BUILDING BLOCKS FOR THE NEW KB E-DEPOT

Hilde van Wijngaarden **Judith Rog** **Peter Marijnen**

Koninklijke Bibliotheek
Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands

ABSTRACT[1]

The National Library of the Netherlands (KB) will renew its digital archiving environment. The current system, the e-Depot with DIAS by IBM as its technical core, has been operational since 2003 and needs to be updated. More importantly, a new system is required because KB has published a new strategic plan with ambitious goals. They require development of an infrastructure that can process, store, preserve and retrieve millions of digital objects, now and for the long term. The digital collections will include e-journals, e-books, websites and digitized master images and will grow from 20 TB currently to 720 TB in 2013. The New e-Depot will also implement tools for digital preservation, as being developed in international collaboration (Planets, JHOVE, etc.).

Together with eight European national libraries, KB defined the architectural framework for the new system. It is based on a modular approach and translated into so-called building blocks for a preservation environment. This paper discusses the building blocks and the rationale for the components-based architecture of the New e-Depot. Currently, requirements for all the building blocks are finalised. A market consultation for the workflow component will starts in the summer of 2010 and the procurement process for the other components will follow in the fall. The first iteration of the New e-Depot will be delivered in 2012.

1. RENEWING THE E-DEPOT

In January 2010, the new strategic plan of the Koninklijke Bibliotheek, the National Library of the Netherlands (KB), was published [1]. This new strategic plan is an ambitious plan with a strong focus on the digital library: digitisation, online access and long-term storage. To put this plan into action, KB needs an infrastructure that can process, store, preserve and retrieve millions of digital objects, now and for the long term. The current digital processing and archiving environment, the e-Depot, cannot fully address the new challenges and will be replaced by a new, improved and extended processing and long-term preservation environment.

Digital archiving and permanent access has been a key priority of the KB since the late nineties of the 20th century. After experiments and prototyping, KB and IBM developed an archiving environment between 2000-2002. In March 2003 the current e-Depot, with the IBM system DIAS [3] as its technical core was taken into production. Since then, more than 15 million e-journal articles from major international publishers have been loaded into the system.

This environment will be renewed for the following reasons:

- KB sets out to process and preserve multiple types of digital collections while the current environment is tailor-made for processing and managing e-journal articles.
- KB needs to upscale its processing and storage environment for:
  - processing at least ten times as many digital items in a limited time frame as it does currently;
  - processing digital items that will be much larger then they are currently;
  - storing and managing at least twenty times as many Terabytes than it does currently (estimation: up to 720 TB in 2013).
- Functionality for identification, characterisation, format-conversion, and other newly developed preservation functionality has to be added to the system to ensure permanent access.
- Software combinations that are used in DIAS have reached their 'end-of-life'. Although all components are standard IBM products and are still supported, their current combination in DIAS is becoming vulnerable.
- The KB-IBM maintenance contract will expire in September 2012.

First plans to renew the e-Depot environment have started in 2007. This included a collaborative effort to set requirements for digital preservation functionalities and services with the Deutsche National Bibliothek (DNB) and the Niedersächsische Stats- und Universitätsbibliothek Göttingen (SUB). During the period March to October 2009 this international collaboration was extended and renamed to the LTP Working Group. Several meetings were held with representatives of eight National Libraries in Europe (Spain, Portugal, Switzerland, Germany, UK, Czech Republic, Norway and the Netherlands) to explore the possible collaboration in developing and

[1] This paper reflects the work and writings of the New e-Depot team at KB, consisting of Judith Rog, Jeffrey van der Hoeven, Aad Lampers, Yola Park, Peter Marijnen, Liedewij Lamers and Maarten van Schie. This paper is a joint paper of the whole group.

implementing a next generation long-term preservation system. Together, the libraries worked on scoping and defined a modular approach and so-called building blocks for a preservation environment. To actually enter into a Request for Information (RfI) process together turned out to be too challenging due to different needs and planning- and budget constraints. However, cooperation was continued on further information sharing and working towards common long-term preservation services [4]. The eight national libraries decided to include each other in their development/procurement processes with sharing information and if possible inviting each other to join in meetings with suppliers.

2. SCOPING THE LONG-TERM PRESERVATION SYSTEM

One of the outcomes of the international working group was what we called the 'two-layered OAIS-model'. When starting to work on joint requirements, we started with a discussion on scope. The OAIS-model describes the necessary depot-functionalities for a long-term digital archive. But what does this mean when translated to detailed requirements? How much does a long-term digital archive have to 'do' when compared to the wider digital library infrastructure, or even the library functions as a whole? In our view, a library consists of a number of depots and the OAIS-functions are applicable to each of them. This starting point opens the need to define which functionality should be realised at library-level and which functionality should be realised at (e-) Depot-level. The previously mentioned LTP Working Group agreed to a two-layered OAIS model-approach as presented in the picture below. It defined which (part(s) of) OAIS-functions should be centralised at library level (i.e. identity management, billing-functionality) and which (part(s) of) OAIS-functions are executed at depot-level. This resulted in the picture shown in figure 1. During requirements elicitation for the New e-Depot, this model has proven to be very helpful in scoping and discussing expectations throughout the different departments of the library [2].

3. PRESERVATION LEVELS

An important requirement for the New e-Depot is that it should be capable of processing and managing multiple digital collections at different preservation levels. Not every collection represents the same value to the library, not every collection is preserved for the same reasons and not every collection needs the same treatment to ensure permanent access. If all digital publications have to be processed and managed at the highest quality level, the digital archiving environment would become unaffordable. KB therefore defined a set of preservation levels and a value- and risk methodology to link preservation levels to digital collections. This policy has not yet reached its approved version, but the general outlines are clear and have to be put in practice by the New e-Depot system (see [2] for more details).

The preservation levels will consist of:

- Level 0 for collections that will not have to be preserved by the library and will only be stored in a presentation environment;
- Level 1 or 'limited' level for collections that have to be preserved for more than five years but will not need to be fully checked on ingest and do not need large scale investment on preservation actions;

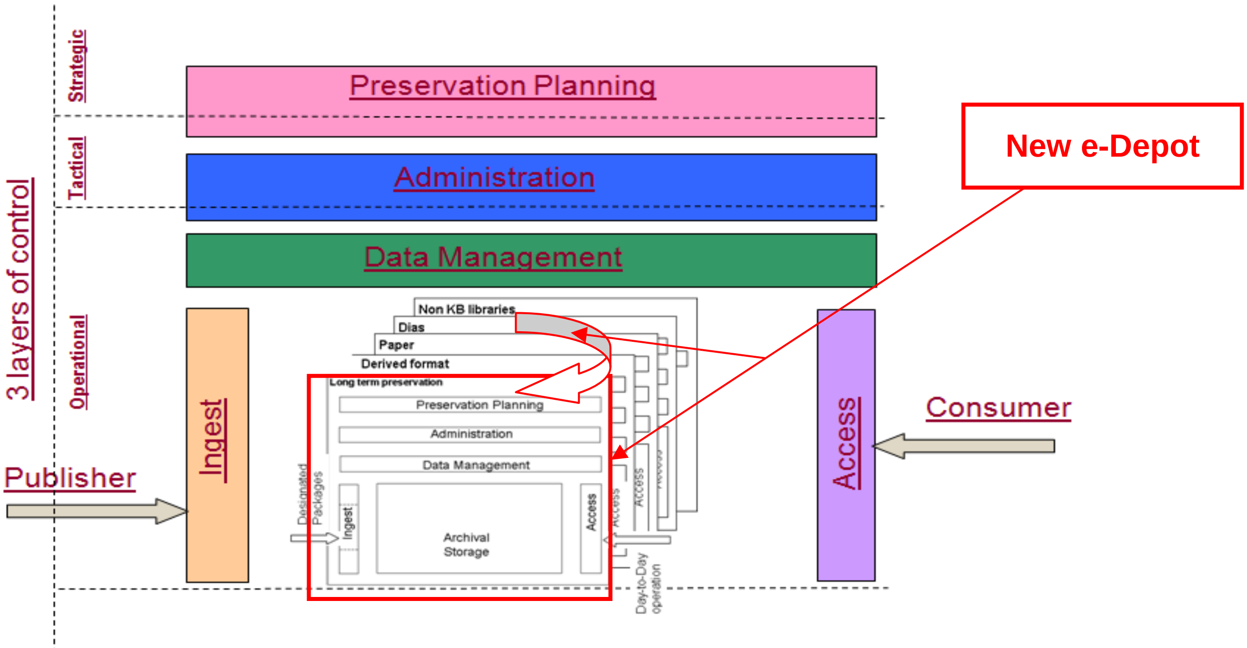

Figure 1. Two-layered OAIS model

- Level 2 for collections that do have to be preserved for the long term (is more than five years), need to be checked on ingest but do not require future access in original file format. These collections will be subject to validation, will be stored on preservation storage, but may require less preservation actions;
- Level 3 for collections that have to be preserved for more than five years, need full ingest validation and preservation actions that secure future access in an authentic way.

4. DEFINING THE COMPONENTS FOR THE NEW E-DEPOT

Based on eight years of experience running the current e-Depot system, on international discussions, on working on digital preservation research projects and on growing insight into the processes that need to be supported, the KB team defined a components-based architecture for its New e-Depot environment. Three basic considerations have led to this set-up.

First of all, digital preservation is not just a matter of identifying technical requirements for secure storage. Far more than that it is the holistic approach of an organisation to achieve its preservation goals. It is defined by the services that institutes deliver, by checks on the publications on ingest, by management of information on the objects and the processes, by closely monitoring ICT developments and assessing what these developments mean, by storage management and the overall architecture of the preservation environment. So it's not just a few extra things you do after you store digital objects, but it is inherent to the organisational approach, the work processes and the automated steps that process, store and use digital content.

This leads to the second consideration, where preservation functionality is to be seen as an addition to more general requirements for a processing and storage environment. As digital preservation is a result of organisational approaches, work processes and systems, not all preservation functionality depends on specific systems. A long-term storage environment has perhaps extra features but is also ‘just’ a storage environment. Ingest for a long-term preservation system does include extra functionality but is also 'just' a processing workflow. It can very well be that standard IT solutions can deliver most of the required functionality.

Thirdly, a components based set-up allows for more flexibility, avoids vendor lock-in and makes it possible to choose the best product for each part of the archiving environment. As the KB is experiencing at this moment replacing a complete and integrated digital archiving system at once is a very challenging task. Choosing a modular approach will allow the KB to extend and improve the new system one module at a time. In the future, components must be replaceable by modern technologies more easily. This will add to the stability of the systems and avoids changes to stored content and metadata. Which brings us back to the first consideration, digital preservation is more than secure storage alone.

Based on these considerations, the KB team started to 'break down' components of a processing and storage environment into separate processes and services and made a translation into what we started to call 'real world building blocks' (not to imply that a library is not the real world...). By defining a combination between generally available IT components and special requirements for digital preservation, it became possible to set up an approach that would allow us to make optimal use of (commercial or open source) off-the-shelf software together with preservation focused services.

Considering the specific characteristics of the work processes that need support from the New e-Depot and quality attributes such as performance, adaptability, resilience to interferences and stability, the building blocks for the New e-Depot were chosen as depicted in figure 2. Each of these building blocks or modules will be described in more detail hereafter.

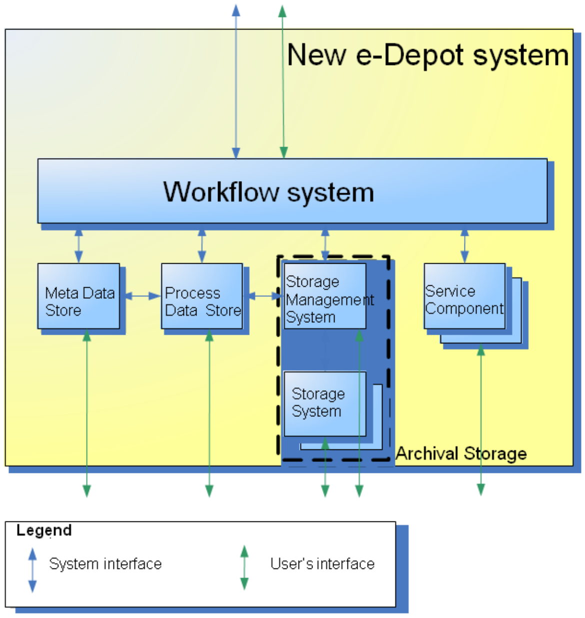

Figure 2. Building blocks New e-Depot

5. ARCHIVAL STORAGE MODULE

Archival Storage is provided by an implementation of a two-layered storage-solution. A Storage Management system abstracts other system components (more specifically the workflow system) from the actual storage provided by Storage Infrastructure (consisting of various storage media and network components).

The storage infrastructure will ensure that files are written and read to the actual storage media, in this way holding the enormous volumes of the actual bits of the material to be preserved. Because of the enormous and ever growing volumes and the crucial role in the archiving system, the storage infrastructure will have to have the following characteristics:

- highly reliable;
- scalable to very high volumes;
- cost effective;
- self monitoring.

Cost effectiveness plays a role for each of the modules, but for the storage infrastructure it is of the utmost importance. The volume of data will only increase year by year. The investments and operational costs for the storage media are by far the highest cost factor in the archiving system and will have to be controlled.

The Storage Management layer has an essential role within Archival Storage. It stores and retrieves files based on assigned unique identifiers, since access can not be based on storage locations or filenames. The reason is that the lifecycle of storage locations (e.g. the precise storage infrastructure and used media) and storage methods (defining constraints to filenames and locator structures) will undoubtedly change. These changes can be driven by the storage management system itself, when creating replica's of the original content on other storage infrastructures to safeguard the content from loss through hardware errors or disasters. Storage Management will also move large volumes of data to new storage infrastructures when older infrastructure or media are phased out. Storage Management should abstract all other system components from future technological changes. If this does not function properly, the stored data may well be perfectly retained in the storage infrastructure layer, but may be no longer accessible.

Another important requirement for a storage management layer is that it does not lay any restrictions on the storage infrastructure that is used in combination with the storage management layer. Therefore a storage management solution will have to be:

- highly reliable;
- independent of underlying storage infrastructure;
- implements a well described method and data store to connect content identifiers to storage locators.

6. WORKFLOW MANAGEMENT MODULE

The processes needed for ingest, access and preservation actions are provided by a Workflow Management system. This system implements ingest, access and preservation functions as defined workflows. It consists of a Process or Orchestration layer, a Mediation layer and a Transport layer to connect all systems to the workflows. The fourth layer is the service layer, that uses Service Components to implement specific atomic functionality to perform amongst others content analysis, content transformations and metadata conversions. The workflow system effectively implements the integration layers of a Service Oriented Architecture (SOA) [5]. This also implies that the workflow system will offer the entry point for all integrations with external systems.

Of the three types of processes the workflow module will have to support, the ingest process will put the highest demands on the system. Each day, tens of thousands of publications arriving at the KB in a large diversity of submission formats, containing several different file formats, will have to be validated and, if necessary, normalised to a more generic format. Depending on the preservation level, during long-term management, several preservation actions will be performed on the material. Next to integrating and orchestrating services, the workflow system will offer functionality to prioritize and parallelize workflows and service executing, perform load balancing to optimize resource usage and offer message persistence and workflow resilience services.

The Workflow module must be capable of:

- processing high-volumes of data;
- support multiple workflows dependent on content types, producer and required preservation levels;
- offer support for manual intervention and repairs of invalid content and metadata;
- run different workflows in parallel maximizing throughput and balancing system load;
- allows for restart and recovery of failed workflow instances;
- halt and automatically resume workflows when services are temporarily unavailable;
- minimizing the development effort to implement new ingest streams (workflows);
- support the easy integration of specific preservation tooling.

7. META DATA STORE MODULE

Conformant with the OAIS model, metadata is stored in Archival Storage with the content. This makes the metadata subject to preservation together with the content. However, it also makes the metadata difficult to use by services that support the preservation action and access processes. To allow direct access to metadata needed to control these processes, it is not only stored as files, but also redundantly maintained in a Meta Data Store.

The metadata stored within the Meta Data Store serves as access mechanism to the data objects stored in the Storage Module. Its data model therefore structures the relationships between stored data objects and their metadata to enable the retrieval of selected Archival Information Packages (AIP). The data model offers placeholders to store identifiers the outside world can use to request stored content, when needed in a specific version or variant. The Meta

Data Store also holds provenance data on actions performed on content and versions created.

The Meta Data Store will offer reporting functionality to query the systems database giving insight in the holdings of New e-Depot. To create these reports all stored metadata attributes can be used in queries and to structure the report. This report will be used as input in the preservation planning process, driving decisions on which preservation actions need to be performed.

The Meta Data Store will only hold a minimal amount of descriptive (or bibliographic) metadata and will therefore not be used directly by end users for requesting content stored in the New e-Depot. A separate bibliographic cataloguing system is available to search for content. Identifiers will be used to link the Meta Data Store of the New e-Depot with the external cataloguing system.

The relationships between stored metadata in the archival storage module and the data in the Meta Data Store are defined in such a way that the Meta Data Store can be rebuild when a disaster occurs using the metadata stored in the Archival Storage.

8. PROCESS DATA STORE MODULE

The purpose of the Process Data Store is to support the Monitoring & Control process for the New e-Depot system. More precisely:

1. it provides information on the execution of processes in terms of:
   a. measuring process execution results in a certain period to enable reporting on Key Performance Indicators (KPI's);
   b. reporting of deviations from the normal flow of operations (relative to defined benchmarks and tolerances);
2. it provides information to analyse trends in the growth and evolution of the collections processed and stored;
3. it provides information to merge process results with collection metadata, thus enabling analysis of the collections and related processes;
4. it provides information to sustain the integrity of all collections stored (both AIPs and Descriptive Information);
5. it supports consistency checking between the Meta Data Store and Archival Storage.

The Process Data Store does not provide the daily monitoring and control of the operational processes which are the responsibility of each system that performs or supports that operational process (i.e. primarily the Workflow Management Module).

The Process Data Store collects and receives data from other modules of the New e-Depot system and transforms and integrates them for reporting purposes. There is no automated feedback loop to these other modules and none of those other modules will depend on the Process Data Store for its proper functioning. The output of the Process Data Store will be used by operators and management for monitoring and control purposes.

9. DEVELOPMENT OF THE NEW E-DEPOT

Each module for the new system has been defined in detailed specification of requirements. On top of that, an overall architecture and data model have been designed. After a final review of the requirements, the procurement and development process will start in June 2010. A request for each component will be placed in the market separately and is expected to be filled in differently. While workflow systems are widely available, both commercially and open-source, storage management is more specific and will see a different number of possible applications. Modules will either be bought, integrated or developed. The success of the approach will be largely defined by how the modules will be integrated, with each other, but also in the KB infrastructure. During the next few months, after the choice for applications and development of services has been made, it will be decided how the integration will be managed. In general, such an integration will look like as depicted in figure 3.

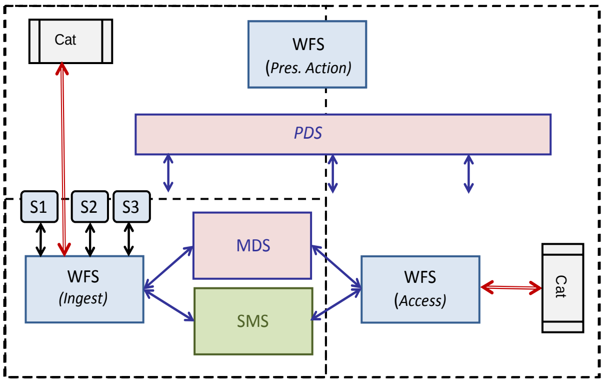

Figure 3. **New e-Depot integration**

The workflow system will be present in several processes (ingest, retention / preservation, access) and interacts with the New e-Depot modules Storage Management, Metadata Store and Process Data Store, and with external systems such as existing cataloguing services and possibly others.

10. REFERENCES

[1] KB Strategic plan 2010-2013, available at: http://www.kb.nl/bst/beleid/bp/2010/index-en.html (accessed 9 July 2010).

[2] Hilde van Wijngaarden, *The seven year itch. Developing a next generation e-Depot at the KB.* Paper to be presented at the World Library and

Information Congress, Gothenburg August 10-15 2010, available at: http://www.ifla.org/files/hq/papers/ifla76/157-wijngaarden-en.pdf (accessed 9 July 2010).

[3] IBM's Digital Information Archiving System (DIAS), available at: http://www-935.ibm.com/services/nl/dias/is/implementation_services.html (accessed 9 July 2010).

[4] *Long-term Preservation Services – a description of LTP services in a Digital Library environment*, Long-term Preservation Working Group, white paper will be available in July 2010 at: http://www.kb.nl/hrd/dd/index-en.html (accessed 9 July 2010).

[5] Bell, Michael (2008). "Introduction to Service-Oriented Modeling". *Service-Oriented Modeling: Service Analysis, Design, and Architecture*. Wiley & Sons. pp. 3. ISBN 978-0-470-14111-3