Embedding Legacy Environments into A Grid-Based Preservation Infrastructure

Claus-Peter Klas, Holger Brocks, Lars Müller, Matthias Hemmje

FernUniversität in Hagen

Universitätsstrasse 1

58097 Hagen, Germany

{Claus-Peter.Klas, Holger.Brocks Lars.Müller, Matthias.Hemmje}@FernUni-Hagen.de

Abstract

The SHAMAN project targets a framework integrating advances in the data grid, digital library, and persistent archival communities in order to archive a long-term preservation environment. Within the project we identified several challenges for digital preservation in the area of memory institutions, where already existing systems start to struggle with e.g. complex or many small objects. In order to overcome these, we propose a grid based framework for digital preservation. In this paper we describe the main objectives of the project SHAMAN and the identified challenges for such a heterogeneous and distributed environment. We on the one hand assess in a bottom-up approach the capabilities and interfaces of legacy systems and on the other hand derive requirements based on the project's objectives. Our investigation is focused to the integration of storage infrastructures and distributed data management. In the end we derive a service-oriented architecture with a grid-based integration layer as an initial approach to manage the challenges.

The SHAMAN Project

As part of the European Commission's 7th Framework Program, the SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) project targets a framework integrating advances in the data grid, digital library, and persistent archival communities in order to attain a long-term preservation environment which may be used to manage the ingest, storage, preservation, access, presentation, and reuse of potentially any digital object over time. Based on this framework, the project will provide application-oriented solutions across a range of sectors, including those of digital libraries and archives, design and engineering, as well as scientific data and information management.

The SHAMAN project will integrate the automated handling of technology evolution with data analysis and representation mechanisms in a way which will uniquely enable multiple user communities to preserve and reuse data objects, in whatever format, which are deposited in the preservation environment.

The project will furthermore provide a vision and rationale to support a comprehensive *Theory of Preservation* that may be utilized to store and access potentially any type of data, based on the integration of digital library, persistent archive, knowledge representation, and data management technologies. In addition SHAMAN will supply an infrastructure that will provide expertise and support for users requiring the preservation and re-use of data over long-term periods of time. Within this infrastructure the project will also develop and implement a grid-based production system that will support the virtualization of data and services across scientific, design and engineering, document, and media domains. Finally three Integration and Demonstration Subprojects (ISPs) supporting the *Memory Institutions* (ISP-1), *Design and Engineering* (ISP-2) and *eScience* (ISP-2) domains are used to analyze their ecology of functional (and non-functional) requirements and to identify a core set of relevant digital preservation usage scenarios. These ISPs foster the systematic integration and evolution of project results towards the targeted SHAMAN framework and its prototypical application solutions, i.e. they drive the horizontal integration of RTD contributions.

In this paper, we will focus on ISP-1 Document Production, Archival, Access and Reuse in the Context of Memory Institutions for Scientific Publications and Governmental Document Collections, which trials and validates the SHAMAN approach along the business purposes of scientific publishing, libraries, and parliamentary archives.

We will present challenges which are derived from the preliminary results of the top-down requirement analyzes of ISP-1. From the bottom-up technology perspective we have conducted an initial assessment of the capabilities and interfaces of the systems employed inside and outside the SHAMAN consortium which hold relevant digital collections, but also solutions supporting access (i.e. searching and browsing), resource discovery and collection management. We will then elaborate on the specific technological challenges of integrating heterogeneous storage infrastructures and distributed data management and present a first conceptual integration-approach based on a grid-based integration layer and service-oriented architectures for resolving these issues.

Integration Requirements

The goal of this paper is to describe digital preservation legacy technology and existing application solutions as well as a draft integration concept for embedding such legacy environments into an overall preservation infrastructure like the SHAMAN framework. To evaluate this integration concept we need to provide an assessment scheme which represents general digital preservation requirements, but also specific challenges derived from integration of individual, complex systems and processes within the SHAMAN context. The following generic integration requirements represent overall conceptual goals or success criteria for the SHAMAN framework, which are refined and complemented by more specific challenges from the ISPs:

• Integrity - The main goal of preservation environments is to maintain the persistence of digital objects. Integrity refers to maintaining their completeness and immutability. A preservation environment has to provide adequate measures for maintaining the integrity of its digital objects.

• Authenticity - Authenticity corresponds to the genuineness of an object. An object is considered as genuine if certain properties can be attested which confirm its identity. A preservation environment must prevent unauthorized manipulation of its objects in order to guarantee their authenticity.

• Search & Browse - Besides safe-keeping its digital objects, a preservation environment also needs to provide access to its collections. This requires persistent identifiers and sophisticated search methods to find and access particular objects.

• Interpretability - Technological advancements leads to the aging of digital object formats. The careful selection of allowed formats according to various criteria enables the long-term interpretability of the content of digital objects. Furthermore, preservation environments need to support strategies for dealing with technological obsolescence.

• Virtualization - The integration of distributed information systems requires coherent management of the heterogeneous systems and collections. A federated preservation environment needs to abstract from the idiosyncrasies of its constituting peers, while maintaining full control over processes and objects, including their significant properties.

Following these general requirements the next section describes the specific scenarios for memory institution in ISP-1.

ISP1 - Memory Institutions

Within SHAMAN's ISP-1 scenario we need to provide long term preservation for three memory institutions (2 libraries, 1 archive), the German National Library (DNB), the Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) and the Flemish Parliament (FP) governmental archive. All of them are running existing individual solutions. As a grid-based data management system iRODS will be assessed and trialed as the successor of earlier SRB technologies, which already is used in Europe and the US for very large file repositories. We will evaluate its appropriateness for virtualizing the storage layer of the SHAMAN preservation framework architecture, also with respect to its capabilities for integrating existing legacy systems with proprietary data schemas. The initial goal of the iRODs assessment, is to embed the existing repositories and archival systems from the DNB, SUB and FP as an active node within the iRODS data grid. In the following sections we will describe the legacy systems and discuss possible solutions based on existing tools and grid systems.

Existing Storage and Long-Term Preservation Systems

The SHAMAN application scenarios require the integration of various types of existing and upcoming systems. Examples of these systems are institutional repositories like Fedora and DSpace, the KOPAL long-term digital information archive, standard database storage systems as well as access support systems such as DAFFODIL or Cheshire. These systems have to be assessed individually, but also the resulting composite infrastructures have to be evaluated according to the challenges described above. We will discuss the above named institutional repositories, archive systems and access systems closer in the next sections. The grid-based systems under evaluation follow the legacy system description.

Institutional Repositories

Institutional Repositories are used for managing documents and collections within scholarly environments, such as universities and libraries. As production systems they need

to be integrated in a transparent way, without impeding or compromising their primary functions.

Current global players on institutional repositories are Fedora and DSpace.

Fedora

Fedora (Flexible Extensible Digital Object and Repository Architecture) represents a repository enabling archival, retrieval and administration of digital objects and their metadata via web services. It is developed at the Cornell University and the University of Virginia. Within Fedora a digital object is a container for different components. These are a unique identifier, descriptive metadata, data streams, and disseminators. Each container consists of at least one data stream including metadata in Dublin Core format. A data stream can also be a URL. An object can also contain disseminators connected to a data stream to generate dynamically different views, e.g. a black/white picture of a color picture.

Fedora also supports integrity via checksums and authenticity of digital objects. Redundancy is based on replication on a second Fedora system. Archiving, retrieval, and administration is based on SOAP and REST web services. To access the metadata an OAI-PMH server is integrated to provide access to other systems. The system is OAIS conform and supports ingestion of SIPs (digital objects with METS data) objects.

There are currently 127 Fedora projects and 25.000+ downloads have been counted last year according to the Fedora Wiki.

DSpace

DSpace is like Fedora an institutional repository developed by Hewlett-Packard and the Massachusetts Institute of Technologies as open source project. Objects (items) are stored in collections structured by communities and sub communities.

Each item represents an archived object, including metadata and further files like thumbnails of the original picture. Here also checksums are used to check the integrity of stored objects. Metadata is supported via Dublin Core and other formats can be transformed. An OAI-PMH supports access of metadata, so DSpace can be used as data provider. Objects can be stored in the local file system or via SRB / iRODS data grid technology.

Search and browse functionality is provided by a web interface and DSpace uses persistent identifiers.

Currently DSpace exists in 324 installations in 54 countries with approx. 2.561.082 Documents according to the DSpace Wiki.

Archival Systems

Long-term archival systems are complex IT systems with idiosyncratic processes and information structures. With their ability to provide bit-stream preservation functions at various service levels, archival systems will be embedded as specialized storage nodes which offer higher levels of data security.

A running long-term archival system is operated at the German national library, called KOPAL. As central archival library and national bibliographic center for the Federal Republic of Germany the German National Library DNB has to collect and archive also all electronic publications appearing in Germany since 2006. To comply with this assignment the DNB builds up in co-operation with other national and international memory institutions an IT-infrastructure for archiving and long-term preservation of digital objects. In its current state this infrastructure consists of a repository system for collecting digital objects, bibliographically preparing them and allowing access for external users. All objects are then archived in a back-end archival system for long-term preservation.

This long-term archival system was developed cooperatively with the Niedersächsische Staats- und Universitätsbibliothek Göttingen, the Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) and IBM Germany within the project KOPAL (2004 - 2007). The technical realization is based on prior work accomplished since 2000 in a joint development project of the Koninklijke Bibliotheek (Royal Dutch Library) and IBM.

The core component of the system is the DIAS archive developed by IBM. DIAS implements core components of the Reference Model for an Open Archival Information System OAIS. Hosted at the GWDG at Göttingen this multi-client capable system provides independent access for each partner from any location via well defined interfaces. DIAS itself consists of standard applications by IBM, DB2, WebSphere, and the Tivoli Storage Manager.

The project partners of KOPAL implemented supplementary open source software on top of the DIAS-Core, the so-called kopal Library for Retrieval and Ingest (koLibRI). Realized in Java KoLibRi provides tools for automating archiving tasks like ingest and access of digital objects in a flexible and modular manner.

Currently the KOPAL archival system is transferred into the productive use by the German National Library. It will be integral part of a more complex repository and archival system which cover the whole process from data collection via data preparation, data archiving, data access to data presentation. This repository system itself is integrated in the library system with its several tasks and services.

For the communication with the outside world there are several interfaces provided or in preparation including web forms, SRU and services based on OAI (Open Archival Initiative), especially the OAI-PMH protocol. With these interfaces the foundations are complied to integrate the DNB repository system into subordinated infrastructure networks as it is planned in the integrated project SHAMAN and to provide and exchange data and metadata within these networks.

The Flemish Parliament document storage consists of several classical databases, which can be searched via a web interface. We have to investigate their preservation proprietary solution.

Access Support Systems

Content-based access support represents a fundamental requirement for SHAMAN, in addition to traditional metadata-based search and browsing functions. The main challenge within a federated environment is to keep the retrieval index consistent and up-to-date.

Cheshire

Within SHAMAN we plan to integrate Cheshire, a full-text information retrieval system based on a fast XML search engine. On the basis of indexes it provides access to the essential search and browse functionality of digital libraries. Cheshire's development started 10 years ago at these UC Berkley and currently is run in version 3 by the University of Liverpool. It supports several protocols like Z39.50, SRW/SRU or OAI-PMH for access of metadata.

DAFFODIL

To provide users with the ability to find and access their preserved information we will utilize the DAFFODIL system .

DAFFODIL is a virtual digital library system targeted at strategic support of users during the information seeking and retrieval process (see [Fuhr et al.2002] and [Klas2007]). It provides basic and high-level search functions for exploring and managing digital library objects including metadata annotations over a federation of heterogeneous digital libraries. For structuring the functionality, we employ the concept of high-level search activities for strategic support and in this way provide functionality beyond today's digital libraries. A comprehensive evaluation revealed that the system supported most of the information seeking and retrieval aspects needed for scientists' daily work. It provides a feature rich and user-friendly Java swing interface to give access to all functionalities. Furthermore a Web 2.0 browser interface enables the main but not all functions of the Java interface for easy access. Besides the main functionality of federated search and browse in distributed and heterogeneous data sources and a personal library, further functionalities like browsing co-author networks, thesauri, conference & journal browser and collaborative functions are already implemented and can be directly used. Through a wrapper toolkit DAFFODIL can access SRU/SRW, Z39.50 and OAI data sources. Besides them access to any web based digital library is possible. DAFFODIL currently support access to the domain of computer science and can be used under http://www.daffodil.de.

Establishing Data Grids for SHAMAN

The goal of SHAMAN is to setup a preservation solution based on data-grid technology. Distributed data-grid technology is used to manage and administer replicated copies of digital objects over time. Data-grid middleware will be used as core data-management technology, mediating between SHAMAN components and legacy systems. Such systems are SRB and iRODS, which we will take under close evaluation, since they are widely used systems.

SRB

The Storage Resource Broker (SRB) is a grid middleware, developed at the San Diego Super Computer Center as commercial product. SRB enables integration and transparent use of different geographically distributed storage systems. A user accessing a digital object is not aware of the current location. A SRB system consists of several zones. A zone itself is represented by an arbitrary number of SRB servers and a central database, called MCAT. A SRB server can manage several storage systems (resources). Besides metadata, the MCAT also stores information about the zones, locations and resources. Clients can access via any SRB server all objects in a zone. The query is automatically routed by the MCAT. Archived objects can be structured in collections and sub collections. Another important aspect is the fact that collections can contain objects from geographically distributed sides in one logical view. Around SRB exist several so called drivers which enable access to other storage systems, e.g. GridFTP in both directions, to access SRB storage from GridFTP and vice versa. Also DSpace can integrate SRB as a storage space. SRB based collections currently hold more than 150 million files worth > 1000 TB of data.

iRODS

iRODS, the integrated rule-oriented data system, is the open-source successor of SRB, also developed by the San Diego Super Computer Center. iRODS contains the same functionality as SRB but, as new feature, introduces a rule engine. Such rules follow the event-condition-action paradigm and run on the iRODS servers together with so called micro services. Micro services can be implemented and integrated via a plug-in feature in iRODS, so there are no limitations on functionality and extensibility. Examples for such micro services are to create a copy of an ingested object or check an object for integrity based on checksums.

Furthermore micro services can then be connected to more complex rules, which can follow again events and conditions.

iRODS is already on the way to being used as a preservation system. In some institutions it is also currently in the migration process, where they change the system from SRB to iRODS.

Requirements for SHAMAN's ISP-1 Integration Concept

In order to design a first integration approach of the identified legacy systems in SHAMAN we will analyze an initial example application-scenario provided by DNB. The scenario is as follows: A memory institution uses a specific long-term preservation system for physical printed books and journals. This system is not intended to be replaced by a new system. But the system is not well equipped for handling digital objects, like web pages, which by law have to be preserved, too. The idea is now to extend the preservation solution through new technologies like a grid-based system, in order to cope with the amount of stored digital information objects. The legacy system will remain to be in use as main preservation system, but access, ingest, and management should be possible in parallel through the grid system with one interface and a set of appropriate internal workflow processes.

Within this scenario, we were able to identify a first set of three initial use cases to integrate the existing system with our grid-based system. These use cases are central access on distributed repositories, central storage on distributed archiving and central management on distributed collections.

In order to analyze, discuss, model, and later implement the three use cases, we need to specify an integration architecture. Therefore, as a first exercise towards building the overall framework's reference architecture, a specific integration architecture for ISP-1 has to be derived. This will be a starting point for the extension and abstraction of this architecture into a more general framework architecture that can serve all three ISPs and in the ideal case support many other future system developments for other application domains and scenarios as a development and deployment framework for DP application solutions.

Such an initial architecture for ISP-1 needs to fulfill certain requirements. One set of requirement is provided explicitly in the SHAMAN project plan. The SHAMAN project requires to establish a very dynamic framework for the development of a stable and reliable preservation environment which is strongly driven by supporting infrastructure independence, the ability to preserve digital entities as a collection, and the ability to migrate the collection to new choices of storage and database technologies.

In addition, the current status of requirements that are driven by all the described legacy systems is as follows:

- The data will be stored in distributed repositories. If customers integrate their systems with new technology data will continue to be stored in distributed repositories.

- The repositories are running on different legacy systems. Therefore, future distributed repositories to be established should still be able to run on different legacy systems, like DSpace, Fedora, KOPAL, or traditional database systems, too.

- The legacy systems provide different protocols. Each future system should be able to provide different searching and browsing protocols, too, as well as different protocols and processes for ingestion.

- The legacy systems use different metadata standards. Therefore, a future system should be able to support these metadata standards, such as Dublin Core, METS, MARC or LMER, too.

Utilizing Service Orientation in the Integration Architecture

The above described requirements make it necessary to use a service-oriented architecture (SOA) because it provides the following features:

Modularity - The upcoming system need to be modular to integrate each legacy system.

Standard - The system needs to standardize the protocols and metadata formats.

Independence of technology - The preservation process should not rely on any technology; it should rather be able to easily adopt new technology for better performance.

Flexibility - Each part of the system should be easily replaceable or adaptable to new needs and future technology.

Reuse - Already existing service should be reusable in other context.

In short, we need to setup a service-oriented architecture in order to provide a modern, agile, flexible and dynamic system to optimize all processes within a long-term preservation environment. Existing services can be reused and new features can be adopted and integrated without disturbing running processes. In this way it will be possible to manage such a feature-rich, complex, and dynamically evolving set of tasks as preservation solutions are faced with. This service-oriented draft of an integration architecture for ISP-1 will be a starting point or the extension and abstraction of this architecture into a more general framework architecture for the whole SHAMAN project.

In the following sections each use case is depicted by a four layered service-oriented architecture. The lowest level *Preservation/Storage Systems* holds all legacy systems as well as the grid-based repositories. The *Wrapper* level enables standardized access to the underlying systems. The *Service* level combines functionalities which represent the workflows and processes necessary to run a preservation system. On the top level the *User and Management Interface* provide users and administrators with access to the system.

Information Integration based on a Mediator Approach

In a distributed environment we need to search and browse several distributed repositories in order to support user queries. If the environment consists of more than one legacy system, a mediator or wrapper is necessary, if it is not possible to directly integrate a legacy system into the grid system. This is e.g. the case with the displayed iRODS driver for the system DSpace.

A multi-layered architecture for such an iRODS driver case

Figure 1: System Integration with iRODS Wrapper

is depicted in *Figure 1*. The legacy systems are located on the lowest level. Via iRODS wrappers/drivers we gain full access on the bases of the iRODS protocol to serve the search and browse queries. The service can rely on defined protocols and propagate the query and gather the results to be presented via the user interface within DAFFODIL. Through these mediator levels, users gain transparent read access to any legacy system.

The (read-only) search and browse process can be described the following way:

1. The user interface of DAFFODIL relies on a specific Search and Browse service and passes any query via the communication platform of the SOA to these services.

2. The service connects to the iRODS MCAT server and if a) a central search index exists, runs the query central b) a distributed search index exists; the query is passed to each repository and performed locally

Figure 2: System Integration with General Wrapper

3. The resulting objects will be accessed by the iRODS driver from the legacy system, e.g. KOPAL's knowledge base and passed through the services to the user.

During this process the syntactical heterogeneity of the metadata is captured on the wrapper level, whereas the semantic heterogeneity of the different search and browse interfaces is captured on the service level.

If we do not want to rely on iRODS only as long-term preservation storage system, it is also possible to abstract from it by implementing general wrappers to access any legacy or grid-based system. The above described process still holds also for this case, but the difference is, that each wrapper has to implement the search and browse functionality formerly provided by the iRODS driver as depicted in *Figure 2*.

As stated in a previous section, a first prototype implementation of this scenario can be completely based on the existing DAFFODIL framework utilizing at the same time a service-oriented architectural approach. On the lowest level we need to implement wrappers for the DNB, SUB and the FP. If they exist, the search and browse functionality is ready to be evaluated. The search service already combines results from distributed heterogeneous data sources and the user interface presents the result directly with query term highlighting, sorting and filtering. It is out-of-the-box possible to store found result in the personal library and many other already existing high-level functions can be used. Within [Klas2007] it was also proven, that the DAFFODIL system raises efficiency and effectiveness of the user during the search and browse process over any other search system.

Distributed Ingestion

Even if the archives of the DNB, SUB or FP are integrated into the grid based system archiving of new objects still takes place in the local repositories. However, the grid system needs to be aware of changes in the local repositories in order to support search and browse functions. The situation is depicted in figure 3. To overcome this problem three solutions can be discussed:

Figure 3: Distributed Ingestion

• Local ingestion and a redundant grid ingestion: Here the local system runs their ingestion process and after success runs the ingestion on the grid system. This is the least complex integration.

• Local ingestion and notification to grid server: The second case ingests also to the local repository, but either sends out a notification message, that a new object was ingested to the grid system or the grid system polls on schedule for new objects in the local repository, e.g. OAI harvesting could be used.

• Grid ingestion and triggered local ingestion: In the third case, the more un-trusted case by the local repository owner, the object is ingested in the grid system and then locally ingested.

In any of the above cases it has to be discussed where the real object is stored. Either it is stored in the local repository and only the metadata information is published on the grid, or the object itself is replicated on the grid. On the management layer the repository manager has to be always aware that the ingestion process was correct and that the integrity and authenticity of the objects is guaranteed.

The quality of this service is different to the search and browse case, since we need write access and access to the access rights management.

Managing Distributed Collections

Besides the integration of information and the distributed ingestion process, managing the distributed collections in the heterogeneous grid environment with all legacy systems is another important challenge. The management is necessary to cope with formulation and implementing of policies, prioritizing and planning, assessing risks or calculating expenses.

The use case here, holding several copies of an object on distributed repositories in order to avoid loss through e.g. a disaster, could be implemented through a replication service. Based on risk calculations and worth of the digital objects, the user states the requirements, to have three copies of the objects in distributed sites.

Whereas in the cases above, the wrapper and services need only to be aware of their local environment, in this case a complete new mediator level needs to be aware of all repositories to find another repository which meets the requirements to replicate the object at that side.

Figure 4: Managing Distributed Collections

In F*igure 4* the management tool within DAFFODIL initiates the replication process, after the replication policies for the user are changed.

1. The task “replicate the object x from resource KOPAL to resource DSpace” is handed to the replication service.

2. The service checks the DSpace repository, if it is available, has enough free space, etc.

3. The service initiates the copy process, which of course contains verification processes, e.g. via MD5.

4. Both repositories have then to indicate if the copy process is completed and correct which is visualized in the management interface. This indication is also logged for legal issues.

The management tool on the interface level will become a master control station in order to monitor processes, policies and archive requirements.

Outlook

Combining the first architectural models from the above three use cases, we can derive a multi-layer conceptual system model based on a service-oriented architecture, as depicted in *Figure 5*.

On the lowest level the preservation and storage systems are located. The main system in the SHAMAN context will be a grid based system. All the existing functionality of this system will be verified and reused. In case of heterogeneity problems, either of syntactical or semantic nature will be handled on the wrapper and service layer. The wrapper layer integrates and enables access to the storage systems. The service layer then supports all necessary functionality not provided by the storage systems. On the top level the user/administration/management interface relies on the lower level to visualize the complex functionalities.

Each functionality is represented by a specific communication protocol and described as a set of services with input/output parameters within the SHAMAN service oriented architecture. In figure 5 the three protocols search and browse, ingestion and management are related to ISP-1, whereas Design and Engineering as well as eScience are related to ISP-2 and ISP-3 within SHAMAN, where the necessary protocols still have to be identified.

The SHAMAN goal to define a The Theory of Preservation can be supported on this conceptual level and we will aim to prove its assumptions based on these services and protocols. The services and protocols define the SHAMAN system and in order to run a future system, a service provider only needs to be compliant to the service description and protocols. Legacy systems need to fulfill only a minimal set of services and protocols in order to be integrated or migrated into our SHAMAN system or they need to have open interfaces to be wrapped, if a customer wants to use their proven system. In order to be compliant with the SHAMAN framework other systems need to implement the necessary SHAMAN services and protocols.

Figure 5: Multi-Layer Model on Service-Oriented Architecture

Summary and Next Steps

In this paper we have described the SHAMAN project, its aims and challenges. Within the ISP-1 we identified the need to incorporate legacy systems, since some customers will not necessarily change their local running preservation environment, but need to extend and integrate new technologies to scope with future requirements. In three realistic use cases we identified challenges that we have to meet. In order to enable these we propose a sophisticated service-oriented architecture based on a multi-layer conceptual model. Doing so, we meet the above stated requirements of modularity, standards and independence from technology. Furthermore this will enable SHAMAN's demonstrators to become independent of any future preservation system, but to fulfill the needs of its users to preserve important information. Going on from ISP-1 to the whole SHAMAN project with the other domains of Design and Engineering as well as eScience we will investigate their needs in order to integrate their requirements and extent, adopt, remodel, and verify this architecture. Furthermore, the next steps to setup and evaluate the grid based SHAMAN system will be:

1. Enabling search and browse functionality on the repositories of the DNB, the SUB and the FP based on the DAFFODIL system. We will reuse existing wrapper and service implementations from previous projects where e.g. The European Library and DNB were projects partners.

2. Integration of institutional repository software DSpace and Fedora, the preservation system KOPAL and iRODS on the wrapper level as storage systems within the DAFFODIL framework in order to implement the graphical management tools and services for the ingestion process

3. Model, implement and setup management functionalities for policy processes as addressed e.g. in third use case for replication.

The best practices gained from these implementations will be evaluated and form impact on the SHAMAN overall conceptual model.

Acknowledgments

Special thanks goes to Jose Borbinha, Jürgen Kett, Alfred Kranstedt, Adil Hasan and the SHAMAN consortium for the discussions and comments. This paper is supported by the European Union in the 7th Framework within the IP SHAMAN.

References

Fuhr, N.; Klas, C.-P.; Schaefer, A.; Mutschke, P. 2002. *Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries.* In Research and Advanced Technology for Digital Libraries. 6th European Conference, ECDL 2002, 597–612. Springer.

Klas, C.-P. 2007. *Strategic Support during the Information Search Process in Digital Libraries.* Ph.D. Dissertation, University of Duisburg-Essen.