The KB e-Depot in development
Integrating research results in the library organisation

Hilde van Wijngaarden, Frank Houtman, Marcel Ras

hilde.vanwijngaarden@kb.nl, frank.houtman@kb.nl, marcel.ras@kb.nl

National Library of the Netherlands

Prins Willem-Alexanderhof 5

2509 LK The Hague

The Netherlands

Abstract

The mission of the KB e-Depot is to ensure permanent access to large quantities of digital resources in a national and international context. Operating an international e-journal archive at a relatively small organization such as the KB asks for a firm foundation of its policy. With support from the Dutch government, the KB has succeeded in setting up two expert teams: an operational team responsible for daily operations of the e-Depot (based in the Acquisitions & Processing Division) and an active research team that secures continuing research and development to secure long-term preservation and perpetual access to electronic information (based in the Research & Development Division). The experience gained in operating an archive is directly used in system- and process improvements. As are the results of the projects in which the R&D team is involved. The organisation around the e-Depot is based on this pragmatic approach.

National Library of The Netherlands

The Koninklijke Bibliotheek (KB, National Library of the Netherlands) is a scientific library and a deposit library. Its mission is to collect published information, preserve it and provide permanent access to the information for use in research, education or for any other purpose in society. In most countries, publications have to be deposited by law. The Netherlands does not have an act or provisions in law concerning depositing. KB works with a voluntary deposit system based upon agreements with the publishers. This has resulted in nearly complete coverage of the print publications produced by commercial publishers in The Netherlands.

Digital archiving system

In the early nineties of the last century, KB began discussing archiving of digital publications. In 1996 an agreement was signed with Elsevier and the first experiments with digital archiving started. The Dutch Publishers Association agreed on a new arrangement in 1999, which covered also online digital publications with Dutch imprint. The traditional model, based on national deposits and geographical boundaries, is no longer valid for guaranteeing the long-term preservation of the international digital academic output. Academic literature is produced by multinational publishers, and has often no longer a country of origin that can easily identified. In line with the international nature of information provision, the KB decided to open up its e-Depot to international publishers in 2002.

In that year, a landmark archiving agreement with Elsevier was signed, including all Elsevier e-journals instead of the e-journals with Dutch imprint. This arrangement turned the National Library into the first digital archive in the world for e-journals published by international scientific publishers. Other publishers followed and currently, KB has agreements with 14 of the most important international publishers. In 2008, 11 million digital objects are stored in the e-Depot, mainly e-journal articles in pdf formats.

The core of the e-Depot is the Digital Information Archiving System (DIAS), developed during a two year project between 2000 and 2002. DIAS is a combination of standard IBM components, with extra functionality that allows the system to interact with the library infrastructure. [1]

Based on the agreements with the publishers, e-Journals are delivered to the KB and ingested into the system automatically. The error recovery procedure is the only manual effort involved. Metadata are delivered by the publisher and converted to the KB format and added to the KB catalogue. Access policies depend on agreements with the publishers. Commercial content can be accessed on site only and trigger events that will allow the KB to open up online are being discussed.[2]

In addition to the e-journal articles, more collections and different types of material will be stored in the e—Depot in the very near future. These new types of material will be more complex, but also be more voluminous.

Apart form the core DIAS, the e-Depot consists of different modules that allow pre processing and access. Especially these components are subject to improvements in the case of adding new types of material.

The national e-Depot

As mentioned above, procedures and workflow were initially designed for the national e-journal archive. But because of the nature of e-journal publications these were quickly dedicated towards an international e-journal context.

A selection of Dutch e-journals, deposited in the context of the KB's depository task, is ingested using the same procedures. That means e-journals of larger publishers, delivered in large quantities.

The national e-Depot is the digital version of the deposit of Dutch printed publications. Because of the broad variety of digital objects, acquisition and processing of these materials will be extended gradually. The first step was to set up digital archiving workflows for national deposit of singular e-journals and monographs. Therefore, web-interfaces were set up to allow any depositor to submit publications, monographs or periodicals. Also procedures were designed to process these singular publications. To allow the system and the organisation to process a growing number of very diverse materials, a number of changes and additions have to be implemented in the e-Depot infrastructure. The next step will be to set up workflows for complex objects like websites and other multimedia objects.

The archiving of digital deposit materials will bring changes in the organisation, especially concerning acquisition and processing of these publications. But also on the technical level extensions to the e-Depot system are needed to handle these new processes.

The Dutch web archive

Of a more complex nature are web sites. In 2005, KB started a project with the goal of harvesting and preserving a selection of Dutch web sites. As for the harvesting of web sites, but also the access to the archived web sites, it was decided to make use of standard tools used by other web archiving projects. By using the toolset developed under the colours of the International Internet Preservation Consortium (IIPC), the KB was able to focus mainly on the preservation issues concerning web sites. A selective approach was chosen because of the large volume of the Dutch web and the fact that it was decided to collect web site in full depth as part of the national deposit. A so called domain approach does not make sense in this case.

The project is concentrated on preservation issues, quality assurance and on setting up a procedure for processing web sites based on the current possibilities within the e-Depot.

To put this new procedure in production, a number of changes to the e-Depot infrastructures, as well as additions to the metadata model, are necessary to ensure the durability of the content. [3]

Digitised materials

Another new type of materials to be processed are the masters resulting from national digitisation projects carried out by the KB, e.g. Dutch newspapers from 1618 and Parliamentary Papers. These materials are well structured, but in other file types as we are used to process, Tiff and Jpeg200.

KB e-Depot phase 2

Operating a digital archive entails continuous efforts in improving workflow and infrastructure, maintenance. But also investing in research activities to develop tools that enable us to really do what the e-Depot was meant for: retrieving archived documents for eternity.

During the first six years the e-Depot is operating, the focus was on processing e-journal articles. Generally these are objects of a same type and form. Over 11 million e-journal articles have passed the processes since 2003. The initial workflow did work for the past six years. However, this same set-up will not be sufficient for the many different types of content and content-suppliers that are on the e-Depot itinerary for the next few years. Currently we work with a pre process which has only basic functionalities and only basic quality control on the materials to be ingested. That will have to be improved and could be improved on the basis of our own research and developments, but also based on the results of the PLANETS projects as described above.

Apart from the steady growth of the e-journal archive, the Dutch digital deposit, the web archive and the KB's active digitisation policy and the goal to preserve digital images are the main drivers for extensions and improvements of the current e-Depot. Therefore processing and storage capacity needs to be scaled-up enormously to be able ingest of new collections which tend to very voluminous.

During the last six years, KB also invested heavily in research. Preservation research delivered the initial design of preservation functionalities that are now ready to be implemented into the e-Depot infrastructure: characterisation tools, improvements to the pre ingest process, a normalisation module, a migration module and new requirements for the metadata model. Tools that are strongly needed to enable the improvement of the quality of processes and the procedures for quality control of the objects to be ingested.

To coordinate implementation projects of new functionality running at the same time as the enlargement

of the system and to control dependencies between these projects, KB has set up a programme. Projects run simultaneously while using each others' input. The programme is now in full speed and will deliver the e-Depot infrastructure 'phase 2' in the middle of 2009.

Operating a digital archive in a library

During the development of the e-Depot system in 2002, a working group was started to develop organizational embedding of the system. It was decided to make the Acquisitions & Processing Division responsible for day-to-day operations and to set up a digital preservation research team within the Research & Development Division. The IT department took on the daily maintenance of the system, with IBM staying closely involved. In January 2003, five people started their new responsibilities in the three departments. Today, six years later, 23 fulltime equivalents are dedicated to the e-Depot.

Still, the e-Depot department is responsible for daily operations. Seven collection managers and one functional manager perform tasks which are focused on processing of objects. These are subdivided into different specialist tasks based on the workflow. The Front Desk is responsible for technical contacts with publishers, analysis of content and metadata, guidelines and the set up of processes. The Pre Ingest group is responsible for the technical set up of the process, which means conversions of metadata, writing scripts and style sheets and quality assurance. The Ingest group is in charge of the actual ingest of materials into the DIAS system and error control.

The Digital Preservation department is responsible for preservation research and development. Their daily activities are directly related to the operational e-Depot while they are involved in different European R&D projects like Driver, KEEP, Parse.Insight and PLANETS.

The IT department is entrusted with the technical maintenance of the system and with the coordination of technical improvements on the system. Most of the work on these improvements is outsourced.

The group of KB-staff that has something to do with the e-Depot is even much bigger than that. Access is the responsibility of the User Services Division, management of the relations with publishers is done by the Acquisitions department and cataloguing by the Cataloguing department. All these departments are closely involved in the e-Depot as well.

We are now in a position to evaluate the consequences of running an operational digital archive for the library as a whole and move on to the next phase of improving workflow, enhancing the system and the quality assurance and take a major step in scaling up our storage- and processing capacity, as mentioned above. This could only

be possible because of the firm embedment of functions within the different places of the organization. Commitment of the library as a whole is vital for this.

The e-Depot is a driving force for renewal and change within the organization. The influence is noticeable in three areas. First, there is the content of the library's collections. Taking up the responsibility for long-term digital preservation has made the KB's collection more international, scientific and more diverse in appearance, now also containing multi-media applications, websites, e-books etc.

Secondly, substantial changes had to be implemented in the technological infrastructure that is now also beneficial to the development of other library services. This also includes changes in metadata modelling and handling.

Thirdly, there is the impact of running the e-Depot on people and the organisation. To organise digital preservation activities across several departments is not an obvious choice. And it has not always been the easiest choice either. It requires special attention to coordinate between different departments and to set-up good knowledge management and quality assurance.

However, after six years, we can say it has been worth it. The digital preservation research team could focus on research issues and set up an active role in international projects, but with a firm practical basis and focus on implementable solutions.

The e-Depot department was a separate team in the Acquisitions and Processing division at the start, but is now growing and becoming more and more interlinked with the rest of the division. While all processes are becoming digital (eg. automatic metadata ingest and processing for printed publications), differences between the digital depot and the traditional depot are becoming smaller. The best example of this integration of separate processes is the automatic handling of publisher-submitted metadata. The e-Depot has been working with submitted metadata since the start, while metadata for the print collections is generated manually at the library (in case the metadata is not yet provided by others in the shared national catalogue). For some collections, KB is now developing import of metadata records, setting up a similar workflow for processing of both print- as digital collections.

The e-Depot department with its staff working with the newest digital procedures in an international environment now co-operates closely together with library staff with 'traditional' library skills in the area of acquisition and cataloguing. The e-Depot meant doing the same kind of work in a completely different way and brought people to the library with a new set of skills. And through the

interchange within and across divisions, people can learn from each other and get familiar with new digital processes.

But running the e-Depot at the KB also brought liveliness, an international atmosphere and a broader outlook on the information landscape, thus making the library a really attractive place to work.

Preservation research

The success of the KB e-Depot is built on two pillars: the operational digital archiving environment and a substantial investment in research. As mentioned before, the first practical results of the digital preservation R&D are now being implemented into the e-Depot infrastructure. This improves the quality of the system and the content to be ingested. Improvements are focused on the different phases of the process.
- delivery of objects
- workflow and management of workflow
- characterization of the objects to be stored
- collection management
- preservation management
- IT infrastructure

Implementation is organized in the program described above, that combines these new additions with the upscale of the loading and storage capacity of the system.

Newly developed preservation functionality is partly the result of extensive international collaboration. KB is an active participant in international projects to develop new tools and services. KB takes part in projects for two main reasons. First, KB is able to bring in Library specific knowledge and practical experience in digital archiving. Second, insight in new technologies is necessary to maintain the e-Depot infrastructure. The results of projects like Planets could be of direct use for the operational processes and infrastructure of the e-Depot.

Within Planets, the KB is responsible for leading the subproject Preservation Action.[4] Furthermore, the KB is participating in several other subprojects. As the Planets project is halfway now, it seems to be the right moment to evaluate the Planets output (in this case and more specifically the Preservation Action output) against the interest of an operational system like the e-Depot. Planets will deliver a sustainable framework to enable long-term preservation of digital content. Either the framework or the individual modules delivered by the project will be of direct use for the e-Depot.

PLANETS

Much has been written and said about the objectives of the Preservation and Long term Access through Networked Services, or Planets project. In short, the main goal of Planets is to increase Europe's ability to ensure long term access to its cultural and scientific heritage. Planets delivers preservation planning functionality enabling organizations to plan their preservation actions in a structured and controlled manner. To characterize digital objects, Planets develops methodologies, tools and services, while preservation action tools will be in place to migrate or emulate digital objects. A testbed is created for the objective evaluation of different protocols, tools, services and complete preservation plans. The Interoperability Framework will integrate these tools and services in a network.

For the KB, Planets means performing the R&D we had planned, but in the setting of a closely collaborating international team. Requirements for Planets tools and services are based on KB's practical experiences and future plans, but also aimed at developing a more general framework. Planets products should not be specific for one organization, but should offer a set of services for a large variety of institutions. In practice, for the KB as participant in Planets, this can cause some tension because resources go into developments that might not be directly implementable in the KB. At the same time these activities are necessary to create the overall framework.

The project Planets being two years on the way, we want to take a look at some of the project results and will evaluate what these products could mean for the further development of our e-Depot environment. Within the scope of this paper we restrict ourselves to products that are/will be delivered by the subproject Preservation Action only. This subproject is concerned with the creation of solutions to perform preservation actions. In other words, this subproject is responsible for making the tools available that are needed for rendering digital objects, either in a different format (migration tools), or in a different technical environment (emulation tools). Next to migration and emulation tools the subproject also includes the development of a Tool Registry and a variety of reports of a more strategic purpose. In the following we will discuss the products delivered by the subproject and the possible value for the e-Depot environment. Subsequently we will discuss the Preservation Action Blueprint, the GAP Analysis, the Tool Registry, the Preservation Action Tools on Emulation and Migration.

The Preservation Action Blueprint

One of the products of the PA sub-project is a Blueprint that can be used by any developer or supplier when developing new preservation action tools, - both migration and emulation. It provides a list of functional requirements that these types of tools should offer. It also presents the workflow that should be followed when incorporating newly developed tools into the Planets framework. This list of functional requirements for newly build, improved or

adapted PA tools will ensure not only a consistent behaviour but a consistent level of quality as well.

Since the Blueprint is very much aimed at guiding and stimulating future development, for now it is just of indirect value to the e-depot workflow. In future, it will guarantee a certain degree of quality of new PA tools. It can (and should…) be used by developers when building new preservation action tools.

GAP Analysis

Within the Planets project, the PA subproject is responsible for providing tools required to perform preservation actions. In order to do so, existing tools can be wrapped and made available within the Planets framework. If no tool for a certain action exists, new tools have to be provided for. To offer a choice of tools to be used for preservation actions, we first need to know which file formats are in use for long term archiving. This is what has been done in the project: we created a list of file formats based on information provided by 76 institutes from different countries. At this moment (August 2008) the list contains 121 used file formats but is still being expanded. By analyzing this inventory we will have a clear understanding of which preservation action exist and/or what tools are needed.

What we have found for the Blueprint is also true for the Gap analysis. There is a certain value for the e-depot, but again it is of indirect value for the e-Depot although it could constitute an important instrument in the future when combined with the tool registry.

Registry

The Planets Preservation Action Registry stores descriptive information about preservation action tools (and services, which are wrapped tools) and how and for what kind of actions to use them. In Planets PA registry, a preservation action tool is a software program that performs a specific action on a digital object to ensure the continued accessibility of this digital object. This action could result in a transformation of the object or a (re)creation of the technical environment required for rendering the object, or result in a combination of these two.

How tools and services could be used is described in a 'pathway'. A pathway is a predefined set of one or more preservation actions operating on a specific input file format and version and possibly (in the case of an 'actions on objects' tool) resulting in a specified output format. A pathway can include at least one or more preservation actions (and thus require at least one or more tools).

Of course, a registry which includes indications of both functionality and quality of preservation action tools contains a very usable overview for the e-Depot. The registry will be of direct use to deploy preservation action tools before or after ingest.

Migration tools

As more and more heterogeneous content will be presented to the e-Depot (think for example about the content that is generated by web archiving), the need for preservation actions becomes more important. One of the main digital preservation strategies is migration. Migration modifies a digital object in order to keep it accessible. There are three types of migration to distinguish. First, there is a type of migration that will take place in the ingest phase. This we call normalization. At the moment, a module for the e-Depot is in development that will convert text based publications that are not delivered in the PDF format, to PDF/A. There is a second type of migration that will be periodically used to execute batch migrations. This kind of migration will be used to prevent already stored digital objects to become obsolete. The third type is called migration on demand. A digital object will be temporarily migrated at the request of a user.

Migration tools, or tools for objects, are essential in the e-Depot environment. They play an important role in the pre-ingest phase, during the storage phase and eventually in the access phase. The results of the Planets project will offer a broader range of migration tools that will allow a digital archive like the e-Depot to perform migrations of a higher quality, including quality control.

Emulation tools

Another strategy to ensure the accessibility of digital documents is formed by emulation. Emulation tools, or tools for environments, change the technical environment in such a way that the original objects can be accessed. In Planets, a modular emulator is developed, based on earlier research and development of the KB. This emulator, named Dioscuri, is especially designed for digital preservation by being more durable and flexible than other emulators.[5] In the design, each hardware component is represented as a module in the emulator. A full emulator is created by combining all the modules.

Emulation tools could play a significant role in the accessibility of digital objects. For KB, emulation is as important as migration, because a growing group of digital collections (interactive, complex objects) cannot be migrated if the original does not work anymore due to their complexity. The development of the emulation tool within Planets is again an investment in the future.

Employability

Several Preservation Action products can be, directly or indirectly, employed within the e-Depot workflow. Obviously, within the Planets project many more significant tools and products with a potential value for the workflow in the e-Depot are developed. For example, in

the Preservation Planning part of Planets the planning tool PLATO is developed, while the characterization module PRONOM will be further developed in the Characterization subproject. However, within the scope of this paper it is impossible to describe all the (potential) valuable tools and products at some length.

Conclusions: Workflow improvement

Six year of operating a digital archive brought a lot of practical experience and knowledge. At the same time years of research started to pay off. The KB R&D department delivered a clear list of functionalities that now have to be implemented to ensure ongoing durability of ingested material. Because of this, the KB has a firm fundament to raise the level of the e-Depot environment and to extend the usability.

People from different departments (e-Depot, Digital Preservation and IT) are now in a program together to implement changes. And at the same time, research is moving forward, working on rendering tools like the emulator Dioscuri, which will be tested on the results of the KB's web archiving project.

The results of the Planets project as described above generate new tools which can be directly implemented in the e-Depot environment. Besides, these research activities create new views on preservation issues and the workflow of KB's e-Depot environment. Practical experience and knowledge from research projects give us a clear focus for further research. Current developments with the program setting up a new ingest process make us feel confident about the phase after that: when current research delivers results and will become implementable solutions for permanent accessibility.

References

[1] Erik Oltmans and Hilde van Wijngaarden, Digital Preservation in Practice: The e-Depot at the Koninklijke Bibliotheek', in: *VINE - The Journal of Information and Knowledge Management Systems*, Vol. 34 (1), pp. 21-26.
[2] Erik Oltmans and Adriaan Lemmen, The e-Depot at the National Library of the Netherlands. In: *Serials*, Vol. 19 (1), 2006, p. 63-67 and Els van Eijck van Heslinga,. 'SHAPING COURSE. The Development of the Strategy for the e-Depot of the Koninklijke Bibliotheek, National Library of the Netherlands, in a National and International Context' In: *New Technology of Library and Information Service (iPres2007)*, 2007.
[3] Information on the KB webarchive is to be found at: http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html
[4] The Planets website can be accessed at: http://www.planets-project.eu/. Information KB in Planets: http://www.kb.nl/hrd/dd/dd_projecten/projecten_planets-en.html
[5] Jeffrey van der Hoeven, Bram Lohman en Remco Verdegem, Emulation for Digital Preservation in Practice: The Results, *International Journal on Digital Curation (IJDC)*, Vol. 2 (2), 2007.