Profession
Mötesplats Borås
Forskning

Elena Maceviciute and Tom Wilson
Swedish School of Library and Information Science, University of Borås
elena.maceviciute@hb.se
2011

# Evaluating the SHAMAN framework by memory and industrial engineering institutions

## Introduction

Today, vast amounts of information in various media are being digitised to create digital libraries of one kind or another: of texts, of images, of maps, of engineering drawings, of films, and of multimedia. Enormous resources are being poured into digitisation programmes by cultural organizations of one kind or another, as well as by commercial companies such as Google. The question that is all too infrequently asked, however, is "How will these digital materials be accessed and used in the future?" Digitisation is sold on the proposition that the material is being preserved for posterity and is being made immediately available to a wider audience than could ever have been imagined fifty or even thirty years ago. But how can we ensure that what is digitised now will actually be readable with the technologies of even twenty years ahead? This is the question asked by digital preservation and this is the question to which the SHAMAN project is trying to find the answer.

Thus, by 'digital preservation' we mean not simple the capacity to recover digital files from storage, but also the capacity to use those files, regardless of the technology with which they were produced. This definition puts it clearly:

> *Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.* (ALCTS 2007)

To achieve this is not simply a matter of appropriate technologies: as the definition indicates an organization must have policies and strategies to support digital preservation, as well as the technological means to accomplish it. Therefore, some digital preservation projects, though mainly tackling the solution of technological problems, have to address digital preservation policies as well.

## The SHAMAN Project

As part of the European Commission's 7th Framework Program, the SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) project focuses on the development of technology for preservation environments and creation of a next generation digital preservation framework.

The aim of the SHAMAN Integrated Project is to investigate the long-term preservation of large volumes of digital data in a distributed environment by developing a preservation framework that is verifiable, open and extensible. The approach will investigate 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.

Furthermore, it is developing corresponding preservation tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives. Three prototypical applications are intended to support trials and validation of the result in memory institutions, industrial design and engineering and, finally, experimentally, also in scientific application domains. The SHAMAN data grid infrastructure was developed in close cooperation with US project partners.

To achieve this aim the SHAMAN framework investigated data grid and cloud technologies, digital library, persistent archive and information knowledge and content representation technologies to create preservation system prototypes, These prototypes characterise the preservation process sufficiently to prove that it is possible to replace preservation services without impacting the data or access to and reuse of it.

The work implemented in the SHAMAN IP falls into four main areas:

1. Identification of user and organisational requirements and of suitable technologies for meeting them.
2. The design, development and prototyping of processes addressing key elements of digital preservation and of relevant data analysis and context representation scenarios.
3. The validation and integration of the elements to create prototype test-bed environments and the evaluation and refinement of these test-beds within and across three domains.
4. Project coordination and exploitation, including dissemination, training, demonstration, evaluation, outreach and take-up and project management (Watry et al. 2008),

Two areas are most interesting from the point of view of digital preservation policies: identification of user and organisational requirements and evaluation of the test-beds by the end users of technology. It is interesting to note that the final evaluation of the test-beds reveals different issues that do not always appear in the first statement of requirements for the system. And the evaluation part is presented in this short paper.

So far two demonstrators have been evaluated: one for memory institutions (libraries and archives), the other for the industrial design and engineering domain.

## Demonstration and evaluation

The demonstrators of the SHAMAN embody the concept of information life-cycle management (fig. 1) assuming that information should be understandable by future users and that any re-use of information, including the development of services, should be possible for future technologies. Therefore, changes in standards for encoding the information and any dynamics in its representation must be managed throughout the life-cycle. The SHAMAN framework should allow sufficient preservation information to be defined so that it will be possible to replace one or all of its components with new or different, but functionally equivalent, components without adverse impact on the preserved digital data. The framework envisages distributed archives accessible through data grid networks.

The life-cycle of preservation starts with the production of digital objects. It is followed by a usual appraisal and packaging of digital objects for preservation system that prepare information for the incorporation of digital objects into a system (pre-ingest). The packaged digital objects are then imported into the system and archived. In the system digital objects are managed to be accessible over time. They can be accessed from the system, unpacked and integrated into different working environments for various purposes and re-used for different purposes from those for which they were created.

The same digital preservation cycle is used for creating scenarios and evaluating the SHAMAN outcomes, however, the scenarios are adapted to the specific domain of interest.

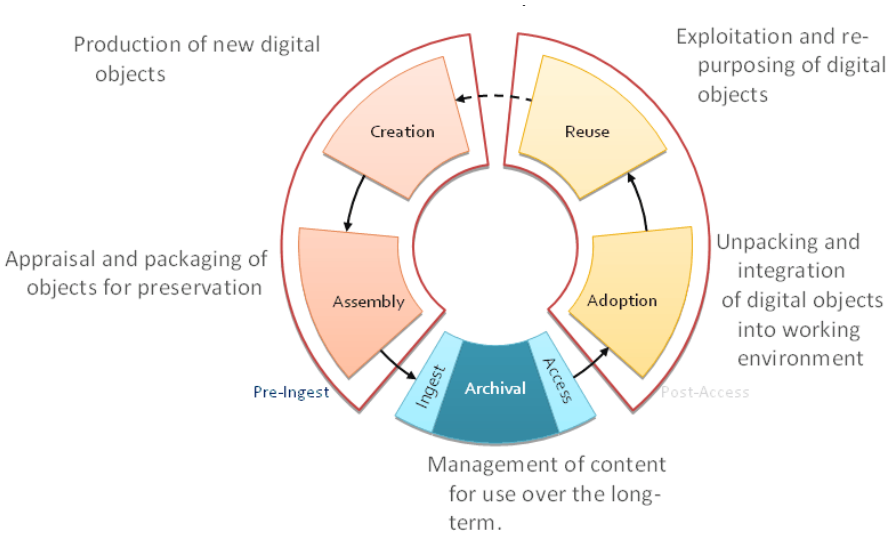
Production of new digital objects
Creation
Reuse
Exploitation and re-purposing of digital objects
Appraisal and packaging of objects for preservation
Assembly
Adoption
Unpacking and integration of digital objects into working environment
Pre-Ingest
Ingest
Archival
Access
Post-Access
Management of content for use over the long-term.

**Fig. 1**: Life cycle of preservation

The memory institutions were represented by librarians and archivists in three focus groups in Frankfurt, Vilnius, and Glasgow.

For the memory institution domain the main scenarios used were: the preparation of the textual digitally born material for inclusion into the preservation system based on the data grid (equals indexing and archiving books in libraries); the ingesting of digitised materials into the preservation system. The processes demonstrated were:

**Assembly:** designing the storage structures for digital objects and collections; creating and uploading the Submission Information Packages to the temporary area of the archive.

**Archival import and ingest:** uploading the Information Packages to the pending area of the archive and enforcing appropriate policies for processing, virus scanning, creating and archiving an Archival Information Package for the digital objects.

**Archival access:** discovery of digital objects using the Cheshire 3 Web interface, which allows generation of the table of contents of the object, requests delivery of the content and ensures security through separate login.

**Adoption:** the Fab4 Multivalent browser renders the data objects without migration to a newer file format and can add functionality to digital objects through behaviour lenses.

**Re-Use:** through layering annotations on the object which constitute new digital objects and may be re-ingested, indexed and archived with the original.

The demonstration for engineering and industrial design was carried out in Philips Consumer Lifestyle division at Eindhoven.

"Analysis of the engineering domain led to two key insights that guide the integration of long-term preservation into the Engineering domain (Fig. 2). As it is the central data management system in design and engineering, the Product Life-cycle management (PLM) system is the one and only software system the archive interfaces with. For integration with design and engineering processes, Release for Production (RFP) is the one and only event right for triggering archival of product data, otherwise inconsistent data states are likely." (SHAMAN 2011: 16).

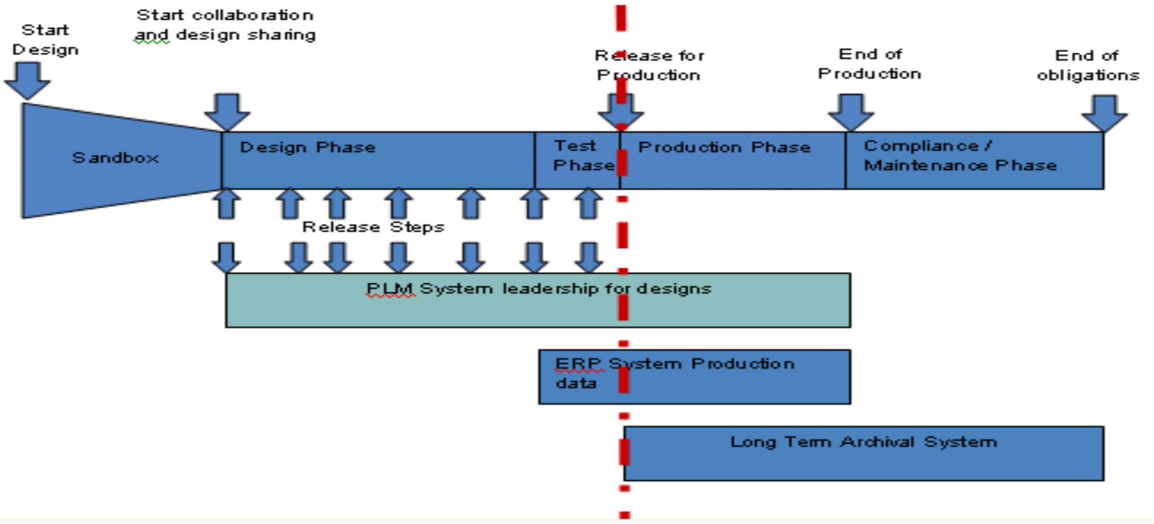
Start Design
Start collaboration and design sharing
Release for Production
End of Production
End of obligations
Sandbox
Design Phase
Test Phase
Production Phase
Compliance / Maintenance Phase
Release Steps
PLM System leadership for designs
ERP System Production data
Long Term Archival System

**Fig 2:** Phases in the engineering process and involvement of relevant data management systems along the phases (ISP2 Presentation)

So, the archival system interfacing with the PLM and capturing the design data was selected to start the presentation, closely followed by the normalization of the captured data and access of it through the iRODS metadata repository.

The central position in the demonstration event was allocated to the capturing of the collaborative context within the design processes and storing the date in the digital preservation together with other product design data. Open Conjurer fusing the ideas of social semantic networks and collaborative knowledge creation exploited a scenario of collaborative decision making and capturing the data on these decisions. It was demonstrated through the concrete example of working on a TV set:

The possibilities of evolving ontologies for knowledge discovery in digital preservation systems were demonstrated in the final part of the presentation, which was based on the Philips ideation process and included comparison and merging of the actual ontologies used in Philips.

## What do the evaluation of digital preservation system prototypes reveal?

*Overall concerns of memory institutions*

All memory institutions participating in evaluation had a mandate to preserve data for the indefinite future ("for eternity"). The preservation objects included a wide variety of formats, not only printed or digital (digitized), but also manuscripts, sound and video recordings, parchments, research and administrative data, etc.

Despite having a mandate, only three institutions had some kind of digital preservation policy. The participants of the focus groups made it clear that in most cases they have digitisation policy and sometimes it includes preservation aspects. In almost all cases there was no official digital preservation policy. This is also supported by the answers to the question about the budget in responding organizations: one declared that it receives adequate budget for planned preservation activities, two had inadequate budget and the rest did not allocate separate financing for preservation at all.

On the other hand, only one customer organization answered that it does not use any preservation technology. The rest named a variety of technological means used for digitisation and preservation including, digitisation (scanners), storage (disc or tape libraries), digital preservation (emulation, migration, digital surrogates) or digital repository (Fedora repository software). In addition, many longer-serving members of the participant groups had ten or more years experience in the preservation of digital objects.

The responsibility for digital preservation is distributed differently over the customer organizations. Top management is involved in four institutions, in two of them it is middle management that takes care of digital preservation. Data curators and librarians are involved in six organizations, IT specialists in four of them. Usually, professionals share this responsibility, i.e., data curators and librarians work together with IT specialists. It seems that customer organizations have competent specialists, as all of them have made a clear and informed distinction between digital archiving and preservation, though the answers show that the difference between the two is perceived in different ways. It can be related to time of preservation, access, function, object or several features at the same time.

Participants with responsibilities for digital preservation held, on average, approximately three interconnected roles. Fifteen of the 23 participants who were involved in digital preservation work were responsible for the formulation of policies and procedures, despite 21 participants (77.7% of the sample group) indicating that they operated without a documented digital preservation policy. Seven participants selected materials for digitisation; the same number managed the digitisation process; 7 of the 23 were responsible for "defining the organizational need for technology"; 5 curated digital data; 6 were responsible for "selecting and adopting digital preservation systems"; 6 managed or provided IT services for preservation; 6 were involved in the development of digital preservation software; whilst the same number again held "other" responsibilities including research, project management and managing archival teams and facilities (both physical and digital).

Three organizations did not specify user groups for digitally preserved materials as their preservation policy was not yet formulated. The others mentioned several groups, mainly corresponding to the groups that they serve generally. The archives target scholars, family historians, citizens as well as government bodies (legal and fiscal), and courts in the first place. The libraries focus on scholars, students, teachers, other libraries. The differences in audiences also depend on the profile of the institution and the nature of its collections (e.g., archives may serve film-makers if they hold a large amount of film materials).

Most of the concerns expressed and questions asked by the professional librarians and archivists who participated in the evaluation, fell within the scope of practical problems which they must routinely solve in their jobs. These problems occur on different levels:

1) *the diversity on the level of a document*: how to preserve different fonts and how to deal with manuscript documents or multimedia documents in preservation; the concern about non-book documents absent from the demonstrator was very high, and how to work with already obsolete formats was a concern;
2) *the diversity on the level of collections*: preservation of large collections of audio and visual documents in a variety of formats;
3) *the level of metadata*: how much to collect, how to collect it in cost-effective manner, how to extract and select what will be *really* necessary for future identification and use;
4) *level of discovery and finding aids*: linguistic support in terms of controlled vocabularies (ontologies and rules engines);
5) *addressing workflows on a more concrete level*, e.g., for institutions applying the principles of value expertise the check modules will be necessary;

6) *cost-effectiveness*; addressing the costs of storage, metadata extraction, URI structures, etc., will affect digital preservation policies and processes.

Despite the concerns and problems expressed, SHAMAN was perceived as promising for long-term preservation. The IT specialists noted that it is the composite of SHAMAN principles that makes it valuable and innovative in this respect. The librarians and archivists thought that a very influential direction for SHAMAN is the *automation* of processes, as all participants expressed concerns over costs. SHAMAN also presents an overall way of thinking about digital preservation processes and their associated policy, even if innovation is not the concern of practice. Practitioners want effective implementation: in this respect, the ease of integration of the demonstrated approaches with existing environments was seen as an asset. However, other factors will ultimately affect the willingness to implement SHAMAN approaches, such as ease of use and learning, conditions of use, users' trust in the system, and the advantages when compared with similar systems.

*Overall response of industrial engineers*

In the case of ISP2 our target customer organizations (industrial design and engineering companies) were represented by Philips Consumer Lifestyle Division and, more specifically, the Audio, Video, Multimedia and Accessories business area. It was pointed out that in this area, the products have relatively short lifetimes and that, consequently, the need for long-term digital preservation was limited. The maximum estimate of the time for which documents would need to be preserved was fourteen years and the potential users of digital preservation software believed that existing systems were perfectly capable of coping with preservation over that period. The nature of the technology and the pace of development within the industry militates against re-use of earlier technologies, although earlier ideas that were originally not capable of being realised in a product could be re-used. Such re-use appeared to depend more upon ensuring that the language in which earlier ideas were expressed was 'understandable' to modern search capabilities in systems. The example was given of '3D television', which was earlier known as a 'stereoscopic display'.

Participants pointed out that, although their needs for digital preservation were limited, other business areas in the division were more likely to have greater needs, for example, the Health and Wellness area, which had legal obligations to retain data. The theoretical justification for the SHAMAN framework appeared to be of little interest to them: their concerns centred on the 'business case', would the costs of implementing a SHAMAN-based set of processes be justified by the savings?

The outcome of the evaluation session as a whole (including the question and answer sessions during the presentations) can be summarised as follows:

a) The participants from this particular business area of the 'Lifestyle' division perceived no need, either personally, or from the company's point of view, for long-term preservation of design and engineering documentation.

b) Estimates of the length of time for which documents were archived at present ranged from "more than three years" through "seven years" to "fourteen years".

c) Improved search capabilities *were* seen as desirable, as add-ons to the existing archiving systems which were based on Product Lifecycle Management systems and Document Management Systems.

d) Automatic meta-data generation was seen as a desirable time-saving for engineers in their daily work.

e) The Open Conjurer sub-system attracted interest, though participants pointed out that they tended to work in face-to-face meetings rather than over networks and that if the system could be developed to capture information from such meetings, it could be of value to the company. They also stated that the usage of the social graph metaphor would be useful in order to initiate collaborative sessions and projects, making it easier to put the groups together, as the information is already present in the social graph.

As in the case of participants in the memory institutions domain, the comments of the engineers participating in the event were strongly influenced by the nature of their work, the types of products upon which they worked and the prevailing systems used in the company. There were frequent comments regarding '*saving the time of the engineer*' and a vision was presented of a desirable system in which all an engineer had to do was to deposit a document (of any kind) in a system, which would then carry out all the necessary operations to identify, index, code with metadata, and archive in the appropriate form without further intervention by the engineer. So, their primary requirement for software was that it should 'save the time of the engineer'. They felt that any systems that required more input from the user than they gave already would not be acceptable and that the optimum system would be one that accepted a document or design and then automatically extracted all the necessary metadata to enable it to be found and used in the future.

## Conclusion

'Digital preservation' is revealed by our work as a non-unitary concept: it depends upon *what* is digitised, the period over which the preserved material is anticipated to be of use, and the future end uses to which it is believed the material may be put. Clearly, memory institutions have the longest view, since they are the storehouses of preserved *physical* materials from, in some cases, hundreds of years in the past. However, even here, question of *which* institutions need to retain digital archives and which may draw upon those archives is a significant question.

On the other hand, our work in industry suggests that the useful time period over which digital objects may be seen as useful will depend mainly upon two factors: the lifetime of a product and legal requirements for preservation. In the case of TV sets, the production lifetime is quite limited, with new models coming out every two or three years, but in the case of health products from the same organization, legal requirements relating, for example, to testing may require records to be maintained for a much longer period of time. Even then, the expenditure has to be justified by the business benefits.

However, regardless of these issues, the need for preservation is high simply because of the pace of change in the technologies of digitisation, storage and access. File formats that were common 15 years ago, may be limited in their application today – think, for example, of the different formats used by Microsoft in its word processor. Still these formats are passed to the archives in organizations and some means must be available to read them in the future, if they are considered to be of use.

Storage technologies, too, change rapidly – when did you last see a 5.25" floppy disc? And of course, the very nature of computer technology changes all the time, with promises of optical computing, new alternatives to the silicon chip, and quantum computing.

In other words, in a rapidly changing technological environment simply to *archive* without ensuring *preservation* is less than adequate.

## Reference

Association for Library Collections & Technical Services. *Preservation and Reformatting Section. Working Group on Defining Digital Preservation.* (2007). *Definitions of digital preservation.* Chicago, IL: ALCTS. Retrieved 13 January, 2011 from http://www.ala.org/ala/mgrps/divs/alcts/about/contact/index.cfm (Archived by WebCite® at http://www.webcitation.org/5vhyJIsGN)

Watry P., Hasan A., Hemmje M. (2008). SHAMAN Vision: Contribution to section 1.1 of a deliverable 1.1. (2008.10.27).