DESIDOC Journal of Library & Information Technology, Vol. 32, No. 4, July 2012, pp. 302-305

Tailoring Digital Preservation Services for Practice: Workflow Development and Cost Modeling

Reinhard Altenhöner

c/o Deutsche Nationalbibliothek/Director IT
Adickesallee 1, D-60322 Frankfurt am Main
E-mail: r.altenhoener@dnb.de; http://www.dnb.de

ABSTRACT

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 organisational 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 organisational solution and a cost model is striking. Against this background, it is necessary to bring together technology and organisation and to anchor in a holistic approach operational aspects of digital preservation and its organisational embeddedness in institutions. Of course needs that are stated in a theoretical manner require practical implementation. Reflecting the history of workflow implementation in the German National Library some central needs and results from these experiences are presented. The new workflow for mass ingest of digital publications integrates requirements from the digital preservation unit in the broader scope of ingest routines and showcases the demand for modular and generic components, which can be integrated in the IT-infrastructure of a library. Results from the project DP4lib show in addition how a service-oriented infrastructure for digital preservation services can be established. By implementing the project was learned that much more important than technical tools is the workflow-related work, and here especially quality assurance, verifiable processes, predictable costs, reliable reporting, and documentation of history.

Keywords: Digital preservation, workflow management, process management, service management, long-term archive, national libraries, ingest

1. INTRODUCTION

In the last decade, an increasing relevance has been attributed to digital preservation in the portfolio of cultural heritage organisations. This is especially true for those organisations that have a legal mandate to archive and preserve (specific collections of) digital objects. This type of institution takes over responsibility for the long-term to guarantee the accessibility of historic material in a perspective of 50 years or more. In a cooperative information infrastructure, a specific layer as a safe place, taking on the function of a last resort in the whole landscape is needed. This causes a set of requirements for cultural heritage organisation, which often becomes responsible for these kinds of tasks.

The example of the German National Library (In German: Deutsche Nationalbibliothek=DNB) illustrates the specific challenges, which have to be addressed seriously which is ased on a new law, which states in its essence, that the library is responsible to collect and preserve digital publications, the library was obliged to enhance its capacity generally and to change its workflow organisation within a relatively short period of time.

Following the way of these developments, this paper presents some insight views and basic results from the different processes which were started to introduce digital preservation in the library workflow organisation. 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.

2. GERMAN NATIONAL LIBRARY

Understanding the situation, be prepared 2006, when the new law for the German National Library

Received 1 June 2012, online published 18 July 2012

was adopted, the library could ground the needed actions on a lot of preparatory work made in the years before.

Beginning in the 90s, the library had started preliminary, project-based initiatives to extend know-how and to implement dedicated new services and specific routines to collect special types of digital objects. 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 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 need was recognised concerning the infrastructure to identify digital objects in a persistent and unique way; so the 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.

2.1 Challangers and Solutions

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 organisational and technological level, which means that they worked successfully for small number of objects, but they were not designed to process large numbers of objects and do not allow to incorporate more employees without specific IT-skills. In addition another property of these tools caused difficulties: Special attention had been paid to the question of how the objects were accessed and this led 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, it was largely independently created software solutions that encompassed the entire range of collection of objects, from the development/description to validation of objects and their archiving. At the latest when the law went into force it was clear, however, that this implementation approach could not be promising. In particular, the large amount of objects and the attendant need to integrate existing staff in large numbers led to the realisation 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.

2.2 Long-term Archiving System

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*[1]. Kopal means 'Kooperatives Archivsystem für die Langzeitarchivierung digitaler Objekte', In English: Co-operative Development of a Long-term Digital Information Archive[2].

The development was based on a commercial asset, developed by IBM and founded on IBM standard software, called DIAS (Digital Information Archive System). On top of it a specific metadata handling was implemented with specific focus on technical metadata[4,5] including dedicated methods for object integrity and controlling. This system was defined for a defined 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)[6]. 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 implemented to obtain data from the archive for the planning and implementation of policies on migration in practice.

2.2.1 Reviewing of Archiving System

But these principal limitations were not the main issues which occurred by transferring the different routines in an enhanced practice: The existing workflow routines to collect objects from the original producer or publisher were not able to verify the technical quality of digital objects. Existing workflow-routines to handle the digital objects were as noted created as an independent workflow within the digital-preservation system and they started after the identification and physical collection of objects. 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 to review the workflow. So two core tasks could be identified:

- Scaling and flexibilisation of ingest routines
- Practical integration of the kopal-solution into the workflow-organisation of DNB.

The situation of DNB is not exceptional for cultural heritage organisations. Many archives are fighting with this kind of hurdles and this is a fundamental problem for the introduction of IT-based solutions into the workflow of libraries (or cultural heritage organisations) in general. The IT-based systems like library management solutions and even digital preservation systems arise as complete solutions including an interface to access archived materials ‘first-generation systems’. Or they arise as extensions of existing repository-solutions, where certain functions are specifically expanded. An open, modular solution that offers the opportunity for a full integration in an existing IT-infrastructure is still rare.

3. EXPERIENCES, NEEDS AND ANSWERS

At the beginning of the initiative, fundamental question had to be answered: Are there any restrictive stipulations, which have to be considered under the new law[7]? This question was answered quite simply: The law, addresses these points very open and explicitly, allows both push- and pull-routines; so the further steps weren’t restricted. 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.

Two main initiatives launched were:

(a) A dedicated concentration on technical workflow development.

(b) Extension of digital preservation to other partners in order to share experience and to offer customisable and flexible services.

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

- Use of standardised metadata formats for the specification and verification of electronic resources in the catalogue 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 in to 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 (Fig.1). 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

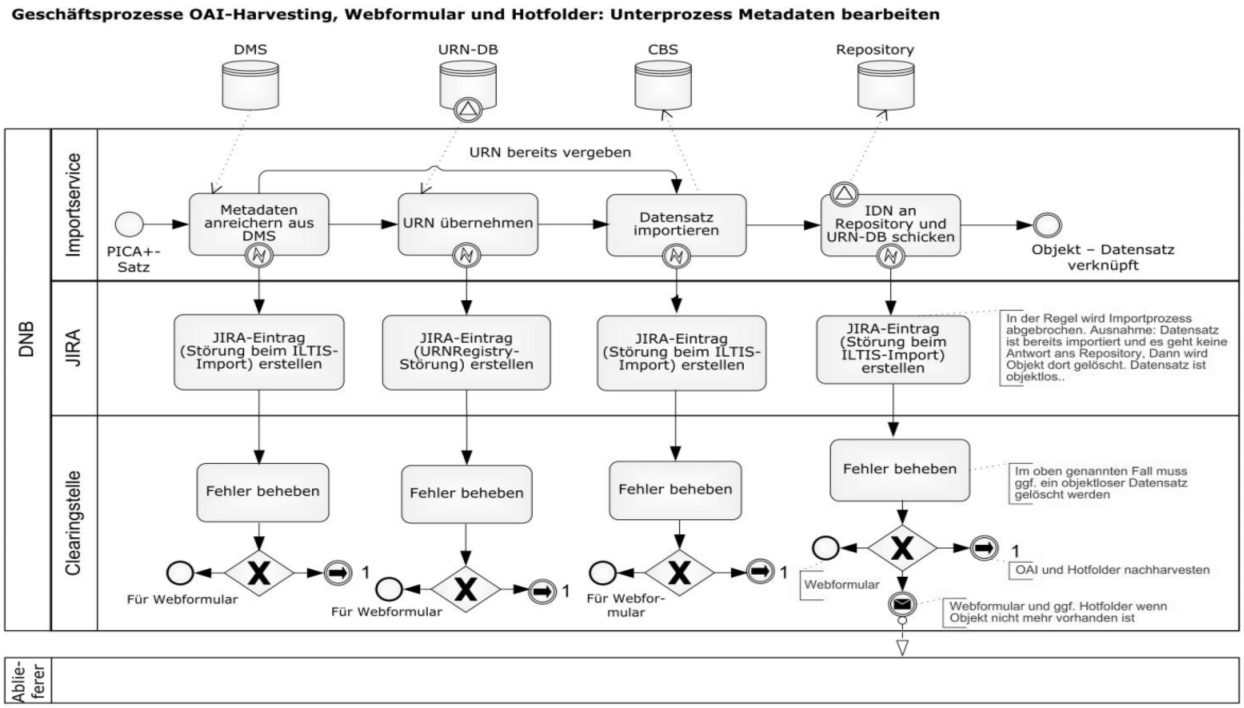
Geschäftsprozesse OAI-Harvesting, Webformular und Hotfolder: Unterprozess Metadaten bearbeiten
DMS
URN-DB
CBS
Repository
DNB
Importservice
JIRA
Clearingstelle
Ablie-ferer
URN bereits vergeben
PICA+-Satz
Metadaten anreichern aus DMS
URN übernehmen
Datensatz importieren
IDN an Repository und URN-DB schicken
Objekt – Datensatz verknüpft
JIRA-Eintrag (Störung beim ILTIS-Import) erstellen
JIRA-Eintrag (URNRegistry-Störung) erstellen
JIRA-Eintrag (Störung beim ILTIS-Import) erstellen
JIRA-Eintrag (Störung beim ILTIS-Import) erstellen
In der Regel wird Importprozess abgebrochen. Ausnahme: Datensatz ist bereits importiert und es geht keine Antwort ans Repository, Dann wird Objekt dort gelöscht. Datensatz ist objektlos..
Fehler beheben
Fehler beheben
Fehler beheben
Fehler beheben
Im oben genannten Fall muss ggf. ein objektloser Datensatz gelöscht werden
Für Webformular
1
Für Webformular
1
Für Webfor-mular
1
Webformular
1
OAI und Hotfolder nachharvesten
Webformular und ggf. Hotfolder wenn Objekt nicht mehr vorhanden ist

Figure 1. DNB interfaces.

combination with a transfer URL submitted within the metadata.

REFERENCES

1. http://kopal.langzeitarchivierung.de/index.php.en
2. Reinhard, Altenhöner & Steinke, Tobias. Kopal: cooperation, innovation and services: Digital preservation activities at the German National Library. *Library Hi Tech*, 2010, **28**(2), 235.
3. Reinhard, Altenhöner. The German National Library and digital preservation: Challenges and opportunities. *World Digital Lib.*, June 2010, **3**(1), 1-9.
4. http://kopal.langzeitarchivierung.de/index_objektspezifikation.php.en
5. http://kopal.langzeitarchivierung.de/index_standards.php.en (accessed on 25.2.2012).
6. http://kopal.langzeitarchivierung.de/index_software.php.en (accessed on 25.2.2012).
7. http://www.gesetze-im-internet.de/dnbg/index.html