preservation lifecycle. We discuss the current state of interoperability between these processes and repositories and outline next steps ahead.

The digital preservation lifecycle

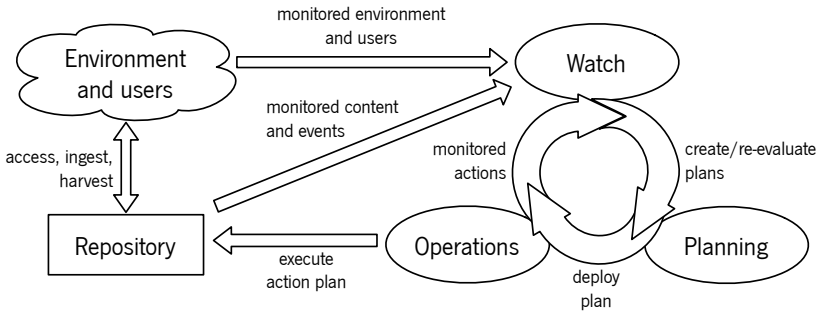

Figure 1: Digital preservation lifecycle

Figure 1 shows the key building blocks of the preservation lifecycle. To preserve content, the preservation risks that hinder the continuous and authentic access to the content need to be identified and continuously monitored. To this purpose, a continuous watch process monitors the alignment of what the repository has and does with its context, the technical environment and its users. Digital preservation starts by understanding what content a repository holds and what are the specific characteristics of that content. This process is supported by the characterization of content and allows a content owner to be aware of content volumes, characteristics, format distributions, and specific peculiarities such as digital rights management issues, complex content elements, or other preservation risks.

The characterization process feeds the key characteristics of the monitored content into the preservation watch process that should cross-relate the results of this internal content characterization with the institutional policies and external information about the technological, economic, social and political environment that the repository is set upon, allowing for the identification of preservation risks and cost-reduction opportunities. Checking the conformance of content with the owner's expectations or policies, identifying format or technological obsolescence in content or comparing content profile with other repositories can reveal preservation risks. Repository events, e.g. ingest and download of content, can also be useful for tracking producer and consumer trends and can be used to uncover preservation risks.

These possible risks and opportunities should then be analysed by preservation planning. The planning process carefully examines the risks or opportunities, having in mind the institution's goals, policies, objectives and constraints. It evaluates and compares possible alternatives and produces an action plan that defines what operations should be implemented and the reasoning that supports this decision.

An action plan is deployed into the operations process that orchestrates the execution of the necessary actions on the repository content, if necessary in large-scale, and integrates the result back in the repository. These operations include characterisation, quality assurance, migration and emulation, metadata, and reporting. The operations process should provide to the watch process information about the executed actions (or monitored actions), such as quality assurance measurements, to be sure that the results conform to the expected. Also, all assumptions about internal and external information taken by the planning process should be continuously monitored so the action plans (to do some action or to remain idle) remain valid. Once a plan becomes "invalid" the preservation planning process should be called upon to re-evaluate the plan, creating a continuous life cycle that ensures content remains preserved.

Recommended tools to support each preservation process

Each of the preservation processes can be done manually, but the common increasing volume and heterogeneity of documents in institutions make it necessary that tools exist to support and automate part of these processes. In this section we present tools to support each of the described processes that work well with each other. This list does not aim to be complete, but to serve as a recommendation based on experience and the development and integration work done in the SCAPE project[1].

There are several tools for content characterization (FITS[2], Apache Tika[3], ffprobe[4], etc.), some are very specific of the file format they work with, other wrap several tools together and work with a larger set of file formats and object classes. These tools provide technical information about the files and their key characteristics. However, these tools do not provide aggregation and analysis of these characteristics, something considered necessary to feed back information into the watch and planning processes. To fill this gap, the C3PO tool[5] was developed (Petrov & Becker, 2012). C3PO collects information from characterization tools and provides a content profile, i.e. an aggregated view of the content characteristics, necessary to support the watch process. Furthermore, the tool analyses the content and allows selection of representative datasets, which are necessary for the planning process. Also, C3PO provides an interface for browsing and drilldown of content characteristics and a programmatic API.

Scout[6] is a preservation watch system that provides an ontological knowledge base to centralize all necessary information to detect preservation risks and opportunities (Becker et al., 2012; Faria et al., 2012). It uses plugins to allow easy integration of new sources of information, as file format registries, tools for characterization, migration and quality assurance, policies, human knowledge and others. The knowledge base can be easily browsed and triggers can be installed to automatically notify users of new risks and opportunities. Examples of such notification could be: content fails to conform to defined policies, a format became obsolete or new tools able to render your content are available.

Plato[7] is a well-established tool for systematic preservation planning. It allows definition of preservation objectives, criteria and restrictions necessary for decision-making and helps with the evaluation of all action alternatives, arriving to a well-determined best solution, documenting all reasoning behind the decisions, and providing traceability, one of the basis for maintaining the authenticity of digital assets (Becker et al., 2009). The result of preservation planning is an action plan that, besides documenting the process itself, defines the necessary actions to perform on content.

[1] http://www.scape-project.eu
[2] https://code.google.com/p/fits/
[3] http://tika.apache.org
[4] http://ffmpeg.org/ffprobe.html
[5] http://ifs.tuwien.ac.at/imp/c3po
[6] https://github.com/openplanets/scout
[7] http://ifs.tuwien.ac.at/dp/plato

If, in one hand, the actions to be performed on content raise feasibility concerns due to the content volume or the action computing intensiveness, scalable platforms need to be taken into consideration. The SCAPE platform (Schmidt, 2012) provides guidelines on how to deploy such a platform to support execution of large-scale preservation actions. On the other hand, if the scalability is not a concern, less complex platforms can provide the same action plan execution features, such as the workflow engine Taverna[8]. Furthermore, example workflows of preservation action plans and components, e.g. characterization, migration and quality assurance, can be found and shared in the myExperiment site[9].

Architecture for repository integration into a full preservation lifecycle

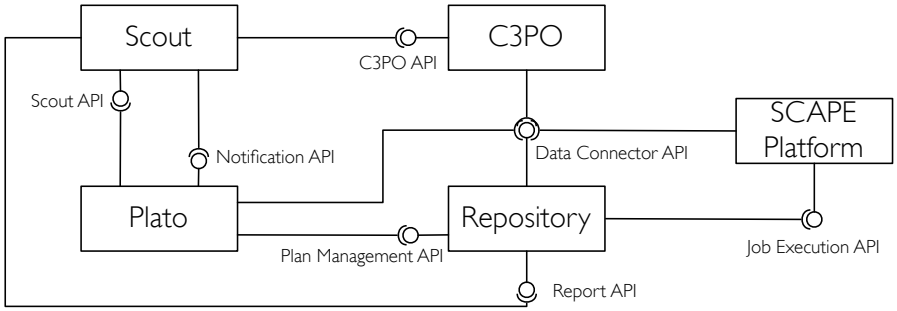

Figure 2: Preservation lifecycle architecture (software components and interfaces)

Figure 2 depicts all software components necessary for the preservation lifecycle (already described in the previous section) and focuses on the interfaces between each component. This is not a strict architecture because any of the software components can be skipped and the process it supports can be done manually or with other tools. Every programmatic interface has analogous human interface that achieves the same functionality. This is, therefore, an open and loosely coupled architecture that can be incrementally integrated into a repository implementation.

A repository can integrate into this preservation lifecycle architecture by implementing three interfaces:

1. **Data Connector API**: Interface to create, retrieve, search and update digital objects within a repository (Asseg et al., 2013).
2. **Report API**: Interface to retrieve information about events that take place on a repository, e.g. ingest, access, and preservation operations (Asseg et al., 2013).
3. **Plan Management API:** Interface to manage and execute preservation plans (Asseg et al., 2013). The implementation of the Plan Management API can use the Job Execution API to actively perform preservation operations as defined by a preservation action plan. The Job Execution API provides an interface for performing and monitoring parallel data processing operations (jobs or workflows) on the platform infrastructure.

Reference implementations of all above APIs are being developed for the Fedora Commons based repositories RODA[10] and eSciDoc[11] with focus on creating reusable components that could help with the development of APIs for other repository implementations.

Conclusion

Many digital repositories lack continuous digital preservation features and consider preservation processes as a manual endeavour, detached from the repository system. But this approach is not scalable and becomes unsustainable as the volume and heterogeneity grows. Integrating repositories with a complete, continuous and systematic preservation lifecycle, streamlining information between all digital preservation processes, becomes necessary to cope with the large-scale requirements of modern institutional repositories.

This paper presents the architecture for bringing together all necessary software components that support the lifecycle of digital preservation, focusing on the APIs that enable integration with a repository. The API specifications and reference implementations in RODA and eSciDoc enable new repository implementations to better integrate with a full digital preservation lifecycle, enabling systematic, semi-automatic, large-scale digital preservation of content. The loosely coupled architecture presented allows partial integration with digital preservation processes, as they can be employed independently of others. The composition of the API

[8] http://www.taverna.org.uk
[9] http://www.myexperiment.org
[10] http://roda-community.org
[11] https://www.escidoc.org

specifications, reference implementations and loosely coupled architecture aims to reduce the obstacles of adding digital preservation features to digital repository implementations.

All presented tools have already a released version and all APIs for repository integration have been formally specified. The reference implementations of the APIs are in progress and will be ready until fall of 2013. A proof-of-concept implementation of the APIs is already available and is used for a round-trip showcase demonstration of the whole preservation lifecycle with a RODA repository instance.

Acknowledgements

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

References

Antunes, G., Barateiro, J., Becker, C., Borbinha, J., Proença, D., & Vieira, R. (2011). *SHAMAN Reference Architecture (version 3.0)*.

Asseg, F., Bacall, F., Barton, S., Castro, R., Hahn, M., Schenck, M., Schmidt, R., et al. (2013). *SCAPE D4.1: Architecture design*.

Becker, C., Duretec, K., Petrov, P., Faria, L., Ferreira, M., & Ramalho, J. C. (2012). Preservation Watch: What to monitor and how. *iPres'12*. Toronto, Canada.

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, *10*(4), 133–157. doi:10.1007/s00799-009-0057-1

CCSDS. (2002). *Reference Model for an Open Archival Information System (OAIS)*.

Faria, L., Petrov, P., Duretec, K., Becker, C., Ferreira, M., & Ramalho, J. C. (2012). Design and architecture of a novel preservation watch system. *ICADL'12* (pp. 168–178). Taipei, Taiwan: Springer. doi:10.1007/978-3-642-34752-8_23

Hunter, J., & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *IJDL*.

Petrov, P., & Becker, C. (2012). Large-scale content profiling for preservation analysis. *iPres'12*. Toronto, Canada.

Schmidt, R. (2012). An Architectural Overview of the SCAPE Preservation Platform. *iPres'12*. Toronto, Canada.