User requirements for a next generation digital preservation framework: analysis and implementation

Brian Aitken, Perla Innocenti; Humanities Advanced Technology and Information Institute (HATII); University of Glasgow; United Kingdom

Seamus Ross; Faculty of Information; University of Toronto; Canada

Leo Konstantelos; Humanities Advanced Technology and Information Institute (HATII); University of Glasgow; United Kingdom

Abstract

The EU-funded SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg, http://www.shaman-ip.eu/) project is investigating 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. During the initial stages of the project, a detailed user requirements analysis led by HATII at the University of Glasgow was conducted across three domains: memory institutions, industrial design and engineering, and e-science. This research pinpointed the needs and expectations that end-users and service providers feel should be met by such a preservation framework. This paper gives an overview of the requirements that were gathered, formulated and adopted by this project. It then discusses the outcomes of this empirical research and indicates both how these outcomes are being implemented within SHAMAN and how external parties may also benefit from the findings.

Approaches to digital preservation are often still ad hoc and based on a single institution focus. They frequently do not take into consideration the needs of the variety of actors who will come into contact with a system throughout the preservation lifecycle. This paper provides an insight into the preservation practices that a broad range of real-world organisations would like to follow and provides a discussion of how SHAMAN intends to meet the needs of the identified users.

Introduction

In order to sustain access to large number of digital objects over the long-term, digital libraries and archives must adopt a preservation framework that meets the needs of current users and is flexible enough to continue to meet the needs of future users. The aim of the SHAMAN Integrated Project (Sustaining Heritage Access through Multivalent ArchiviNg, http://www.shaman-ip.eu/) is to investigate and develop a long-term next generation digital preservation framework, together with corresponding application solution environments for analysing, ingesting, managing, accessing and reusing information objects and data.

On the basis of this framework environment, the project is building three prototype application solutions, testing and validating outcomes in three 'Domains of Focus': Memory Institutions (DOF1); Industrial Design and Engineering (DOF2); and e-Science (DOF3). Figure 1 demonstrates how these three domains will interact with the core methods for supporting digital preservation that SHAMAN will investigate.

In the development of the digital preservation framework, the project is employing and expanding upon a variety of existing technologies: data grid technologies will be employed to offer a future-proof digital preservation strategy for multiple organisations and communities; basic control mechanisms in a distributed environment will introduce transaction capabilities to digital preservation services; digital library services will provide access to data held within preservation environments; data modeling methods will be integrated to enable the rendering of obsolete data and media formats; the Multivalent document model will be incorporated to assist in the development of domain-specific uses for digital media; and a model for capturing, representing and managing context will be defined.

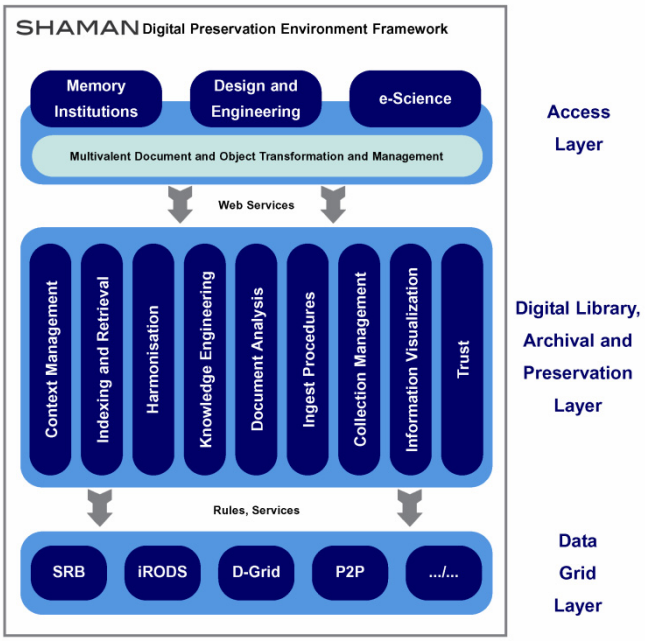

Figure 1. The SHAMAN Digital Preservation Conceptual Framework © SHAMAN Project

To ensure that the framework and application solutions developed by SHAMAN meet the needs and expectations of the three targeted domains, a period of user requirement analysis led by HATII at the University of Glasgow was undertaken during the initial phase of the project. This investigation built on and further advanced the approaches of previous relevant projects [2, 3, 4, 8] and practices [5, 7]. This process focused on the preservation policies that are currently in place at a selection of real-life

organisations including national libraries, governmental archives, multinational technology companies and scientific data grids. Additionally, the functionality these organisations would expect from a next generation preservation framework was also defined. Interviews were conducted at the targeted organisations and from these a series of abstracted use cases and user requirements were formulated. The results of this investigation have been presented in detail within the first part of the project deliverable SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol [1].

Findings of the SHAMAN user requirements process

The user requirements process defined the high-level requirements of the SHAMAN Preservation Framework: the fundamental properties that any system which adheres to the components of the SHAMAN Theory of Preservation must take into consideration. During the analysis of the interviews conducted at the eleven representative organisations across the three domains of focus it became apparent that the extracted use cases formed clusters based on their function as follows: General Regulations, Ingest Preparation, Ingest, Data Management and Archival Storage, Preservation Planning, Access and Post Access. These categories are broadly analogous to the OAIS functional entities.

The SHAMAN Preservation Framework will be firmly grounded in the conceptual and technical reference architecture provided by the OAIS model [6] (Fig 2). OAIS has served as a starting point for the infrastructure of SHAMAN, guiding the specification of relevant interchangeable system components. For this reason the grouping of user requirements by their corresponding OAIS functional entities was deemed to be the most effective approach. However, it is the intention of this project to refine and extend the OAIS model, especially in areas such as pre-ingest and post-access and the preservation of contextual information at these stages, therefore a number of user requirements for these proposed new stages were also documented.

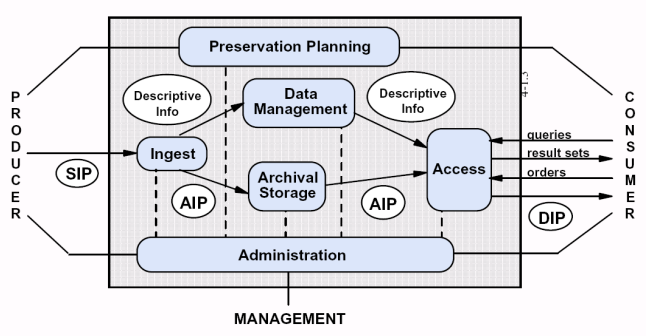

Figure 2. OAIS Functional* Entities (from *Reference Model for an Open Archival Information System (OAIS),* 2002, Fig 4-1*)

The specific user requirements that were documented can be found in the publicly available SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol [1], and an overview of the findings is presented here.

General Requirements

Requirements in this section relate to general administrative tasks such as user account management, error reporting and the definition and management of services that will be deployed as part of the Preservation Framework. Requirements in this category may be generally considered to relate to the 'Administration' OAIS functional entity.

The required general administrative tasks were found to be shared by all three targeted domains, with only a couple of exceptions. All three domains required user account management options and facilities for the definition and management of multiple user roles. Across the three domains authentication measures were also required, with authentication based on user accounts and user roles.

Similarly, all three domains required error reporting and error management procedures, plus the logging, validation and verification of all operations carried out within the long-term archive. Each domain also placed importance on ensuring the interoperability of the components utilised by the archive and desired facilities to manage these individual components.

Some domain-specific requirements were however noted. For example DOF2 required relationships to external systems to be managed and maintained, and both DOF1 and DOF2 specified a desire for facilities to define and manage acceptable representation information and data formats.

Pre-Ingest Requirements

Requirements in this section focus on the activities that may take place before digital objects are ingested into the long-term archive. Within this category are requirements relating to the production of materials, the definition of ingest workflows and any tasks carried out on digital objects and their accompanying metadata prior to ingest. Requirements in this section would generally be considered beyond the scope of the OAIS functional model.

A thread of commonality can be found throughout the identified pre-ingest requirements across the three targeted domains. Each domain required facilities to define, manage and maintain ingest preparation policies and ingest workflows for both the digital objects and their accompanying representation information. All three domains also expressed a desire for facilities to assemble digital objects and representation information prior to ingest. This process would include the transformation of data and representation information based on defined submission policies.

DOF2 and DOF3 also specified requirements relating to the validation of ingest workflows with sample data. This process would take place at the pre-ingest phase as a means of checking the suitability, completeness and consistency of the defined ingest workflows. In addition, DOF2 also required facilities to enable the configuration of automated ingestion processes.

Ingest Requirements

This section contains requirements relating to the ingestion of digital objects and their accompanying metadata into the long-term archive. Requirements in this category generally correspond to the 'Ingest' OAIS functional entity.

The core high-level ingest requirements were found to be shared amongst all three domains, specifically requirements relating to the management and maintenance of the ingestion of data and the management and maintenance of the ingestion of representation information. In addition to these high-level requirements, DOF1 and DOF2 specified some additional, shared requirements. These requirements relate to the need for data to be ingested from both external systems such as CAD programs, and externally hosted systems owned by a third party, such as a publisher. DOF1 and DOF2 also stated that the ingest process must be able to handle the ingestion of new versions of data that is already stored within the long-term archive and the association of new versions with previous versions.

Both DOF1 and DOF2 also specified that as part of the ingestion process it must be possible to include verification and validation facilities in order to ensure that ingest was successfully completed and that any encountered errors were logged.

Data Management and Archival Storage Requirements

These sections contain requirements relating to the management, maintenance and storage of data and accompanying representation information. The requirements in these sections correspond to the identically named OAIS functional entities.

As with previous requirements sections, there are a core set of requirements relating to data management and archival storage that are fundamental to all three domains. Each domain required facilities to define data management policies and specified high-level requirements for the management and maintenance of digital objects and representation information once ingested. These included a variety of data management processes such as replication, deletion and integrity checking. There was also consensus across all three domains that disaster recovery processes and recovery from backups were essential features for a long-term archive, and that testing workflows should be defined for such recovery processes.

Each of the three domains also recognised a need for a long-term archive to provide facilities for updating both the ingested digital objects and accompanying representation information, and for enabling the import of new representation information.

In addition to the requirements that were identified in each of the domains, there were also a number of further requirements that were only identified in the first domain. These included the specification of further data management policy details, such as the requirement that the long-term archive must be able to mirror its content over geographically disparate locations, and the suggestion that monitoring services for the underlying hardware and software would be required. DOF1 also suggested that data management policies for capturing and tracing the preservation processes that have been applied to digital objects would be required and that facilities should be established to enable the auditing of both the content of the archive and the processes carried out within the archive. The first domain also noted that it should be possible to define trigger events within the archival storage, for example when the copyright of a digital object has expired.

The first two domains also noted that a long-term archive should offer multiple levels of storage and enable the definition of multiple storage sections. Once defined and implemented it should then be possible to move data and representation information between such sections, depending on the data management and security policies. Also identified in both DOF1 and DOF2 was the need to connect to or integrate with external systems in order to support the management of data. DOF1 noted that connections to file format registries would be required to facilitate the identification and validation of data formats and DOF2 stated that classification systems and ontologies would be required, and would need ongoing management and updates.

Preservation Planning Requirements

This section includes requirements relating to the detection of digital objects that are facing technical obsolescence and the processing of such digital objects once identified. Such requirements correspond to the 'Preservation Planning' OAIS functional entity.

There was almost universal consensus across the three domains about the preservation planning facilities that would be required. Each domain required facilities that could detect the technical obsolescence of data and representation information formats. Similarly, each domain specified requirements for the definition, management and maintenance of preservation plans and workflows.

In addition to the above requirements, the first domain also noted an additional requirement concerning any modifications made to digital objects or representation information as the result of the execution a preservation plan. DOF1 desired that the originally ingested versions of the digital objects and the representation information should never be overwritten as a result of a preservation action; if new versions are to be generated then these should be stored within the long-term archive and linked to the pre-existing versions.

Access Requirements

Requirements in this section relate to the querying and retrieval of digital objects and representation information from the long-term archive. These requirements correspond to the identically named OAIS functional entity.

The access requirements specified by each of the three domains were all broadly comparable. Unlike some previous sections, there were no major differences between the access requirements specified. Each domain specified requirements to enable the querying of the representation information and the retrieval of both the data and the representation information, with restriction to be placed on querying and retrieval rights depending on user account configuration. In addition, all three domains noted that interfaces would be required to enable remote access to the data and representation information, and also that future users of the archive may require data to be retrieved in different formats, a process that may have to be handled in real-time.

Post-Access Requirements

This section includes requirements that relate to the future use of digital objects after they have been retrieved from the long-term archive. Requirements in this section would generally be considered beyond the scope of the OAIS functional model.

Across the three domains of focus the requirements identified in this section generally relate to two particular aspects. Firstly those requirements that relate to ensuring the usage rights of the data continue to be upheld and enforced once the digital objects are retrieved from the long-term archive, and secondly those requirements that relate to enhancing the digital objects with knowledge in order to ensure they remain understandable by the designated community.

Implementing the requirements

An implementation of the SHAMAN framework will be developed in the various Research and Technology Driven (RTD) workpackages of the SHAMAN project. The requirements described in the above sections will be adopted and further elaborated upon by these workpackages in order to guide the subsequent phases of analysis, design and implementation. As mentioned previously, high-level user requirements relating to a component will be assigned to one or more workpackages and these will then elaborate more specific software requirements from this starting point.

It should be noted that the above requirement sections are not uniquely paired with specific RTD workpackages and in many cases the same requirement will need to be satisfied by components that will be developed in more than one SHAMAN workpackage. However, it is possible to identify sections that will be of particular interest to the activities of certain parts of the project. It may be observed that the Data Management and Archive Storage requirements will be of particular importance to the workpackage that deals with data grid implementation. The requirements contained within the Access sections will of primary interest to the workpackage that is investigating the management of shared collections, and also to the workpackage that is dealing with Multivalent [9] preservation interfaces and media engines. Additionally, it is expected that workpackages that are looking into harmonisation, basic analysis and ingest, and advanced information extraction and knowledge engineering will be particularly interested in the Pre-Ingest and Ingest requirements.

Conclusion

Planning and developing a software framework or reference architecture that can be broadly applicable yet meets the needs of a number of specifically targeted domains is an ambitious undertaking. For it to be a success the process must involve several iterations of requirements and continued verification that the direction taken during development is not at a tangent to user expectation.

The approach taken by the SHAMAN project, as described in this paper, was to initially pinpoint high-level user requirements for the framework that the project intends to develop, and for these user requirements to then be extrapolated as the project progresses in order to formulate more concrete software requirements.

The methodology adopted in order to elicit the user requirements has ensured that a representative range of organisations and user roles across the three domains of focus could be targeted. The process of interviews, user scenarios, use case and finally requirements definition has resulted in both a generic set of universally applicable high-level requirements for a preservation framework and domain-specific sets for each of the three targeted domains. From each of these sets it is possible to trace the evolution of the high-level requirement back through the requirements elicitation process to pinpoint exactly where the basis for a requirement was first formed.

The set of high-level requirements discussed in this document represent not only a crucial starting point for the SHAMAN project; they also represent the basis of a validation framework for the project outcomes. The subsequent work carried out by SHAMAN will be informed by the user requirements and these user requirements will form a checklist against which the applicability of the outcomes can be evaluated. By presenting the SHAMAN user requirements analysis and implementation, we expect to contribute to the work of other preservation system development efforts and offer the community a working example of a comprehensive and refined methodology.

Acknowledgments

SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) Integrated Project is co-funded by the European Union (Grant Agreement No. ICT-216736).

References

[1] Innocenti, P. Aitken, B. Hasan, A. Ludwig, J. Maciuvite, E. Barateiro, J. Antunes, G. Mois, M. Jäschke, G. Pempe, W. Wilson, T. Hundsdoerfer, A. Krandstedt, A. Ross, S. 2009. SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol, SHAMAN Project. http://shaman-ip.eu/shaman/sites/default/files/SHAMAN_D1.2_Requirements%20analysis%20report_0.pdf

[2] Thibodeau, K. 2007. The Electronic Records Archives Program at the National Archives and Records Administration. First Monday, Volume 12, Number 7. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1922/1804

[3] Knight, S. 2009. Early learnings from the national library of New Zealand's National Digital Heritage Archive project, IFLA. http://www.ifla.org/files/hq/papers/ifla75/146-knight-en.pdf

[4] NARA ERA Documentation. http://www.archives.gov/era/about/documentation.html#requirements

[5] Rational Software White Paper. 2001. Rational Unified Process, Best Practices for Software Development Teams. http://www.ibm.com/developerworks/rational/library/content/03July/1000/1251/1251_bestpractices_TP026B.pdf

[6] Consultative Committee for Space Data Systems (CCSDS). 2002. Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0b1.pdf

[7] Software Engineers Standards Committee of the IEEE Computer Society. 1998. IEEE Recommended Practice for Software Requirements Specifications. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=720574&isnumber=15571

[8] Aitken, B. Helwig, P. Jackson, A. Lindley, A. Nicchiarelli, E. Ross, S. 2008. The Planets Testbed: Science for Digital Preservation. Code4lib, Issue 3. http://journal.code4lib.org/articles/83

[9] Multivalent. http://multivalent.sourceforge.net/

Author Biographies

Brian Aitken joined the HATII in 2001 as a Systems Developer. He has worked as sole developer and as leader of a development team on a number of successful projects, most recently EU-funded Planets, for which he is leading the development of the Testbed. In the EU-funded SHAMAN he has undertaken requirements definition. Previously he has managed and developed online tools and content management systems for the Digital Curation Centre, DigitalPreservationEurope and DigiCULT, and for several successful digitisation projects.

Perla Innocenti is co-Principal Investigator in the EU-funded projects SHAMAN and DL.org, and Research Associate at HATII, University of Glasgow. Her research interests include digital preservation methodologies, preservation of media art, audit and risk assessment for digital repositories, digital libraries design and usage models, digitization methodologies. She was involved in repository design, audit research as part of DigitalPreservationEurope (DPE) and Digital Curation Center (DCC), co-ordinating activities and development for the DRAMBORA) Toolkit, usage models research within the EU-funded project Planets, and the investigation of the potential application of the DRAMBORA toolkit in digital libraries within the DELOS project.

Seamus Ross is Dean and Professor, Faculty of Information, University of Toronto. Formerly, he was Professor of Humanities Informatics and Digital Curation and Founding Director of HATII (Humanities Advanced Technology and Information Institute) (1997-2009) at the University of Glasgow. He served as Associate Director of the Digital Curation Centre (2004-9) in the UK, and was Principal Director of ERPANET and DigitalPreservationEurope (DPE) and a co-principal investigator such projects as the DELOS Digital Libraries Network of Excellence and Planets. He recommends Digital Preservation and Nuclear Disaster: An Animation [video] (2009) and "Digital Archaeology" [PDF] (1999).

Dr. Leo Konstantelos is a Preservation Resources Officer in the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. He has conducted research into preservation of interactive and ephemeral digital content for the Planets project, and into a methodology for software validation for the SHAMAN Integrated Project. Leo holds a PhD in Humanities Computing in the area of user studies for Digital Art in Digital Libraries. He has delivered a number of seminars on digital libraries, user studies, statistical methods and the digital arts. He was a member of the DELOS Network of Excellence on Digital Libraries.