Date submitted: 06/07/2010

The seven year itch
Developing a next generation e-Depot at the KB

Hilde van Wijngaarden
National Library of the Netherlands (KB)
The Hague, The Netherlands

Meeting: **157. ICADS with Information Technology**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden

http://www.ifla.org/en/ifla76

Abstract:

In 2003 the digital archiving system e-Depot of the National Library of The Netherlands (KB) became operational. This system was developed together with IBM and at the time of implementation it was the first long term preservation library system, based on the OAIS reference model, running globally. Today, seven years later, the system has processed and stored over 15 million digital objects, mainly e-journal articles. And, after seven years of operation, KB has started the process to replace the system.

Building on experience and at the same time building a system with much broader capabilities, has lead to new choices about approach and architecture. Two important features of the new set-up are a components-based approach and support for differentiated processing, making use of preservation levels. The current design is the result of two years of requirements gathering, research and discussions. Very import for this process was the collaboration with a group of European National Libraries that worked together on a common understanding of how to design a preservation architecture.

In this paper I will discuss the reasons for replacing the current system, the requirements and approach for this new system and how we worked together with other national libraries.[1]

Reasons for replacing the current system

The current IBM DIAS implementation for the KB e-Depot was based on requirements set in the late nineties. Taking those requirements at this time, it was remarkable to see how much of those will also apply to our new system. But this comparison also shows that many of the requirements did not make it into practice for the first system, simply because tools and functionality that are needed for long term preservation of digital publications did not exist

[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 and Liedewij Lamers. Many thanks to them and also to Marcel Ras, head of e-Depot for his comments and additions.

yet in the year 2000. Today, more functionality does exist (although not nearly enough operational solutions are available yet), but more importantly, we have a better picture of the components that build up a flexible, durable preservation environment. This results in a different approach towards developing a preservation system this second time around.

There are several reasons the KB has chosen to develop a new preservation system, the 'seven-year-itch' or the system already 'living' longer than most other IT systems, only being one of them. In 2012 our maintenance contract with IBM runs out and components of DIAS will no longer be supported, but there are many changes within the library that call for a new system as well. These changes are not specific for KB, but are a general development in libraries' digital collection management and are a consequence of digital library developments. In short, these changes are:

- scale: digital publishing, webarchiving and digitisation has lead to enormous growth of digital collections
- requirements for digital collection management: while preservation was first focused on special parts of our collections, with the growth of digital collections, preservation has become a core requirement of libraries collection management
- progress in digital preservation R&D: new tools have become available that allow us to better process and manage digital collections (eg. Tools for identification, characterisation, migration and emulation)
- diversity of digital collections: digital publications (including websites) have become container formats with all types of multi media components embedded. These formats are a challenge for permanent access.

In January 2010 KB published an ambitious new policy plan, with three main goals: give access to all Dutch publications (which means digitize everything), preserve these publications for the long term and facilitate a national infrastructure for Access to all these publications. For the year 2013 'smart' milestones were defined that could be reached within three years.

The ambition to preserve all born-digital and digitised publications for the long-term sets a major requirement for the new e-Depot system: this has to process and store all types of digital collections. The current e-Depot has been focused on e-journals and can process digital monographies, the next system will have to process and store e-journals, e-papers, e-books, digitised images

Preservation levels

When faced with the enormous challenge of having to process and preserve hundreds of millions of objects of very diverse nature and future use, you need a system that is not just scalable but also flexible. And you need a set-up that is pragmatic enough to offer a realistic approach. It will just not be possible or affordable, to process and preserve all digital collections at the highest quality. But that is not necessary either. Not every collection carries the same risk or needs the same treatment to be accessible over a long period of time.

The very obvious example here is the difference between a collection of digitised books and a collection of born-digital e-journals. If digitisation is carried out within control of the library, using the quality settings the library has defined, checks on ingest for the preservation system can be limited. To establish the value of the digitised collection, and through that value the

preservation goal (short- or longterm), this could depend on the original book and the intention of scanning it. In the case of preservation scanning, the digitised version is the preservation copy and has to be preserved for the long-term. When scanning is done for access and the original paper copy is in good enough shape to be scanned again, the digital copy may be preserved for a shorter period.

When we look at born digital e-journals, as the other example, there is no paper version, so the digital version is the preservation copy by nature.[2] Production has not been controlled by the library and it is difficult to influence the digital production line. Publishers are very willing to cooperate and try to follow preservation guidelines, but they will not restrict use of new formats, complex objects, moving images etc. within their publications. E-Journal publication are becoming more complex everyday and set a challenge for ingest processing. If these publications are to be preservable, we at least need to know and register what they're made of. And the complexity of e-journal article is even relatively small compared to websites.

Looking at websites, as a third example, it may be obvious that it is more difficult to preserve these for the long term than preserving a plain text or pdf file. A website contains all kind of functionality, look-and-feel, structure, behaviour and probably video, sound and images. The risk of loss of one of these elements is much higher and the effort for preservinf it could be much higher.

What preservation actions (migration, emulation or any other activity to assure future use) we need to have ready for our collections, does not just depend on the file format of the publication, but also on the value that publication or collection represents. And what that value is about. What do we need to preserve, in terms of content or look-and-feel and how 'preservable' is the file format? [1]. The answer to this question defines the required investment in preservation planning and development/implementation of preservation action tools.

These previous paragraphs have become common ground for those of us that work in digital preservation. What we have to do now is turn theory into practice. How can the diversity of collections, goals and challenges can be managed in a preservation environment? At KB, we are defining a process that includes value and risk assessment of the digital collections and concludes with setting a preservation level. Assigning a preservation level is based on:

- the value of the collection
- the complexity of the collection (structure, file formats)
- the required authenticity and significant properties to be kept for the long term
- knowledge on file formats
- the availability of tools and services which enable preservation and preservation actions
- the availability of knowledge and experience
- the costs and available budget
- the requirements of the (future) end users

[2] There is still a question of the number of copies that are stored all over the world because not every digital copy has to be preserved for the long term. But this is a matter of dividing responsibilities between national libraries…

Part of this process are policy documents like our Strategic plan, collectionplan and collection management plan. The Strategic plan is available, the collectionplan will be as soon as possible and the collection management plan is being written at this time.[2] This paper gives some preliminary directions in the setting of preservation levels, that might change when plans have become final.

All though definite preservation levels have not been set yet, the main levels will consist of:

- Level 0 for collections that will not have to be preserved by the library;
- Level 1 for collections that have to be preserved for more than five years but do not need to be fully checked on ingest or do not need large scale investment on preservation actions;
- Level 2 for collections that do have to be preserved for the long term (= more than five years), need to be checked on ingest but do not require future access in original fileformat;
- 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.

Preservation level 0 to 3 each refer to specific treatment in the New e-Depot, as expressed in the table below:

Level	Storage regime	Ingest validation	Preservation actions
0	Presentation	No	No
1 (limited)	Durable storage	No	No
2 (active)	Durable storage	Yes	No
3 (pro-active)	Durable storage	Yes	Yes

The process from value- and risk assessment to setting the preservation level takes place before an object has to be processed and stored. Once the preservation level is set, it will be included in the stream profile, that describes the content stream and defines the business rules for processing and storage.[3] And by choosing a process like this, one of the major requirements for the New e-Depot, being able to process and store different content streams differently, was set.

International collaboration

Flexibility and scalability are major requirements for the new e-Depot of the KB, as discussed above. Other requirements setting is based on our experience with our current system, but also on discussions and requirements setting with international colleagues. In 2008, KB, the National Library of Germany (DNB) and SUB Göttingen worked together on requirements for a next generation preservation system. This collaboration was based on the joint use of the IBM system DIAS. In the beginning of 2009, KB and DNB sought co-operation with other national libraries within Europe in order to share experience, knowledge and resources. But also to see what was possible if we could join forces. Commercial off-the-shelf solutions are still not widely available and the solutions that are available bring the risk of a lock-inn scenario. If a group of national libraries would define requirements together, and possibly

[3] This process is similar, and partly inspired by, the process of the BnF system SPAR, as described in the paper [3], to be presented at the same session at IFLA Gothenborg.

even tender together, this could trigger commercial suppliers to invest more in developing solutions that answer to the general requirements.

In the spring of 2009 the national libraries of the UK, Germany, Norway, Spain, Portugal, Switzerland and the Czech Republic set up workshops to discuss:

- standardization of interfaces
- the modular approach for the system architecture
- operational tools for characterization and identification, quality control and preservation actions

The result of these workshops was an architectural outline, with as main features the use of a two-layered OAIS model and a components-based set up of the preservation system. In October 2009 it was decided that it would not be possible to enter into a joint tender, caused by different timelines, but UK, Norway, Germany and the Netherlands did have another workshop to work on the definition of services that would be required within a long term preservation environment.[4]

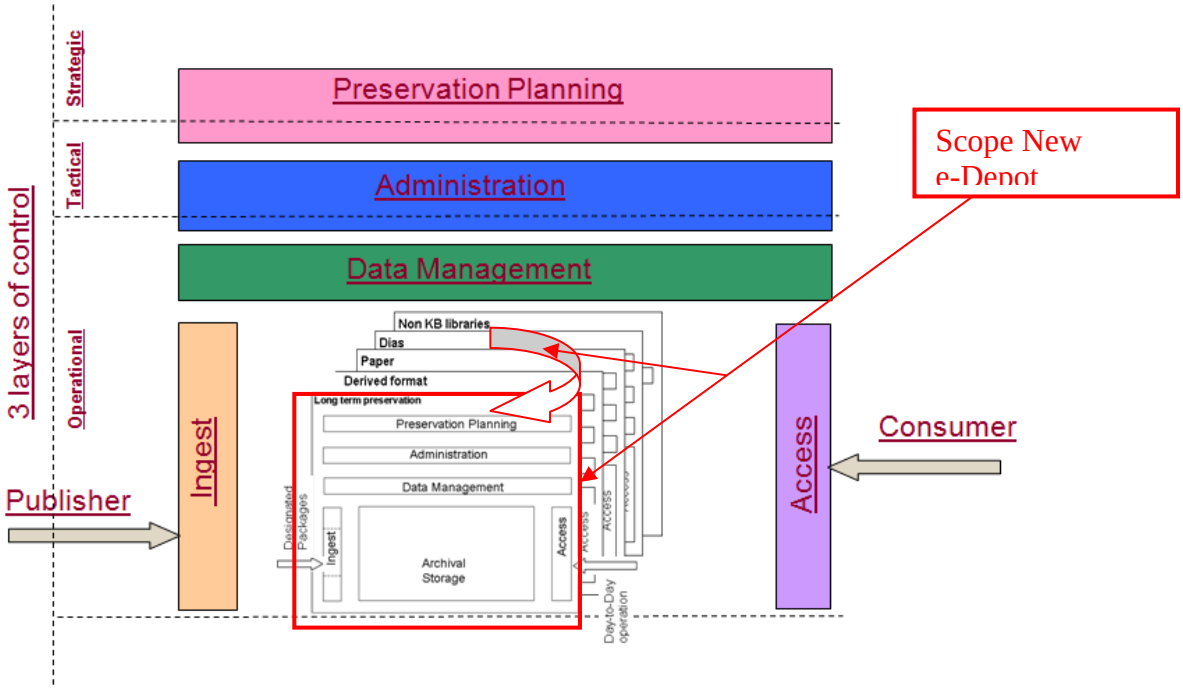
For KB, the international workshops have been crucial for requirements building and working on architecture and datamodelling. The set-up for the new e-Depot system, consisting of separate components, is largely inspired by the work of the joint national libraries.

Two-layered OAIS-model

The OAIS-model describes the necessary depot-functionalities for a long-term digital archive and, obviously, a requirement for the New e-Depot is that it will be OAIS-compliant. However, discussing what that means in practice, it's especially scoping that becomes a problem. How much does a long-term digital archive have to 'do' when compared to the larger digital library infrastructure, or even the library functions as a whole? The OAIS does not contain a multiple depot approach. In our view however a library consists of a number of depots, at least a paper-depot and an e-Depot. The OAIS-functions are applicable to each depot. This starting point opens the need to define which functionality should be realized at library-level and which functionality should be realized at (e-) Depot-level. The previously mentioned international working group agreed to a 2-layered OAIS model-approach as presented in the picture below. It has been defined which (part(s) of the) OAIS-functions should be centralized at library level (i.e. Identity management, billing-functionality) and which (part(s) of the) OAIS-functions are executed at depot-level.

[4] The definition of services will be published as a white paper in the summer of 2010.

This resulted in the picture below:



During requirements building 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. It is assumed that the OAIS-functions (Ingest, Archival Storage, Datamanagement, Access, Administration and Preservation Planning) should be implemented at 2-levels, at library-level and at depot-level. For example access to the New e-Depot for end-users is divided over at least two systems, an (functional) access function at central library level (end-user access) and an (technical) access function at depot level (depot access).

Components or building blocks

KB has chosen a components-based approach for the new e-Depot (components can also be called ‘building blocks). This has been based on a number of arguments:

- A components-based approach allows for the combination of the best solution (either open-source or commercial) for each building block.
- The KB wants to avoid a ‘lock-in’ scenario with one supplier.
- IT-systems do typically have a lifetime of appr. 5 years, much shorter than the expected duration of the storage of the digital content (infinitely). To cope with this problem the new e-Depot-system must be prepared for its own adaptation and replacement. The preferred way to do so is to build the ability to replace individual components of the system asynchronously, when new and better technology emerges asynchronously for each component.

The definition of the components is based on the process and scope for the New e-Depot and on a translation of required components to what would be available (commercial or open-source) on the market.

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 multiple **Storage Infrastructure.**

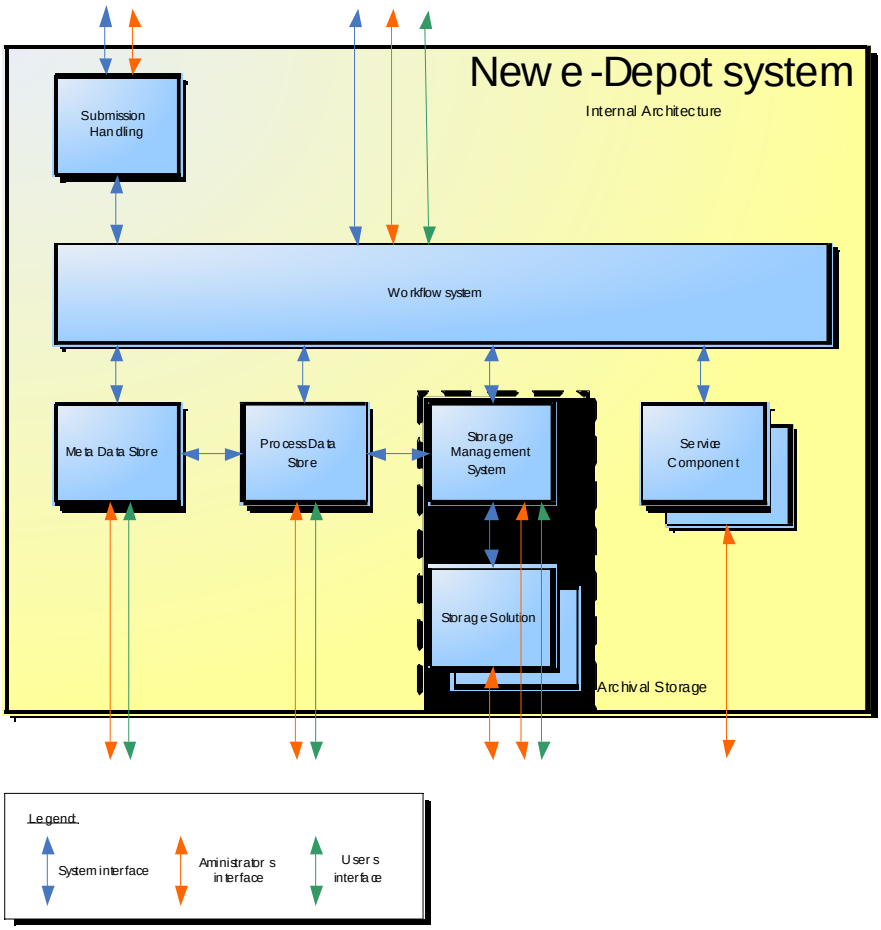
The processes needed for *Ingest* and *Access* and *Preservation* actions are provided by a **Workflow system**. The workflow systems implements ingest, access and preservation functions as defined workflows. It consists of Integration, Messaging and Orchestration layers to connect all systems to the workflows. The workflow system uses **Service Components** to implement specific atomic functionality to perform content analysis, content transformations and metadata conversions.

The Metadata produced by Ingest and needed within Access and *Preservation Management* to search and browse thru content is stored in a **Metadata Store**.

To support monitoring & control the metadata store is complemented by a **Process Data Store** that acts as a data warehouse for all administrative and process data produced by individual components. It supports combined searches over the Process Data Store and the Metadata Store.

Finally a separate **Submission handling** system provide for functions to retrieve or receive Electronic Publications from Publishers, implementing both Push - and Pull interfaces. It will also act as input cache functions.

This set-up is presented in the picture below:



Projects are organised around the components and as a group of projects they are managed by the KB Digital Library Programme. With definition of set-up, scope and requirements, as presented above, an important step towards developing the new environment has been delivered. Currently, the team is working on market consultation for the separate components of the system. Official tendering will start in the fall, starting with the workflow- and

storagemanagement system. A first release of New e-Depot components is planned before the start of 2012.

References

[1] Caroline van Wijk and Judith Rog, *Evaluating file formats for long-term preservation*, in: New Technology of Library and Information Service, iPRES 2007 special edition, 2008, nr 1 & retrievable from http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf

[2] See http://www.kb.nl/bst/beleid/bp/2010/index-en.html for the Strategic Plan.

[3] Emmanuelle Bermes, Louise Fauduet and Sebastien Peyrard, *A data first approach to digital preservation; the SPAR project*, at: http://www.ifla.org/files/hq/papers/ifla76/157-bermes-en.pdf