Demonstrating the Next Generation Preservation Framework

Thomas Wollschlaeger, Attila Zabos, Denise E. Kiefer; Deutsche Nationalbibliothek (DNB); Frankfurt am Main, Germany

Abstract

The goal of the SHAMAN project is the development of fundamental concepts and technologies for the next generation Digital Preservation network system. The German National Library and its partners have created a demonstrator that showcases the distributed instances of the envisaged archival concept. The demonstrator is aligned in accordance with the new archive-centric information lifecycle model and connects new innovative technology of the international partners, especially of notable universities and technology enterprises, to foster the distributed ingest into an enhanced preservation framework and to adapt services that are used to manage, preserve, enrich and access digital data.

The paper will outline the developmental goals and focus especially on the development work of the demonstrator. By highlighting these aspects, the innovative approach of SHAMAN and the objective to provide a next generation Digital Preservation framework will become visible.

Introduction

Since 1913 the German National Library (DNB), acting as a legal deposit library, has archived over 25.4 million hardcopy media. The DNB acts as a library but also as a legal deposit, hence it is obliged to collect German related publications. Due to the legal obligation of national libraries, they - in most cases - must also archive *digital* book-like publications nowadays. Accordingly, in 2006 the German legal deposit law („Gesetz über die Deutsche Nationalbibliothek“) has been expanded to include all kinds of born-digital materials. That includes e-journals, online academic publications and e-newsletters as well as Web sites.

To maintain a responsible care for their digital collections that have been grown considerably over the past year, universities, libraries and other memory institutions need suitable and consistent archives for the electronic materials. The DNB in particular has the task to provide unimpeachable long-term access to all of its electronic documents and collections. To fulfill that task, the DNB has been actively engaged in several projects and initiatives regarding the long-term preservation (LTP) of digital content, on a national basis as well as on international levels. On a national level, the project *nestor*[1] (Network of Expertise in Long-term Storage of Digital Resources) was to create a competence network of long-term archival storage and long-term availability of digital resources in Germany, with the DNB in charge. The aim of the competence network is to establish an infrastructure that enables long-term archival, protection and utilisation of digital resources published in Germany. Through national and international co-operation, a contribution was achieved towards safeguarding our global cultural heritage.

The objective of the *kopal* [2] project was the practical testing and implementation of a co-operatively created and operated long-term archive system for digital content of the DNB and other partners. The technical realisation of the functions in kopal is based on prior work of a project that was carried out at the Koninklijke Bibliotheek (Royal Dutch Library). It was further developed towards a co-operatively maintained system equipped with standardised interfaces within the framework of kopal. The ingest & retrieval software produced by the network partners (koLibRI) has the status of Open Source software. Comprehensive and heterogeneous data have been fed into the system during the term of the project in order to prove the concept's capacity and usefulness. Since mid-2007, the archival system has been transferred by the DNB into production state.

The SHAMAN Project

On the basis of its archival system, the DNB decided to take an active part in the SHAMAN integrated project. SHAMAN is an EU-funded research project with partners in business, science and memory institutions [3]. Its goal is the development of a conceptual and technical base of a next generation Digital Preservation network system. Based upon the Open Archival Information System (OAIS) [4] reference model, the aim is to create an open and extensible framework. This framework defines all components, services, interfaces and specifications in the context of uniform and comprehensive long-term preservation standards in a way that facilitates their reuse. Furthermore, a distributed archiving infrastructure using GRID technologies will be set up. SHAMAN develops concepts, technologies and services which are then evaluated prototypically in test environments and practical scenarios. SHAMAN will feature three prototype applications. The validation of the SHAMAN framework is focused on scientific publications in libraries and documents in government collections, industrial design and engineering-related digital objects, plus data sources from eScience applications. Within the consortium, the DNB took on the responsibility for managing and completing the "Document Production, Archival, Access and Reuse in the Context of Memory Institutions for Scientific and Governmental Collections" work package. Here it oversees the development work on the archive prototype.

SHAMAN Information Lifecycle

For the development of a SHAMAN software demonstrator the “Archive Centric Information Lifecycle” (see *Figure 1*) has been used as a basis. This lifecycle model was developed by the project partners [5] and it describes the lifecycle of electronic publications from the SHAMAN point of view, considering and utilizing several technologies and software components.

The model illustrates exemplarily the five distributed instances of the SHAMAN archival concept. The various instances, also called as phases of the lifecycle, are:

- Creation
- Assembly
- Archival
- Adoption
- Reuse

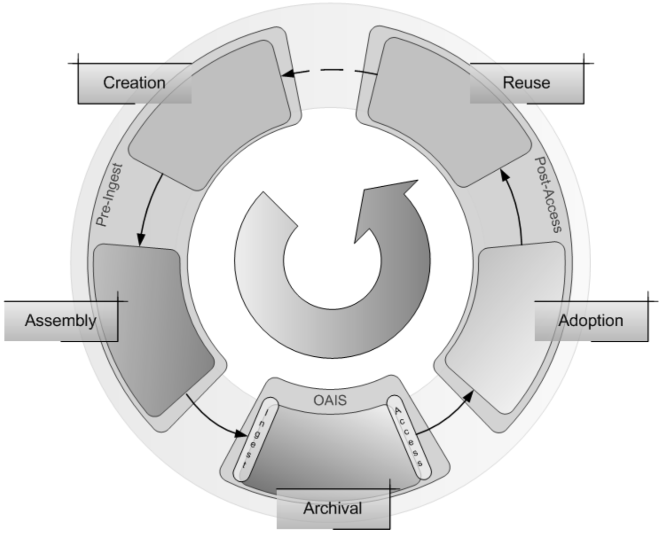

***Figure 1**: SHAMAN Information Lifecycle model*

The cycle starts with the *creation* of digital objects. It is the initial phase during which new information comes into existence. In the best case, the producer also delivers context information (e.g. details about production environment or databases linked to the digital object). The generation of context sensitive metadata allows a better identification of the materials and simplifies their reuse.

Creation processes can be rather complex and involve a multitude of stakeholders, with only chunks of the resulting information worth considering for archiving. Usually, digital objects are not created for the purpose of archiving. Their shape aims at the use within regular business models of the producer. Therefore, the term *use* means the exploitation of information according to the original purpose the object was created for. Traditionally, objects are archived right after *creation*. From the perspective of the archive, *use* is a concurrent thread in the life of a digital object that also starts with completion of *creation*.

The *assembly* phase denotes the appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future, meeting the presumed needs of the designated community. *Assembly* requires in-depth knowledge about the designated community in order to determine objects relevant for long-term preservation together with information about the object required for identification and reuse some time later in the future.

Archival addresses the life-time of the object inside the archive. In most archives, policies prohibit irreversible deletion of content. Hence, preservation is a perpetual activity and the *archival* phase is open-ended - unless digital objects are to be irrevocably removed from the archive. Archives do not only preserve objects to end in itself but also have to ensure that objects are being provided for future use. Basically, information disseminated by the archive must enable the designated community to use that information.

The *adoption* phase encompasses all processes by which accessed archival packages are unpacked, examined, adapted, transformed, integrated and displayed to be usable and understandable by the consumer. This includes also emulation activities if needed. *Adoption* might be regarded as a mediation phase, comprising transformations, aggregations, contextualisations, and other processes required for re-purposing data. Additional information beyond that provided by the archival package could be used to assist in this phase.

Reuse means the exploitation of information by the Consumer. In particular, re-use may be for purposes other than those for which the digital object was originally created. Reuse of digital objects can lead to the *creation* of other, novel digital objects. *Reuse* also may initiate the addition or updating of metadata about the digital object held in the archive. For example, annotation changes information content and affects the relationships existing between the object and other digital objects. In collaborative working environments, there is a continuous flow between access and ingest, in that retrieved digital objects are reused and/or modified, resulting in new revisions and additional (composite) digital objects which have to be preserved, along with their provenance information.

SHAMAN Demonstration Scenarios

It is expected that digital documents will be published at a rate far higher than is the case with hardcopy equivalents. Thus, under the responsibility of the DNB, the appropriate methods required for the digital archiving of book-like publications are chosen for exploration within the first scenario where the DNB is responsible. The other two scenarios focus on digitizations and heterogeneous interlinked materials.

SHAMAN Scenario “Archiving of book-like publications“

This scenario (the key scenario for the DNB) describes the workflow involved in the indexing and archiving process of book-like publications in *Depot Libraries* (DLs). It provides the fundamental requirements for the implementation of a demonstrator for ISP1. The workflow depicted in *Figure 2*, shows the processes as they appear in the context of a national library.

The workflow begins with either the reception of a digital object that is intended to be archived in a long-term preservation system, or with a request to access already stored objects, i.e. based on web-forms implying publisher authentication and submission, or OAI harvesting. The main process of the DL, denoted as *Preservation*, consists of *Ingest*, *Retention* and *Access* processes.

In the first case, commercial and non-commercial *Publishers* deliver digital objects for archiving to the DL. After a successful authentication, the Publisher can submit the digital object as *Submission Information Package* (SIP). Within the *Ingest* process, the SIP is imported, verified and enriched with additional metadata. The metadata contains technical information (e.g. about images, table of content, etc.) but also links to authority files and bibliographic information in the catalogue. The last step of the Ingest process is the generation of an *Archival Information Package* (AIP) that contains the delivered digital object and the additional metadata. Finally, the AIP is transferred to the archive where it will be stored and migrated during the course of time as part of the recurring *Retention* process.

In the second case, the *End-User* sends a request to the DL in order to access certain archives. Note: the *Access Request* process does not include the search but only a dedicated request for a known archived object. The *Access Request* process receives the *End-User's* query and delivers him the requested object from the long-term preservation system. *Working* with the retrieved objects means that the end-users can examine their content in a multivalent client. This client provides the technology to render various digital objects, but also to carry out operations that were initially not planned for the archived object.

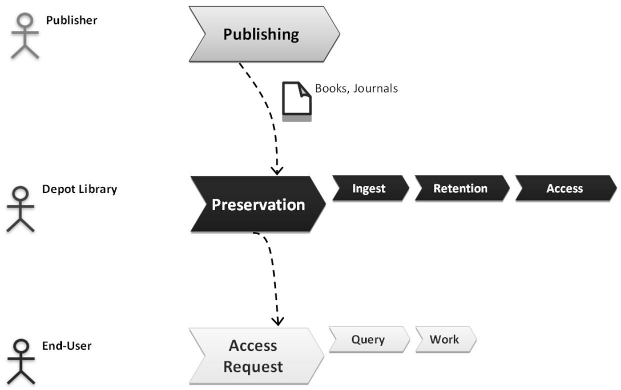

***Figure 2**: Book-like publication workflow in the depot library domain*

DNB test collection

The test collection used in this scenario consists of a large set (containing more than 30,000) of digital PhD theses complemented by metadata in METS files. The DNB has been collecting theses and dissertations (ETD's) since 1997. Their total number exceeds 85,000 at present and forms the largest ETD collection in Europe. The ETD's are delivered from German universities via the DNB portal, leading to a widespread access and use of these ETD's.

Corresponding to the scientific value of these documents and their excellent availability for scientific research, the ETD's form the most used and most respected digital collection of the DNB. Therefore, it is the major task of the DNB to preserve that collection for long-term use. The major challenge for their long-term preservation arises from the fact that the German ETD's are delivered in numerous file formats. The first of the older file types are beginning to become hard to access due to the disappearance of suitable viewer applications. It has therefore become imperative to transfer the German ETD's into a suitable long-term archive. If the ETD's at the DNB were successfully long-term preserved and their future accessibility ensured, it would be likely that other digital publications could be subjected to the long-term preservation efforts as well. These considerations led to the use of the ETD's as pilot materials for the DNB's long-term preservation efforts within the SHAMAN project.

In the legacy system of the DNB, the archived theses contain links to various other data (see *Figure 3*). Apart from the extracted abstract section and appendices, this data contains *publication context* and *technical metadata*. Furthermore, the publication context consists of information about the deliverer of the digital object, title, subject, publisher, creator, etc. The technical metadata is comprised of information about the digital object itself, e.g. content type, size and checksum for validation purposes. Finally, digital PhD theses are linked to authority files that contain information about persons/authors, corporate bodies, places, URIs, etc.

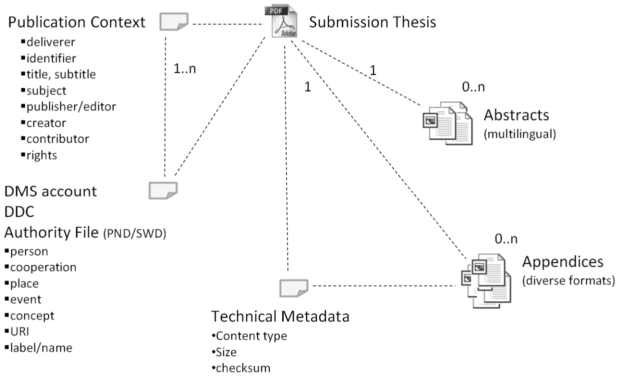

***Figure 3**: Book-like publications demonstrator business objects*

Scenario "Indexing and Archiving Digitisations" [6]

The second scenario focuses on digitized materials. Several memory institutions have set up automatic digitisation facilities, notably the State and University Library of Goettingen (SUB) and the Bavarian State Library in Germany. Since these facilities have advanced technologically in recent years, it is now a challenge to align a high volume of output with the scale of the operation, in terms of size and the number of files generated.

The scenario is being pursued by the SHAMAN partner SUB Goettingen. In the past, at the SUB digitization facility, i.e. at the Center of Retrospective Digitization in Goettingen, the production volume of digitisations was approximately 150,000 pages per year. The size of each page in uncompressed TIFF format was around 25 Mbytes. For the year 2010 it is estimated that the production volume will increase to 200,000 pages, resulting in 10 TBytes of digitised material per month.

Therefore, the scenario will address not only the problem of archiving large volumes of data but also the searching within and the migration of digitised materials, these being important activities for long-term preservation. Within the scenario, the workflow can be triggered by several events. Following an initial selection, the materials are scanned, the scans post-processed and optimized. After a quality check, metadata are added and, if feasible, an OCR process initiated. These steps are subsumed under the term "production phase". Finally, after the production phase, the content is transferred to the archival process.

Scenario "Scientific Publishing and Archiving Heterogeneous Interlinked Material" [7]

Scientific congresses remain an integral part of work and discourse in the field of science. Since they produce and publish numerous content, many activities are common to their own processes and those of other scientific journal and book publishers. They create and produce a broad variety of content during their discourses - for example, abstracts and proceedings, presentation slides, posters, audio-visual materials and reviews. Their workflows are quite typical of those to be found within scientific discourse and publishing more broadly.

The third scenario accordingly focuses on capturing the production context of congress publications that support the scientific publication process. Within the SHAMAN consortium, the GLOBIT is responsible for supervising that scenario. The scenario workflow starts with a research and practise phase where scientific content originates and is being submitted. During the following congress phase, the content is being processed, submissions are added to the conference, further added by talks, programs and the like. Conference Websites support communication and interlink several materials. After the congress, publication processes occur. Finally, all contributions are subject to a long-term preservation process.

SHAMAN Demonstrator

The SHAMAN demonstrator shall showcase a specific implementation of the flexible and adaptable SHAMAN framework for long-term preservation in the context of memory institutions.

Motivated by the scenarios and workflows of the DNB and the other partners, a SHAMAN demonstrator integrating and evaluating new technologies is under development. Details about the current status of the demonstrator are presented in this section.

Software Architecture

The development of the demonstrator is guided by an iterative process that ensures the production of a functional software prototype with increasing complexity during the course of the project. The needs of the stakeholders in the domain of memory institutions build the fundamental basis for the software architecture. These needs were specified in the description of the scenarios and can be summarized for a future LTP system as:

1. support the archival of the expected high volume of digital data in near future,
2. provide a collaborative set of services for long-term preservation using the service-oriented architecture paradigm, supporting existing business processes in various memory institutions,
3. support for flexible and extendable data storage to accommodate the increasing volume of digital objects,
4. automated metadata generation and retrieval for the long-term preservation, especially context information about the production environment to support rendering or migration of the archived objects.

The software architecture (see *Figure 4*) addressing the aforementioned needs is presented in the following and describes the structure of the first SHAMAN demonstrator. The software architecture is composed of three major components:

1. Application server. This component hosts the web applications and services that provide the interface between the user and the LTP system.
2. Data manipulation, search and retrieval component provides functions to ingest, alter and discover object in an archive.
3. Data storage performs the tasks of an archive for digital objects.

To advance the development of the SHAMAN demonstrator, certain existing software components were integrated into the demonstrator. The reuse of components provided the required functionality for the demonstrator to showcase the support for the workflows eminent in memory institutions. In particular, Glassfish, Cheshire3 and iRODS were used as application server, data manipulation and retrieval, and data storage components, respectively.

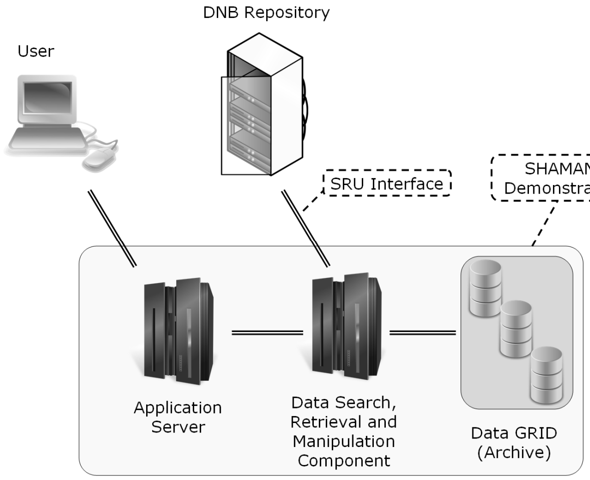

***Figure 4**: Software Architecture of the SHAMAN Demonstrator*

Software Prototype

Considering the envisaged software architecture, the fundamental components and services of the SHAMAN demonstrator were implemented, resulting in a first functional software prototype.

The initial feature set of the software prototype covers the functionality of the system that is required in the domain of memory institutions. However, following the iterative development process, the prototype will be continuously developed and successively extended.

Demonstration

In order to demonstrate the functionality of the SHAMAN prototype, a test collection consisting of digital PhD theses was provided by the DNB. This test collection was considered as a representative set of digital objects that need to be archived by memory institutions. The test collection contained the digital object (usually either as a PDF file or a set of HTML files) and an associated METS files with basic information.

As a data storage for archival, the iRODS data grid was setup and utilized. In iRODS the stored data is organized in logical collections that have similarities to regular operating system directories or folders. In an initial step, the ETD's were uploaded into a collection that was created as a temporary location to store data before ingest.

Before the digital objects were archived in the dedicated iRODS collection for long-term preservation, their basic metadata was enhanced with various information. On one hand, additional structural metadata was created and on the other hand, bibliographic data was fetched from the DNB's repository and stored within the objects' METS files. As part of the automated ingest process, these two steps enhanced each digital object's metadata with information valuable for long-term preservation.

Additionally to the digital PhD theses, scanned documents provided by SUB Goettingen were also archived. These objects were available as TIFF images.

From the LTP perspective, two different approaches were showcased demonstrating the re-use of digital objects. On one hand, using multivalent technology the SHAMAN demonstrator showed the potential of a dedicated browser to render archived documents. The browser not only renders various document formats but also provides features that are not available with legacy document viewers (e.g. fast reader, text-to-speech synthesis, OCR in the case of scanned documents, etc.). On the other hand, the scanned documents available as TIFF images were used to demonstrate the migration process. In the case that the maintenance of legacy document viewers is discontinued, it is still possible to migrate the affected archived objects to a newer more up-to-date format.

In summary, we provided an overview of a future long-term preservation system that has been under development within the SHAMAN project. The motivation and requirements of memory institutions for such an archive system and an overview of the current state of a functional prototype were presented.

Acknowledgements

This paper was funded in part by the SHAMAN Integrated Project (Grant Agreement No. ICT-216736). We would also like to thank the SHAMAN partners for their cooperation on the software demonstrator.

References

[1] nestor Website: http://www.langzeitarchivierung.de/

[2] kopal Website: http://kopal.langzeitarchivierung.de/

[3] Project partners are INMARK, University of Liverpool, InConTec, The Swedish School of Library and Information Science (SSLIS) at Göteborg University and University College of Boras, Xerox Research Centre Europe, FernUniversität Hagen, Philips Innovation Lab, University of Strathclyde, Goettingen State and University Library, Otto-von-Guericke University Magdeburg, Industrious Media, GLOBIT, Humanities Advanced Technology and Information Institute (HATII) of the University of Glasgow, INESC-ID.

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

[5] The model has been developed within the task of „Implementation of context capturing mechanisms within production environments and federated server architecture for temporary storage before ingests", namely by the SHAMAN partners DNB, University of Liverpool, GLOBIT, SUB Goettingen, FernUniversität Hagen, University of Strathclyde, Xerox, HATII, INESC-ID and Industrious Media.

[6] Scenario short description based on the detailed preparation of Jens Ludwig from Goettingen State and University Library written in the context of a SHAMAN project deliverable.

[7] Scenario short description based on the detailed preparation of Gerald Jäschke from GLOBIT written in the context of a SHAMAN project deliverable.

Author Biography

Dr. Thomas Wollschlaeger is an information specialist and scientific librarian, working at the Department of Information Technology at the German National Library in Frankfurt am Main. He has worked on several projects within the context of long-term preservation of digital data, notably DissertationOnline from 2003, and the kopal LTP project from 2005 to 2007. Since then, he is in charge of the IT project management group at the DNB.

Attila Zabos has a degree in Computer Science (2001) from the University of Applied Sciences in Darmstadt, Germany. He has worked for more than 5 years in the industry on the development and documentation of embedded real-time and safety-critical systems, and from 2007 for more than 2,5 years at the University of York on an EU-funded research project in the area of real-time systems. Currently, he is with the German National Library in Frankfurt am Main and involved in the development of new technologies for long-term preservation systems.

Denise E. Kiefer received her M. A. in book science from the Johannes Gutenberg University Mainz, Germany (2010). Since 2007 she worked at the German National Library in Frankfurt am Main at the Office for Library Standards and since 2009 she is working at the Information Technology department and is primarily concerned with long-term preservation, particularly in the SHAMAN project.