http://conference.ifla.org/ifla78 Date submitted: 24 May 2012

Libraries as service-brokers for digital data curation: Practical insights from the DFG project DP4lib (Digital preservation for libraries)

Reinhard Altenhöner
Deutsche Nationalbibliothek / Director IT
Frankfurt am Main, Germany
E-mail: r.altenhoener@dnb.de

Session: **116 — *The role of libraries in data curation, access and preservation: an international perspective* — Science and Technology Libraries**

Abstract:

Introducing digital curation / preservation services into the portfolio of libraries requires a lot of general process know-how, technology and investment. Out-of-the-box solutions for the technology offered by some market participants are hopefully easy to implement technically, but the challenge is to establish a sufficient and durable customer service based on them. Supported by the German Research Foundation (DFG) the project Digital Preservation for libraries (DP4lib) pursued the goal of a cooperative service and organizational model for the provisional long-term archiving - understood as an organizational and financial framework to share costs and efforts for digital curation / preservation. In order to allow libraries and other participating institutions a systematic exchange of know-how and valuable services the project therefore processes mutual cooperation relations as an exchange of functional performance and working results into a service model. In addition to the technical implementation of the project it was necessary to develop a comprehensive and pragmatic cost model for service provision between participating organizations. The project is implemented as a generic process model for network-based services: On the one hand technical requirements and tool development to ensure quality controlled hand-over of digital objects including dedicated metadata-sets, on the other hand accounting procedures and processes, which allow partners to define the demanded quality level and correspond the costs and the financial consequences they have to expect.

1. In 2006 the legal mandate to collect and to archive digital publications was directed to the German National Library (In German: Deutsche Nationalbibliothek, DNB), exposed as a revised version of the law stated for the national library in 1969. Archiving means **long-term** archiving expressed by the capacity to preserve the availability and accessibility of digital objects. This causes a set of requirements with institution-specific challenges, which have to be addressed seriously: Based on the new law, which states in its essence that the library is responsible to collect and preserve digital publications in general, the library was obliged to enhance its capacity and to change its workflow organization within a relatively short period of time. These different tracks of experiences encouraged the library to initiate the project DP4lib[1] and to invest further efforts to improve the achievements so far extended to other institutions.

But this was not the starting point for the library: 2006, when the new law for the German National Library was adopted[2], the library could ground the needed actions on a lot of preparatory work made in the years before. Beginning in the 90ties, the library had started preliminary, project-based initiatives in order to extend know-how and to implement dedicated new services and specific routines to collect special types of digital objects.[3]

Although not specifically targeted to digital preservation demands, this helped to understand the needs and specific requirements for the collection of digital objects. Most of the solutions created in these days were specific in the sense that they were implemented for specific type of objects, e.g. a solution to harvest newsletters or a dedicated project to ingest the production especially of e-journals of the German publisher Springer. In another project, a co-operational network to encourage and collect e-theses and dissertations was implemented. It resulted in a lot of experience in the area of metadata, data enrichment and workflow development. A growing demand was recognized for the persistent identification of digital objects in a digital environment. Therefore DNB initiated a URN-based persistent identifier service that resolves today more than 7 million unique object identifiers. So the infrastructure in general seems to have grown steadily with the requirements.

On closer inspection, however, the fact that many of the applications were implemented separately and differently for specific classes of objects has proved to be a problem. What does this mean practically? Several of these individually established workflows were not integrated on an organizational and technological level, which means that they worked – manually maintained - successfully for a small number of objects, but they were not designed to process large numbers in

[1] Please refer to the project homepage (mostly in German) http://dp4lib.langzeitarchivierung.de/. Some additional information in English can be found here: http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-One-Pager-08-eng.pdf.

[2] The law is available only in German: http://www.gesetze-im-internet.de/dnbg/index.html. [checked 2012-02-25] The mandate covers both Internet publications with printed versions and corresponding such as e-books and e-journals, and web-specific media works such as Web pages. Even the collection of music files is in-scope

[3] Reinhard Altenhöner: "The German National Library and digital preservation: challenges and opportunities." In: World Digital Libraries, Volume 3, Issue 1, June 2010. pp. 1-9

automated processes and they do not allow the incorporation of more employees without specific IT-skills. In addition another feature of these tools results in difficulties: Special attention had been paid to the question of how the objects were accessed and this leads to some different and ‘competing’ interface-solutions. Another important aspect was entirely absent: the validation and technical analysis of the objects on their format and data integrity was missing in the ingest routines.

In essence, DNB had in place largely independently created software solutions that encompassed a broad range of tasks: the collection of different types of objects, validation, storage and presentation of objects for the user.

When the law went into force it was clear, however, that this implementation approach could not be adequate for the future: In particular, the large amount of objects and the attendant need to integrate existing staff in large numbers led to the realization that a fundamentally new approach was needed. Associated with it, it also became clear that the maintainability of countless individual routines quickly pushes its limits and does not provide sufficient guarantees for the safe operation. This, however, affected not only the development of IT-based services themselves, but especially the new and improved working procedures of DNB.

2. All these activities were started to address the legal obligation to collect, to index and to offer access to the digital objects. But the challenge to address the need of digital preservation was unanswered yet. So in 2004 the development of a long-term archiving system was begun, leading in 2006 / 2007 to an extended prototype solution called kopal[4]. The development was based on a commercial asset, developed by IBM and founded on standard software, called DIAS (Digital Information Archive System)[5]. On top of it a specific metadata handling was implemented with specific focus on technical metadata including dedicated methods for object integrity and controlling. This system was defined for a set of types of digital objects ready to become ingested into the long-term-archiving system. Of course the objects had to be validated whether they were technically fit and logically consistent. This happened in dedicated routines in the pre-ingest area, addressed in an Open Source Library called koLibRI (kopal library for Retrieval and Ingest). Except for this Java-based software library, which could be changed and adapted to different usage scenarios of DNB, the DIAS solution is a "black box" in the sense that IBM has the complete responsibility for further development of the software, for change management and error handling as far as this relates to the software. In addition, first steps have been undertaken to obtain data from the archive for the planning and implementation of policies on migration in practice.

[4] kopal means “Kooperatives Archivsystem für die Langzeitarchivierung digitaler Objekte”, in English: Co-operative Development of a Long-Term Digital Information Archive. Cf. http://kopal.langzeitarchivierung.de/index.php.en [checked 2012-02-25] . Further: Reinhard Altenhöner / Tobias Steinke: "Kopal: cooperation, innovation and services: Digital preservation activities at the German National Library", Library Hi Tech, Vol. 28 (2010) Issue 2, pp.235 – 244.

[5] See DIAS specification, http://kopal.langzeitarchivierung.de/index_downloads.php.de.

After these developments the situation was as follows:

On the one hand a few different ingest routines were available for dedicated classes of objects, independent from each other and difficult to maintain. On the other hand there was the long-term solution as an isolated approach, which has to be coupled to the ingest workflow. An example may illustrate the situation:

By transferring the different routines in an enhanced practice it came out that the existing workflow routines to collect objects from the original producer or publisher were not prepared to verify the technical quality of digital objects. Existing workflow-routines to handle the digital objects were implemented as an independent workflow within the digital-preservation system and they started after the identification and physical collection of objects. For technical reasons it was impossible to move functions into the ingest routines easily. Facing the challenge to process big amounts of data and on the other hand to optimize facilities to handle different (and new) types of objects and easy ways to configure and enhance workflow routines, it came out that DNB had to start a new development process in order to review the workflow. This happened in the following years: A completely new infrastructure of generic ingest routines (including ticket handling and process engine) was set up. On the same hand the digital preservation issue should be addressed.

From its beginning the development of a digital preservation infrastructure was driven by the idea that cooperation with other partners helps to save money and resources. Beside of the technical development (here specifically the tool-library koLibRI was implemented as an open source solution, available in the net, commonly developed with other partners) the network of expertise “nestor”, a platform for exchange, know-how transfer, working and standardizing activities, which joins in addition to libraries archives, museums and other cultural and scientific organizations, was established.[6] All members know that many of the tasks needed in the future to enable long-term archiving, such as risk management for file formats, or the implementation of preservation actions need different types of expertise, which can be found in different organizations. So the task to organize practical collaboration came on the table.

Two mayor needs could be identified in the perspective of the DNB:

a. Scaling and advancing the flexibility of ingest routines and practical integration of the digital preservation into the workflow-organization of DNB.

b. Extension of practical cooperation in order to share results and to save resources.

3. Two main initiatives were launched: On the one hand a dedicated concentration on technical workflow development, on the other hand the extension of digital preservation to other partners in order to share experience and to offer customizable and flexible services.

[6] http://www.langzeitarchivierung.de/Subsites/nestor/EN/Home/home_node.html

a) Technical workflow development

The implementation of automated routines is founded on three basic requirements:

- Use of standardized metadata formats for the specification and verification of electronic resources in the catalog or search system

- Definition of quality levels for file formats from the perspective of digital preservation

- Definition of transfer interfaces to receive the objects and the metadata from the producers

Besides the creation of metadata and object quality management policy, the transfer of objects and metadata into the DNB has been addressed: Currently DNB provides three interfaces for delivery: a web form for single objects and two automated methods, one to push and one to pull objects. The push method uses a delivery account (called “hot folder”), the transfer is handled by using SFTP or a WebDAV interface. Each delivery package is a single transfer container, in which both the object (optionally also composed of many files) and an associated set of metadata are zipped. The pull method is based on the OAI Protocol for Metadata Harvesting in combination with a transfer URL submitted within the metadata.

The whole process is documented in written form and in addition by using BPMN (Business Process Model and Notation) graphs, which helps, to visualize the workflow and clarify the need for technical components.

In addition, some steps towards a more systematic approach regarding the collection of online publications were taken: The organization was established in an own inter-departmental task force. The work was focused on working with institutions, not on the collection of individual objects. So aggregators and service-providers became involved and this group has been extended to software vendors, who provide publishers with management systems. The software development itself was defined in a specific development environment and focused on performance and high throughput.

b) Digital preservation for libraries (DP4lib)

4. The aim of the project (funded by the German Research Foundation (DFG), lasting from 2010 - 2012) with a total of 8 partners was the organizational and technical extension of the named kopal solution to an integrated service that will help to render the varying needs of partners in the preservation of digital data on a unified technology base. The partners had different roles: While individual institutions are willing to act only as users and rely entirely on the service-offering partners other partners are willing to take an active role and become responsible for dedicated tasks. This had to be balanced and integrated into an accounting and cost model and organizational framework. The technical

infrastructure, the bit-stream level and the software-service backend layer was outsourced to another partner. So the project was able to focus the cooperation needs between different service providers establishing and to develop a cooperative service and organizational model for the provisional long-term archiving. The basic idea was to customize and share digital preservation services in a cooperative environment. In practice a set of well defined and formalized relations between different units are necessary to enable the exchange of services functionally integrated into a service model. In addition to the technical implementation within the project, there was the need to develop process structures and a cost model as an economic basis for service provision between organizations.

In order to create a generic solution that can be implemented in many heterogeneous environments and integrated as a part of the working policy of cultural heritage organizations, an open concept with modularized service packages was targeted. So the following goals were defined:

- Creation of a flexible long-term preservation infrastructure adapted to the needs of (smaller) cultural heritage organizations and their service providers
- Technical enhancement of the existing solution (kopal), conforming to the partners' requirements
- Implementation of a reusable process model and preparation of a handbook to introduce long-term preservation in (smaller) cultural heritage organizations

W.r.t. the adoption and integration of workflows the example of DNB was used to implement a generic approach built by single technical modules, reusable by other partners, but even integrated into the IT-infrastructure of DNB.

The selection of the partners took account of their first objective, the interest and expertise of the institutions involved themselves in various ways but even in cooperation in the field of digital preservation. This expertise should be brought together and at the same time bundled in a way that the goal to bring in life a process model for the establishment of such cooperative structures can be trained and stipulate cross-important needs and issues. This means that it was also about the creation of concrete, practical handouts, which relates not only to technical and functional requirements of long-term access, but also to their operational and organizational design.

Therefore, at the beginning a systematic analysis of requirements was done, which initially identified the existing needs in the various institutions and departments, and stood at the end of an overview of the existing materials and quantities to internal policies and requirements for long-term preservation and access scenarios to objects. Early in the project it turned out that issues of quality management and documentation as well as the aspect of risk management, although there are not primarily technical aspects, are of great importance. This proved to be a good guideline for future work in the project, which revolved around the question of the organization of long-term preservation services.

The analysis itself was an extensive questionnaire that helps to formulate the basic positions and self-referencing. So each organization has to clarify, which collections they want to keep permanently, what they can work out in advance etc.

In a further step the collection of requirements was summarized and compressed in order to get clear and comprehensive stating about the requested features. Some of the documented functional requirements contain provisions which go beyond the projects in the long-term archiving area and their orientation on purely technical features. The way here was more process-oriented on maintenance, or termination of service, on continuity management; here are some examples:

"If a transfer package has been transmitted through the mass delivery interface, the system must provide to be capable of delivering the institution of a (still closer to be specified) report on the success or failure (e.g. transcription errors, disconnections, etc.) of data transfer."

"The system should be capable of the validity of the documents delivered in respect of: file format, corruption, password protection and further still to be specified protection mechanisms (such as write protection, copy protection) in order to check and report back to the supplying institution success or failure of the respected archive process."

"If a migration of the original digital object or its current version migrated earlier has been performed, the system must be able to deliver both the whole object history and specific objects to the associated institution."

"Termination of service: The system must be capable of delivering the institution to provide an exit strategy available that can be passed with the help of which the delivering institution is able to rehearse upon the working lives of documents including the migrated versions again."

Other stipulations included, for example, the definition of a list format, data volumes, metadata, administrative data, transfer packages, interfaces, exceptions / error handling etc..

A very important aspect in addition to the definition of working items was the introduction of a quality management system and the establishment of requirements for this particular in the field of documentation and reporting. Based on this preliminary work, business processes (workflows) become specified and modeled. In essence core processes like ingest, access and preservation planning were identified (not really surprising if one considers OAIS) and above this level an additional granular segmentation into sub-processes took place. This helps to enable technical developments. At the same time, various reporting levels were defined and implemented in the project.

These definitions should not be understood as a rigid construct, but as a starting point to analyze the needs and to adapt changes. This means continuous review and evaluation of the whole process.

5. Technically the enhancements were largely based on the existing infrastructure (see above), which was either incurred in the kopal project or are present as a part of the infrastructure for ingest in DNB: For example, on the one hand an existing SFTP server (hot folder) was expanded, on the other hand technical processing was directed to an existing OAI-PMH interface. For the workflow itself it must be ensured that the guaranteed level of integrity of digital objects in the ingest process can be held at any time on the defined height. And secondly, a risk assessment is recommended during the recording of digital objects that can serve as a preparatory measure, among other things, for the preservation planning.

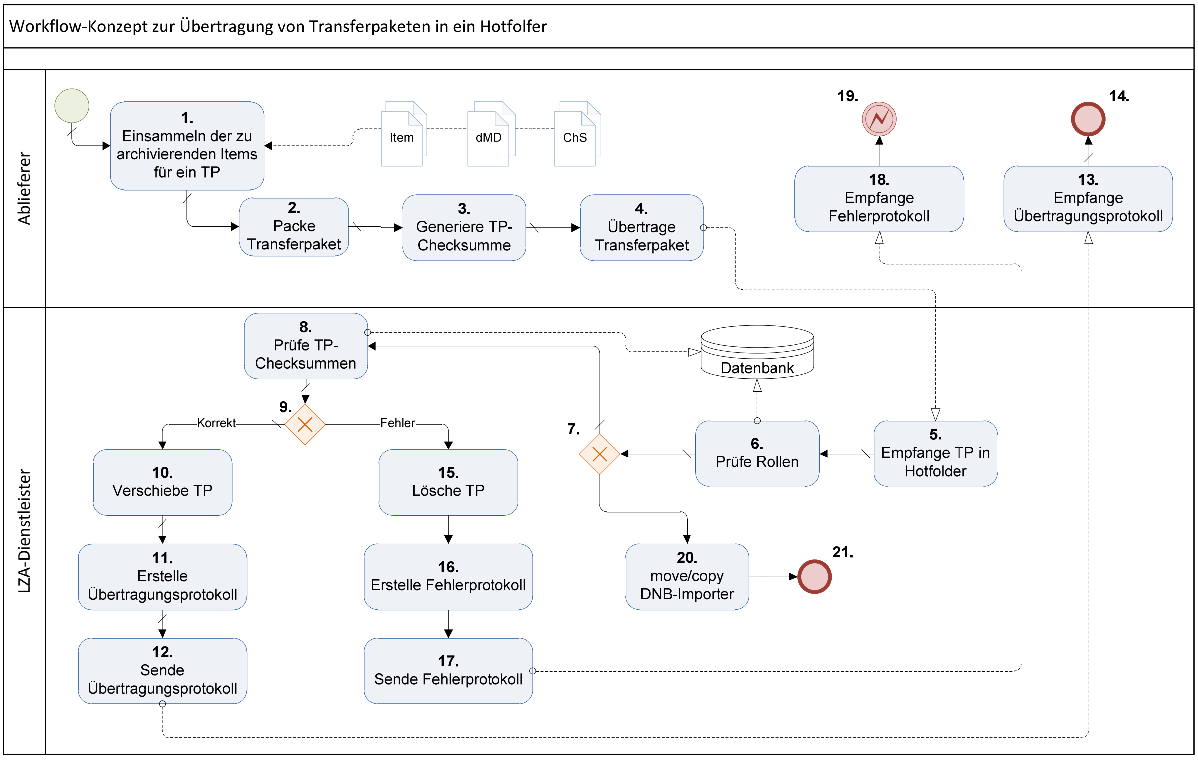
Workflow-Konzept zur Übertragung von Transferpaketen in ein Hotfolfer
Ablieferer
LZA-Dienstleister
1. Einsammeln der zu archivierenden Items für ein TP
Item
dMD
ChS
2. Packe Transferpaket
3. Generiere TP-Checksumme
4. Übertrage Transferpaket
19.
18. Empfange Fehlerprotokoll
14.
13. Empfange Übertragungsprotokoll
8. Prüfe TP-Checksummen
Datenbank
9.
Korrekt
Fehler
7.
6. Prüfe Rollen
5. Empfange TP in Hotfolder
10. Verschiebe TP
15. Lösche TP
11. Erstelle Übertragungsprotokoll
16. Erstelle Fehlerprotokoll
20. move/copy DNB-Importer
21.
12. Sende Übertragungsprotokoll
17. Sende Fehlerprotokoll

Figure: Process steps to preserve the file integrity during transfer of digital objects between service providers and service providers[7]

A crucial element in this case constitutes the continuous documentation of the steps, here concretely visible in the creation and delivery of qualified transfer

[7] Cf. Langzeitarchivierung – Ein Handlungsleitfaden für Dienstleister und Dienstnehmer (V. 1.0) March 2012 (unfortunately only in German available). http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-Handlungsleitfaden_v1.0_light.pdf Here page 24.

protocols (step 11, 12, 16, 17). These protocols result in a standardized form, which is machine-readable. They are based on a ticket system that monitors all activities related to individual packages.

Another important development was the need to carry out a risk assessment at an early stage of the process. To address both aspects, the requirement to maintain file integrity on the one hand and on the other hand the implementation of risk assessment the idea of the so-called ingest level for digital objects was designed. This actually works as a classification concept for digital objects, which derive from a multistep procedure and is in fact a statement on the certainty, we believe to guarantee for the long-term accessibility of digital objects. The corresponding statements are based on the results of the technical object validation in the review process of the service. Criteria are inter alia data integrity (Do we get the same bit-stream sent out by the supplier of data?), the identification, access restrictions (for example, protection mechanisms), the feasibility of extracting format-specific technical metadata, the format validity itself). Based on these results and the success level the ingest level is automatically rendered, which can be used in the subsequent long-term archiving process supporting a management tool.

The ingest levels follow on each other in a hierarchical way. A high ingest level means implicitly that the expectation, to preserve specific digital objects is higher. The ingest-level grading ranges are from 0 (only the secured data transfer is guaranteed in the archive) to level 4, in which the criteria are all fulfilled, e.g. “a PDF document is using the agreed checksum procedure, is checked for file integrity and it’s improved in an analysis tool as limiting free. Even the metadata generator could process and has delivered information on format characteristics transformed in technical metadata. In addition the validity of the file format could be confirmed for PDF specification."[8]

Another work package in the project was to define organizational and communication structures: This is particularly the written record of each work and service step, but also the definition of the documentation depth, the frequency and detail of reports, error reporting chains, initial specifications to policy-building for preservation planning activities. This has to be considered even for scenarios, which result from the access area: how we organize the return dubbing of all stored objects ever received by one partner, if for example a repository is to be rescheduled. And how such a process is initiated, monitored and controlled?

In a mostly technological view some communication interfaces were agreed; for example the web service for access (build as a SOAP / REST-solution). Special attention was spent to automate communication between partners (data exchange process using SFTP and OAI, for reports a PUSH interface was provided, but even a pull-interface is offered.

[8] Cf. DP4lib report, page 30.

To assure the quality of management and to keep the process transparent and understandable, a series of reports was agreed within the project: It may be on the one hand machine-readable messages, on the other hand one can create systematically reports on cross-cutting issues, such accumulations over a period of time or for merging evaluation of several partial reports, aimed at human readers then.

As a crucial component and challenging in specific way preservation planning was set up as an issue in the project: All efforts, dedicated to objects in order to collect the needed information are accumulated for preservation activities. The intention in the project was to agree on a common performance and service catalogue, which however, is based only on basic requirements in principal defined on the shoulders of OAIS and the modules described therein.

The corresponding workflow is as follows: To support the comprehensive process development appropriate areas of work has to be defined and the early provision of appropriate communication and documentation tools has to be ensured. The broadening of expertise by intensifying the working connection to nestor, the German competence network for long-term preservation is another step, processed as the establishment of a working group. In addition, process models for the identification and documentation of significant properties, which have to be developed, can be derived from the concept of the working group “Digital Preservation", which has published a paper for the systematic identification and classification of "conservation groups". This extends the OAIS model in pragmatic and reusable way.

Special attention was given as stated to the organizational framework for a digital preservation service, exposed by dedicated organizational conditions: The integration of digital preservation in the organization of an institution is a challenge itself. The approach to handle this within a collaborating group needs the design of formal relations between the partners, and specifically between service providers and employees. This includes a whole range of practical procedure stipulations: First, have an accounting procedure established for billing of services provided to others. Before an accepted service can accounted, verifiable and transparent evaluation procedures have to be evaluated in monetary terms, especially in order to address individual requirements of single institutions. In addition, it requires a contractual safeguarding to verify the relationship between the different partners, including legal liability issues and a clear allocation of roles, mandates and decision-making authority. The settlement was regarded not only as an exchange of services between partners, but also within an institution. Practically spoken the settlement follows a transparent cost model, which is continuously adapted on changing conditions and cost factors. It is planned to set off certain services to each other, so for example preparatory work in the area of metadata is included in the assessment of the customer services.

For various aspects of organizational security cooperation, the project master plans and tools are developed and integrated into the real cases. In cooperation with the partners one could therefore extend the validity of the created process forms, which are practically tested and improved. In particular, the pattern for legal texting is depending from German legal advice, but some generics are transferable.

Cost model for digital preservation service: Overall, the development of the cost model proved to be very work-intensive. In DP4lib it was pragmatically distinguished between hardware, software, personnel, premises and external services, which are then mapped to the core processes of ingest, curation and access as stated before. These were further divided into sub-processes, in detail

- Ingest: Reception of the objects, metadata handling, SIP handling, reporting and logging system, storage of the objects formed.
- Curation: digital lifecycle management, conservation, integrity check and maintenance, retrieval (search and access) (within the system).
- Access: authentication, search, providing.

Based in these assumptions the distribution model can be processed. It works but it came not yet in action with other partners, because the long-term preservation service for external customers hasn't started in an operative mode so far. But our findings from the given implementation situation in the DNB show in comparison with other similar studies, that the ratio of the cost percentage of each of the main processes (curation and ingest is very close together (34 and 36%), while the access process is only 29% of the cost).

6. Current status of DP4lib and outlook

- Technical implementation is complete, even the extension of existing workflow routines is made
- Testing with partners is successfully done, covering multiple cases w.r.t. throughput and complexity of objects
- Integration of the external storage platform is prepared.
- The organizational framework including cost modeling, contractual preparations and defined and tested controlling reports / ticket system is running
- Workflow redesign and implementation of a new processing order and management system is prepared. The move to the new system is currently prepared by staff training and exercising.

We plan to transfer the service into practice in the course of 2012, in collaboration with some of the participants in the project. In 2013 the service will be expanded to include new partners and transferred step by step into a regular service.

7. Basically one may consider that the shared use of long-term-archiving facilities by different institutions is an easy approach, because all of the

participating institutions are doing the same business. But in practice it came out that different institutions can have the same software, but their way to use the software differs a lot. In addition the sharing of services needs a clear understanding of mandates, tasks and roles within a given environment.

Our systematic approach was to generate a model to process the needs and requirements of organizations in a systematic and comprehensive approach.

As a result we can consider that the introduction of digital preservation services into the operations of institutions where existing workflow processes are established affects the process infrastructure in many ways. This leads in itself to mutual adaptation needs. If not only internal but also external organizational units are involved - and that is often unavoidable in a complex process such as the long-term preservation - the importance of a comprehensive service including an advanced organizational solution and a cost model is striking. The construction of such value-added network usually starts with the common objective to reduce costs through cooperation, to shorter development time, to benefit from experience or to distribute the risk. This is followed by the definition of the individual services, the associated roles and their delivery. In this way the start of a work-sharing operation of long-term preservation within a defined organizational setting is requested. Of course, theoretically stated needs require practical implementation. The project DP4lib is a first step to implement a service-oriented infrastructure for digital preservation services. By implementing the project we learned that much more important than technical tools is workflow-related work, and here especially quality assurance, verifiable processes, predictable costs, reliable reporting, and documentation of history.