A Research Infrastructure for Preserving Digital Objects (Extended Abstract)

Rainer Schmidt, Ross King
AIT Austrian Institute of Technology
Vienna, Austria
Email: firstname.lastname@ait.ac.at

Andrew N. Jackson
The British Library
Boston Spa, West Yorkshire, UK
firstname.lastname@bl.uk

1 Introduction

Managing the ever-increasing volumes of data produced and stored in data repositories and digital libraries provides a research problem across many academic areas. Examples are data produced in areas like science, engineering, and the arts and humanities. The Planets project is developing a service-oriented environment for the development and evaluation of automated preservation strategies in the area of human-centric data. It focuses on the question of logically preserving digital materials, as opposed to the physical preservation of content bit-streams.

2 Overview

The Planets infrastructure provides an e-research and problem-solving environment that supports decision support for preservation planning as well as a testbed for the execution and scientific evaluation of repeatable preservation experiments. It supports a range of preservation tools for migration and emulation, automated metadata extraction, and ontology-based comparison of object metrics in order to automatically validate preservation strategies. The distributed environment is implemented as a service-oriented architecture that allows one to register data, add/discover preservation services, and execute defined experiments and complex workflows. The core of the system - known as the Planets Interoperability Framework (IF) - provides a set of core services as well as a graphical web portal. The portal framework supports a single-sign-on mechanism and integrates the end user applications with data repositories and a number of preservation action and other services, such as services for data/metadata management, preservation, information, and workflow execution [1]. The infrastructure is designed as a research environment for digital preservation and does not intend to implement a preservation archive. It aims to provide an integrated environment that allows a community of researchers to collaboratively explore digital preservation strategies based on a number of shared resources.

A common data abstraction, called Planets Digital Object, is employed in order to organize different data sources within the central data registry. The service framework also enforces a technical contract between the provider and the user of a Planets service. Planets preservation services expose lightweight interfaces that share a set of common features and are simple to implement. Provenance and other preservation metadata are automatically collected by the workflow execution engine and expressed through the digital object model.

3 Example Use-Case

Archived data must usually be first checked out of a managed repository environment in order to make the digital items available for experiments. The type of repository system and its internal data model highly depends on the data and the organization with custody. Therefore, it is required to incorporate data from different digital object management systems, data repositories, storage environments, and other data sources with the research environment. The retrieved records are then mapped to a generic object model, stored within the metadata repository, and made available to the user's workspace. In order to develop an appropriate preservation strategy for a larger digital collection it is necessary to perform a number of experiments that are typically based on a characteristic corpus of files. An example of a Planets workflow would be: for given file, first identify a file format, then validate the format, characterize it, migrate it to a new format, then characterize the new file, and compare with the original. Each of these steps may be carried out by a different component in the system architecture and must be recorded and reflected within the metadata model.

References

[1] R. Schmidt, R. King, A. Jackson, C. Wilson, F. Steeg, and P. Melms. A Framework for Distributed Preservation Workflows. In *Proceedings of iPres 2009, 5-6 October 2009, San Francisco, USA and the IJDC (to appear)*.