QQML 2009
International Conference on
Qualitative and Quantitative Methods in Libraries
26 - 29 of May 2009 Chania Crete Greece
www.isast.org

# A next generation distributed preservation framework for digital repositories: first results from the SHAMAN project

Perla Innocenti, Brian Aitken, Seamus Ross
HATII at the University of Glasgow
{p.innocenti, b.aitken, s.ross}@hatii.arts.gla.ac.uk

How can we deliver infrastructure capable of supporting the preservation of data, as well as the services that can be applied to those data, in ways that future unknown systems will understand? How can we provide a characterization of services that can be applied to data within a rule-based (as opposed to primarily metadata-based) system?

SHAMAN, an EU-funded FP7 Integrated Project (http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects-shaman_en.html), aims to provide a tangible answer to these questions, investigating the long-term preservation of large volumes of digital data in a distributed environment and developing a next generation preservation framework. This framework will be verifiable, open and extensible, with corresponding application solution environments for analysing, ingesting, managing, accessing and reusing information objects and data across for memory institution (library and archives), design and engineering segments, and e-science domains. SHAMAN explores data grid, digital library, persistent archive, content representation technologies and all aspects of digital preservation, from ingestion to dissemination in an environment where the collections, producers, consumers, and curators are geographically distributed and the content of the collections is of a dynamic nature.

SHAMAN's first major activity was to identify user and organisational requirements and suitable technologies for meeting them. This paper describes the outcomes of this investigation, which was based on empirical research ensuring the requirements reflect ground truth. HATII at the University of Glasgow led this process.

Until recently, much research and thinking in digital preservation and curation has been based on a narrow and often single institution focus. These analyses often have not reflected the consistent application of proven data collection methodologies, and in particular the application of industrially established techniques for system analysis and design, although we found four counter examples and have noted these for comparative purposes (i.e. CASPAR, Planets, ERA, and National Library of New Zealand). The SHAMAN team, to ensure that we have the broadest understanding of what would characterise an efficient and effective preservation system, have looked more broadly by investigating the needs of communities across the three SHAMAN domains: libraries and archives, escience, and engineering. In conducting the fifteen case studies in representative organizations across Europe, we also followed practice that proved valuable in earlier research conducted by ERPANET of ensuring that in each institution in our target pool we interviewed a cross section of individuals including preservation experts, mediators, and end-users. In addition we aligned the interview questions with the elements of the OAIS functional model. This provided a breadth of understanding of user needs and expectations. It has made it possible for us to draw some conclusions about the conceptual requirements for the SHAMAN infrastructure which cut across institutional domains.

Through the interviews and their analysis we developed a picture of the types of digital objects (e.g. documents, databases) and kinds of representation information (e.g. metadata) that organisations use in their business operations and to understand the policies that govern their use within these contexts. We supplemented our empirical work with empirical work done elsewhere, such as BRICKS, CASPAR, ERA, Planets, ICSTI and CENDI, NDIPP and the National Library of New Zealand. In considering the use cases and requirements developed by these initiatives, we were able to validate some of our findings and to extend our understanding of the digital preservation arena, laying a solid foundation for subsequent work in the SHAMAN project. The SHAMAN requirements model that we have developed provides a framework for libraries, archives, and centres handling data will underpin the long term accessibility of digital materials in distributed and policy driven preservation environments.