Viewpoints on digital library issues

World Digital Libraries 1(1) v–x

A brief overview of the ongoing EU project SHAMAN: Sustaining Heritage Access through Multivalent ArchiviNg

Gobinda Chowdhury*

Dennis Nicholson**

In this special column, we aim to provide a brief overview of an ongoing EC (European Commission)-funded large-scale digital preservation project called SHAMAN (EU project IST FP7 216736). A collaborative 'Integrated Project', or IP, involving 18 international partners, SHAMAN began in December 2007 and will run for a period of approximately four years. It aims at developing a next-generation digital preservation framework for archiving and managing digital information, together with tools for enabling its implementation, focusing, in particular, on ensuring its ability to facilitate long-term access to digital information for users in memory institutions, the domain of engineering and re-engineering, and eScience. Beginning with a quick introduction to the concept and issue of digital preservation, this paper highlights the major goals of the SHAMAN project, providing a brief overview of its major activities and focusing, in particular, on preserving the preservation environment.

Introduction

For the past few years, large libraries and archives around the world have been actively engaged in the preservation of their digital materials. There has also been a good deal of supporting activity in the area in the form of standards work and research and development initiatives. In the standards area, the most influential effort has been the development of the OAIS (Open Archival Information System)[1] reference model. Accepted as a standard by ISO in 2003, this specifies a reference model for an open archival

[1] <http://public.ccsds.org/publications/archive/650x0b1.pdf>

*Professor, Information and Knowledge Management, University of Technology, Sydney
**Director, Centre for Digital Library Research, University of Strathclyde, Glasgow, G1 1XH

information system and has influenced a wide range of activities in the area. A good deal of research and development funding has also been targeted at digital preservation (see, for example, the work of NDIPP,[2] NARA (National Archives and Records Administration),[3] and OCLC[4] in the United States; Cunningham (2007) on the Australian scene over the previous 10 years; and the PADI portal[5] as one excellent source covering all kinds of information and resources on digital archiving around the globe). Not surprisingly, this high level of activity has also been paralleled in the European Union, where a range of DP projects have been funded, including DPE,[6] Planets,[7] CASPAR,[8] and, more recently, LiWa,[9] PROTAGE,[10] and SHAMAN.[11]

The field is still in its infancy, relatively speaking, and there is a continuing development (often within projects like those listed above) of even basic concepts. The AHDS Digital Preservation Glossary[12] offers two alternative definitions of digital preservation, describing it as either 'storage, maintenance, and access to digital resources over the long term' or 'ensuring the usability of a digital resource through changing technological regimes, with a minimum loss of the resource's intellectual content'. Understanding of the term is still developing to some extent. For example, the ALA's PARS (Preservation and Reformatting Section) (2007) recently conducted a discussion on the topic and came up with short, medium, and long definitions,[13] of which the medium one is possibly the most helpful here: '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 a period of time'. The digital content concerned can encompass almost anything in digital form, including, but not necessarily limited to, various formats of video and audio files, still images, 3D constructs and diagrams, animations, databases, eScience data sets, CAD (computer-aided design) drawings, websites, text files, and so on. Materials likely to need digital preservation may include any digital resource relevant to the business of an organization; for example, in a university, it may include materials relevant to research, record-keeping, and archival activities. In most (possibly all) cases, the aim is not simply preservation but preservation for access (as at least two of the definitions above make clear). In some cases, the difference is made explicit. For example, the digital preservation activities at TNA (The National Archives) of UK is based on two sets of activities: passive preservation, which provides secure storage, and active preservation, which ensures the continued accessibility of the stored records over time and across changing technologies (Brown 2007, p. 5).

The term digital preservation is closely associated with the (arguably slightly wider) term digital curation. It is used to denote the various activities involved in managing digital materials (research data, digital drawings, maps, audio, video, documents, and so on) so

[1] <http://public.ccsds.org/publications/archive/650x0b1.pdf>

[2] <http://www.digitalpreservation.gov/library/>

[3] <http://www.archives.gov/preservation/index.html>

[4] <http://www.clir.org/pubs/reports/pub107/bellinger.html>

[5] <http://www.nla.gov.au/padi/index.html>

[6] <http://www.digitalpreservationeurope.eu/>

[7] <http://www.planets-project.eu/>

[8] <http://www.casparpreserves.eu/caspar-project>

[9] <http://liwa-project.eu/index.php/about/>

[10] <http://www.protage.eu/>

[11] <http://shaman-ip.eu/>

[12] <http://ahds.ac.uk/preservation/preservation-glossary.pdf>

[13] <http://blogs.ala.org/digipres.php>

that they can be accessed and used by current and future generations of users. Beagrie (2006) notes that the term digital curation is relatively new and incorporates aspects of existing concepts such as 'data curation' and 'digital preservation', used primarily by the scientific and digital library communities, respectively. Other researchers have commented that 'preservation is an aspect of archiving, and archiving is an activity needed for curation. All three are concerned with managing change over time' (Lord and Macdonald 2003, p. 12).

Context and preserving the preservation environment

It is worth noting that the focus of digital preservation extends beyond the preservation of the digital materials themselves. For a variety of reasons, the preservation of metadata about the objects is also a requirement. One example of why this is of particular relevance to SHAMAN relates to the idea of context. As Cunningham (2007) points out, the meaning and value of records[14] depend on the contextual relationships surrounding the creation and use of records. In consequence, there is a recognized need to preserve information related to the environment and the context, along with the materials themselves, in order to ensure proper management and use of the archives by future generations. Context-related metadata can be as important in digital archiving as discovery and preservation metadata-and not only in the more traditional forms of digital library and archive. The need is also recognized in scientific fields. For example, Beagrie (2006) notes

> In the biological sciences, the term curation had been applied to the maintenance and publishing of databases such as the human genome and was therefore already implicitly digital. In this context added value is derived from annotation, linkage, and the management, validation, and editorial input of domain specialists employed to curate and publish the database.

SHAMAN particularly focuses on the need to preserve not just the digital object itself but its context as well, extending the idea of context to encompass not just the knowledge required to maintain an understanding of a digital object into the far future but also the technical environment required to store the object, manage its preservation, and render[15] it appropriately for future generations, that is, it has a focus on preserving the preservation environment itself.

Some of the ideas behind this are already represented in the literature. Examples are Moore and Smith (2007), who describe a new rule-driven approach, which enables all preservation processes (not just metadata) to be migrated onto new technologies; Moore (2008), who explains that the major challenge for an ideal preservation system is 'how to incorporate new technology effectively while conserving preservation properties such as authenticity, integrity, and chain of custody'; and Watry (2007, p. 42), who presents the SHAMAN view that the true test of a digital preservation system is whether

> it describes the entire preservation information context sufficiently well that the records can be migrated into an independent preservation environment without loss of authenticity or integrity. This requires migrating not only the records, but also the characterizations of the preservation environment context. The new preservation environment would have to apply the same management policies, the same preservation processes, use the same logical name spaces, and manage the same persistent state information.

[14] A term often used in DP circles to refer to digital resources generally

[15] Roughly speaking, preserve the ability to correctly represent its original look and feel.

The view taken is that an ideal digital preservation system should be able to pass on the information generated in the past to the future while reliably and consistently maintaining the authenticity, integrity, and provenance of the records, and that the preservation of context in the wider sense just described is key to that enterprise.

What is SHAMAN?

SHAMAN stands for Sustaining Heritage Access through Multivalent ArchiviNg. It is a large-scale digital preservation research project funded by the European Union under the Seventh Framework IST programme. The main aim of the SHAMAN project is to develop a next-generation digital preservation framework by developing appropriate preservation tools for analysis, ingest, management, access, and reuse of information objects and data across libraries and archives. The framework and the corresponding tools and technologies will be trialled and validated in three domains: scientific publishing, parliamentary archival, and industrial design and engineering.

SHAMAN is based on the collaborative efforts of 18 international research institutions/centres from nine countries around the world (see the SHAMAN website for a full list of participants) and has the following four major objectives.

1 To establish an open distributed resource management infrastructure framework enabling grid-based resource integration, reflecting, refining, and extending the OAIS model and taking advantage of the latest state-of-the-art virtualization and distribution technologies from the fields of grid computing, federated digital libraries, and persistent archives.

2 To develop and integrate technologies to support contextual and multivalent archival and preservation processes, which are adapted and significantly extended from the fields of content and document management and information systems.

3 To develop and integrate technologies to support semantic constraint-based collection management to target one of the key challenges in automating one class of digital preservation core functions.

4 To support the managing of future requirements by securing interoperability with future environments and maintaining essential properties of the preserved content.

The activities in the four-year long SHAMAN project are divided into 18 work packages, each with one or more deliverable(s), as follows.

1 Requirements analysis and identification of user scenarios
2 Design and specification of the SHAMAN digital preservation framework
3 Context capturing, representation, and management
4 Multivalent preservation interface and media engines
5 Data grid implementation
6 Harmonization, basic analysis, and ingest
7 Advanced information extraction and knowledge engineering
8 Managing shared collections
9 Interoperability with future environments
10 Maintaining essential properties
11 Document production, archival, access, and reuse in the context of memory institutions for scientific and governmental collections
12 Simple and connected object production, archival, and reuse in the industrial design and engineering domain
13 eScience data acquisition and harmonization testbed
14 Demonstration and evaluation
15 Training

16 Scientific/academic-oriented dissemination and exploitation
17 Industrial digital preservation technologies outreach and take-up
18 Project management

SHAMAN work to date

The project is still in its first year, and most of the work undertaken has been on the Requirements analysis and identification of user scenarios work package (called WP1) and on the Design and specification of the SHAMAN Digital preservation framework (called WP2). The former is led by HATII [16] at the University of Glasgow and the latter by CDLR[17] at the University of Strathclyde. Both work packages include participation from most of the other participants-University of Liverpool, UK; InConTec GmbH, Germany; the Swedish School of Library and Information Science, Sweden; Xerox Research Centre Europe, France; FernUniversität, Hagen, Germany; Philips Innovation Lab, the Netherlands; the German National Library, the University of Göttingen, Germany; the University in Magdeburg, Germany; Industrious Media from the UK; the University of Illinois, USA; and INESC-ID, Portugal, coordinated by INMARK, Spain.

WP1 is currently focused on the requirements and expectations of a variety of stakeholders, including service professionals and end-users. It is using interviews with stakeholders in the three domains of focus for informing the development of SHAMAN and preparing an evaluation framework for use in the later stages of the project. WP2 is delivering outcomes on *The assessment and extension of OAIS definitions and metadata standards and sets, Deficiencies and synergies of data grid technologies and tools, and The specification of the SHAMAN reference architecture*. To date, the major part of the work conducted is focused on the first deliverable on the refinement and extension of OAIS definitions and metadata sets, the aim being to inform about later work on the reference architecture and tools through (1) a state-of-the-art analysis of OAIS and the DP field generally and (b) the identification of key areas to be addressed in developing a next-generation digital preservation framework and, in time, the tools to implement it.

Summary

Digital preservation is a relatively new field and is still developing. A good deal of initiatives have been undertaken across the world on standards, research, development, and implementation. SHAMAN is a European project working at the leading edge of DP developments and focuses on the need to preserve not just the digital objects but also the environment in which they are being preserved. It is still in its initial stages. To date, the SHAMAN research team has primarily focused on WP1 and WP2. As it develops further, it will utilize new and cutting-edge technologies like data grid, iRODS, and multivalent data storage and rendering tools, along with a myriad of new technologies and practices to develop a novel next-generation digital preservation system.

Further information on the SHAMAN project and its outcomes will appear on the pages of the website as the project develops.

[16] <http://www.hatii.arts.gla.ac.uk/>
[17] <http://cdlr.strath.ac.uk/>

References

Beagrie N. 2006
Digital curation for science, digital libraries, and individuals
The International Journal of Digital Curation **1**(1)
Details available at <http://www.ijdc.net/ijdc/article/view/6/5>

Brown A. 2007
Developing practical approaches to active preservation
The International Journal of Digital Curation **2**(1): 3–11
Details available at <http://www.ijdc.net/ijdc/article/view/37/42>

Cunningham A. 2007
Digital curation/digital archiving: a view from the national archives of Australia
[DigCCurr 2007, International Symposium on Digital Curation, Chapel Hill, NC, 18–20 April 2007]
Details available at <http://www.ils.unc.edu/digccurr2007/papers/cunningham_paper_7.pdf>

Lord P and Macdonald A. 2003
e-Science Curation Report: data curation for e-science in the UK – an audit to establish requirements for future curation and provision
[Report prepared for the JISC Support of Research Committee]
Details available at <http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf>

Moore R. 2008
Towards a theory of digital preservation
The International Journal of Digital Curation **3**(1): 63–75
Details available at <http://www.ijdc.net/ijdc/article/view/63/82>

Moore R and Smith M. 2007
Automated validation of trusted digital repository assessment criteria
Journal of Digital Information **8**(2)
Details available at <http://dspace.mit.edu/handle/1721.1/39091>

Watry P. 2007
Digital preservation theory and application: transcontinental persistent archives testbed activity
The International Journal of Digital Curation **2**(2): 41–68
Details available at <http://www.ijdc.net/ijdc/article/view/43/50>