The National Library of the Netherlands in a Changing Information World

Paul Doorenbosch
Head Product and Service Development (R&D)
Koninklijke Bibliotheek, National Library of the Netherlands
POBox 90407, NL-2509 LK The Hague, The Netherlands
Email: paul.doorenbosch@kb.nl

Topics: infrastructure; digitization, services

Summary
The Koninklijke Bibliotheek (KB), the national library of the Netherlands, has restructured its complete data infrastructure during the last two years. The architecture is now based on international standards like DC(X), MPEG21-DIDL, OAI-PMH, SRU. The implementation of the basic components is completed and the larger collections are available via the new infrastructure.

This new architecture has many advantages, two of the most important are:
- that the (meta)data can be accessed from outside the KB, without knowledge about the KB’s infrastructure;
- that the entire architecture is much better maintainable because of its simple and standardized construction than the former, more fragmented infrastructure.

Due to large scale digitization projects, KB’s data will increase drastically over the next few years. The KB is therefore still working on the parallel restructuring of hardware (storage) and software (search engines, databases).

The KB’s next step will be to create new services on top of the new data infrastructure and to stimulate the use of our data in already existing services outside our own library.

The shift in focus from the traditional physical library activities to an increasingly virtual oriented presentation calls for a thorough reconsideration of the library’s position in the information chain. The renewal of the data infrastructure is part of this process, the purpose being to explore and facilitate new ways of interaction between the user and the information.

Introduction
The Koninklijke Bibliotheek (KB), the national library of the Netherlands manages a relatively large collection of printed materials, an electronic deposit-collection and precious collections of old, curious or unique documents and manuscripts. Apart from the later category, these collections can primarily be seen as pieces of information; the physical object itself is only of secondary importance.

Information in physical form limits access and use, compared to its digital sibling. Digitization is,

1 www.kb.nl
therefore, the best way to facilitate access to printed information for future users, and the best way to ensure that knowledge from the past can and will be used over time.

The KB is situated in the Hague, a city which does not have a university. This implies that the KB’s main user group, researchers, will have to travel to visit the library, mostly to gain access to our special collections. The general public, however, seeking information or interested in Dutch culture, history and society, will probably first visit a local, public library, seeking out the KB only as a last resort. Although the KB puts much effort into making the library an attractive place of study and meeting place for researchers, users will only visit us if absolutely necessary, for instance to look up copies of our deposit collection or other the unique parts of the collection.

This trend has been growing over the last decades, even before the internet took such an enormous flight. Before, interlibrary loan (ILL) and document delivery were the means for the KB to bring the information it collected to the remote user. Digitization has, of course, largely replaced ILL and other forms of physical information transport. Some major internet companies are currently undertaking enormous digitization activities, on their own or in co-operation with libraries. An important task for university and national libraries from minor cultures is to guard that their scientific and cultural heritage will not be overlooked but remains incorporated in the global flow of digital information. Although the Netherlands is a country with a very high internet penetration, Dutch will always remain a niche language compared to countries with English or Chinese as their mother tongue. Yet it is an interesting niche. Dutch history, culture and science are closely intertwined with the history and culture of many other countries in Europe and the rest of the world. The same is true for most of the smaller countries in Europe. Globalisation is nothing new, it is of all times. Only the scale differs.

The KB not only digitizes information to serve its own user groups. A second goal is to ensure that our information is part of a European and even a global information environment. Our mission is to grant any citizen and researcher\(^2\) free access to the available information, restricted only by possible demands of the author, and, moreover, to provide him with the opportunity to manage and use this information in any way he sees fit. This calls for an open system and a set of simple standards for our data. This may seem obvious, but the process of change in this hybrid world of independent parties, different strategic goals and sometimes long-lasting customs has proven to be no sinecure.

\(^2\) An interesting report about the value of digital cultural heritage resources for the humanities, with recommendations for the building of a common infrastructure was published by the American Council of Learned Societies (Our cultural Commonwealth) in July 2006: www.acls.org/cyberinfrastructure/acls.ci.report.pdf

Renewal of the data infrastructure

Change takes place by convincing people in charge, by gaining support for ideas, and by finding early adaptors. But without a working example or prototype it is hard to show possibilities and potentials. In 2001 the KB joined the European Commission funded TEL-project, which has now become The European Library and may become part of the European Digital Library in the near future. The European Library, a CENL-product, is developed and hosted by the KB and as such the KB was one of its founding partners. By being a co-creator of TEL, the small group of early adaptors in the KB learned how to open up data in a standardized way and how to bring together data from different repositories. But even in TEL only part of the ideas how to best serve information-seekers could be realised. Many of these ideas have become generally accepted only a few years later, especially in Web 2.0 applications. The KB, forerunner in TEL, stored its metadata in several different legacy systems in 2004. It was able to provide data to The European Library, but in a rather inefficient and performance-reducing way. There was an XML-database in place for parts of the data, but with several different schemas. Besides this hybrid XML database data was also stored in Access, DBtext and other database systems. Such a fragmented infrastructure is not uncommon to most libraries, archives and museums.

In 2005 the KB started a project to renew the complete data architecture, based on the same philosophy as the TEL-architecture. The infrastructure was to be service oriented, which means it consists of relatively independent elements that communicate by standard protocols. There were two main reasons for rebuilding the infrastructure. The first was the accumulating maintenance effort needed for all those different infrastructures. The other and most important reason was the need for a data infrastructure on which a service infrastructure could be developed needed for the future of the national library as part of the world-wide information area.

The new infrastructure is there to support services and end users, not for administrative or managerial purposes. The data itself is created and maintained in databases designed for that special purpose, although we try to limit the amount of local source databases.

The architecture for this new data infrastructure is based on the following principles:

- if available, a standard is used; preferably standards that are commonly accepted by more parties then the library world alone;

---

3 Relevant documents are published on: research.kb.nl (only part of the information is in English)
4 www.theeuropeanlibrary.org
6 Veen, Theo van. Renewing the Information Infrastructure of the Koninklijke Bibliotheek. In: Dlib, March 2005: www.dlib.org/dlib/march05/vanveen/03vanveen.html

- a simple standard is preferred above a complex standard, and the standards chosen should be flexible and extendable, without compromising the standard itself; so in principle any desired functionality can be supported;
- all metadata in the KB must be searchable in one action;
- access to data and services is via http to facilitate web 2.0 functionality;
- any end user service (website) must have the opportunity to use data originally created for an other service; any service from an other party must be able to use the KB-data in the same way we use it ourselves;
- any party outside the KB wishing to make use of our data must be able to do so without knowledge of the internal organisation of the data and without requiring support from our IT department other than the URL to access the SRU-service for search and retrieval.

These are the standards currently in use:
- Dublin Core terms as data-format, expressed in XML, with extensions for extra functionality
- MPEG21 DIDL for the description of complex objects (like digitized books)
- SRU for search and retrieval
- OAI-MPH for metadata-exchange by harvesting


---

The new infrastructure currently contains:

- the *general catalogue*: metadata for books, journals, newspapers etc. are created and maintained in a nation-wide library cataloguing system and on nearly real-time base converted by style sheet transformation to DCX, stored and indexed;

- the *e-Depot*: metadata delivered by publishers is automatically converted by style sheet to DCX, stored and indexed;

- *Memory of the Netherlands* (the national digitization programme for museums, archives and libraries): metadata is converted to DCX by hand, stored and indexed; structural metadata is created semi-automatically;

- A number of large *digitization-projects* the KB has started, will result in nearly 20 million full-text records within 7 to 10 years: metadata is created directly in DCX and structural metadata in MPEG21. Some of the projects already use the new system to store their metadata.

The total amount of metadata in the database today is 13 million records. The new infrastructure must be highly scalable to store and process all the newly created data. The central index is an important link in the process of granting access to this data. The KB is currently trying to answer the question whether it is wisest to upgrade the present search engine or to install a new one. This is an urgent matter, because the performance of the search engine is expected by users to match the Google-performance. The KB’s search engine has to process not only millions of metadata records but also millions of full text records. It must provide faceted searching and has to support complex queries. This is the real challenge: a library nowadays needs to perform traditional tasks like lending, conservation, providing information by telephone etc., but next to this, it needs to support a variety of services on the internet, built by ourselves or others. To execute both activities is the most challenging task for the coming decade, especially for a library that is not directly connected to a physical user community like a university.

The use of Dublin Core to offer access to metadata ran into some resistance within the KB. A library community prefers the traditional MARC-like metadata schemas, which are much more detailed, to the 'limited' Dublin Core. By adding extensions to Dublin Core, however, it is possible to provide the richness of MARC to support all the needed functionalities. Moreover, a multi-format database is used to store descriptions in several formats. This allows for the use of different formats in specific environments. The main advantage of the use of Dublin Core as a record schema is that it is seen in several sectors as a useful format for data storage, which enhances interoperability between sectors and specific data files.

Organisational change

To develop and execute policy for a library part of an information community that is becoming increasingly digital takes different skills and other attitudes than are necessary for a physical library. The research and development part of digitization and the activities necessary to bring information online will change the organisation and the kind of personnel employed in a library. There will be a growing need for people with a background in information science, especially in combination with skills like (in our case) a scientific study in Dutch history, language technology, knowledge of user behaviour on the internet, or internet marketing. It will be necessary to re-evaluate the position of the library in the information chain. Will e-publications remain the domain of publishers, or partly shift to universities and other communities? Libraries could focus on digitization, storage and providing tools and methods for access. Building dedicated interfaces, combining tools and presenting services to end users are activities that could be left to user communities or to other parties closer to the public, like museums, public libraries, news sites, and individuals. To quote Theo van Veen, data infrastructure architect at the KB:

"Enrichment of information by means of the integration of data originating from different sources under the user’s control and, if needed, on the user’s desktop, is the key issue here.”

This does not imply a library should never act as a publisher or as a vendor of information, but it does say something about the intended focus.

This shift in organisation and skills is currently occurring throughout the KB, but is most visible in the vast expansion of the R&D Division. This division has grown from 30 to nearly 80 people over the last two years, comprising a fourth of the total amount of personnel at the KB. In this division we conduct research in the field of digital preservation to ensure permanent access to our own digital resources, but also to attribute to the world-wide quest for knowledge in this area. Furthermore, the KB researches the digitization workflow, image-quality, and specifications on text-digitization. The KB is partner in CATCH, a scientific research programme in which heritage institutions and

---

8 Veen, Theo van. Renewing … [see above]
9 CATCH (A computer science research programme for Continuous Access to Cultural Heritage): […] Within the framework of this research programme scientifically relevant methods are being developed to acquire new fundamental and applied knowledge about IT-bases solutions. […] The CATCH research strategy concentrates on three research themes: Semantic Interoperability through metadata, Knowledge enrichment through automated analyses and Personalisation through presentation. It focuses on the development of tools and methods to speed up the back office processes, i.e. tools and methods that will enable collection managers of the cultural heritage institutions to do more in less time and acquire higher qualities.
In the CATCH programme, interaction between computer science and the cultural heritage is of great importance. […] To support a vivid interaction between the researchers and the prospective users, research teams are physically located in the cultural heritage institutions. […] (from the fact sheet brochure, 2005)
The programme is funded by the Netherlands Organisation for Scientific Research Councils for Physical Sciences and Humanities, and runs from 2005-2010.
information science research groups at Dutch universities work together to develop new tools and methods to improve interoperability between heritage collections and to improve searchability and accessibility of cultural artefacts. As mentioned earlier, the KB is partner in TEL and EDL projects because of the conviction that an integral information space is more than a theoretical possibility. To put theory into practice the KB has realised this new data infrastructure to accommodate its own data flow based on accepted standards.

**Digitization of information**

The KB develops and manages large digitization and preservation programmes, not only to facilitate the KB collection, but with a broader view to further development in the entire field of science and cultural heritage. The KB’s digitization activities become increasingly important and have a growing impact on our resources. Most of the projects are executed with the help of additional grants, but maintenance and support of the infrastructure is funded by the yearly government grant. The amount of storage is expressed in hundreds of terabytes and in a few years a technical infrastructure must be in place that can facilitate these enormous amounts of information.

The following information will be stored at the KB:

The metadata and images of the *e-Depot*\(^\text{10}\); the system for long term preservation and permanent access, already containing over 10 millions of articles from STM-publishers like Elsevier, Springer, Sage, Oxford University Press, Taylor and Francis and others. New publications of these publishers are continually added. The same system will be used for *web archiving* purposes. Currently, a pilot project is being executed, in which the KB has recently decided to archive (parts) of Dutch websites on a regular basis.

The *archival images* (TIFF’s or JPEG2000) produced by the KB’s digitization projects will also be stored in the e-Depot and in the near future the KB will explore the possibilities of expanding the e-Depot to create an archival storage service for Dutch heritage institutions.

**Digitization of the Dutch Parliamentary Papers.** This project (2005-2011) will digitize over 2.5 millions of pages, the oldest of which date back to 1814. These documents, representing Dutch democracy and history, are made available in OCR-ed full-text of very high quality. This is necessary because the website will not only serve the general public, but is also designed to be used by members of the parliament themselves, by researchers, lawyers and by other parties that will only use the information if they can completely rely on it.

**Databank Digital Newspapers.** The project (2007-2010) will deliver a databank that contains 8 millions pages, full-text searchable. Current OCR-techniques yield poor results for newspapers because of bound volumes, deteriorating paper and bad printing, though most results will at least

---

\(^{10}\) E-Depot: [www.kb.nl/dnp/e-depot/dm/dm-en.html](http://www.kb.nl/dnp/e-depot/dm/dm-en.html)

---

Research subjects are: OCR of handwriting, alignments of thesauri, database enrichment, music information retrieval, collections integration, semantic annotation and searching multimedia oral history content, enrichment by automated analyses of images and video, automated visual analyses, personalisation in art presentation.
satisfy the needs of the general public. Currently we heavily invest in R&D to improve the quality of OCR software and solutions to meet the requirements of the different scientific communities like historians and linguists.\(^\text{11}\)

*Dutch Books from the 18\(^{th}\) century.* This project (2007-2009) will deliver 1.2 million digitized pages. Partners in this project are the university libraries of Leiden and Amsterdam. Different typesetting and different languages are some of the challenges in making these books available in full text.

*Printed Radio News Bulletins.* This project (2007-2008) contains 1.8 million pages, produced with a typewriter between 1937 and 1989. The material also contains marks and remarks from the radio presenter aiding him to read the text correctly. The texts are information dense. The material would gain value if accompanied by news pictures. However, these typed pages are often hard to decipher, even by the human eye.

*Preservation scanning:* national preservation programme Metamorfoze\(^\text{12}\). In one or two years time, the Dutch national programme for the conservation of printed material is expected to shift from microfilm carriers to digital images or texts for preservation purposes. Digitization techniques and digital preservation have improved to the extent that they can match preservation standards for some types of materials. Besides that, the great advantage of digital copies as opposed to microfilm copies is accessibility. We expect to produce ten million pages every year during the coming decade.

*Memory of the Netherlands\(^\text{13}\):* this national digitization programme for museums, archives and libraries has resulted in some half a million digital objects and approximately 350 hours of audio and video. Each object is stored as a master image in TIFF as well as an access copy. The programme started out by providing hands-on aid in organising (and financing) the digitization process at cultural heritage institutions. Two years ago the position of *Memory of the Netherlands* as a national programme was reconsidered and the focus shifted to developing services with the digital information (for instance for educational purposes). Starting fall this year, further development of the programme will be stimulated by handing out grants (again) to finance local digitization projects and related services, continuous dissemination of knowledge on digitization methods and standards, and by

\(^{11}\) KB is co-initiator of IMPACT (Improving Access to Text), proposal under the EU 7\(^{th}\) Framework Programme. […] With the i2010 vision of a European Digital Library, the EU has launched an ambitious plan for large scale digitisation projects transforming Europe's printed heritage into digitally available resources. However lack of institutional knowledge and expertise slows down the pace with which this vision can be realised. The state-of-the-art in OCR performance and machine understanding of the original document is inadequate, especially for historical important material with archaic fonts and spellings, newspapers with complex layouts, bound volumes, microfilm, typescript.

[…\] The project will let them [i.e. the partners] share their know-how and best practices, develop innovative tools to enhance the capabilities of OCR engines and the accessibility of digitised text and lay down the foundation for mass-digitisation programmes that will take place over the next decade. This project will facilitate a more collaborative approach to mass-digitisation. […]

(from the summary of the project proposal, 2007)


\(^{13}\) Het Geheugen van Nederland: [www.geheugenvannederland.nl/index_en.html](http://www.geheugenvannederland.nl/index_en.html) (website both in Dutch and English)
making the Memory infrastructure available to ready digitized collections from other parties. The KB continues to promote the digitization of (large) collections in stead of individual objects.

**Education**

Although in the past scattered effort was put into the development of digital services for secondary and university education, the Memory of the Netherlands-programme is the KB’s first successful coherent approach in this area. Secondary education is highly fragmented and therefore hard to reach in an efficient way. A few commercial vendors of educational materials control the market for schoolbooks and educational methods and they are reluctant to share. When the Memory of the Netherlands was started there were only a few, mostly commercial, digital educational applications. Publishers were hesitant to develop these applications due to high costs and a short expected life cycle. The government that provided our grant, however, demanded that each of the published collections within the programme should be accompanied by an educational application. The goal was to stimulate other parties to do the same. Nowadays many publishers have incorporated electronic teaching aids in their learning methods. This coincides with the fact that infrastructure at schools has much improved and teachers have become more familiar with new electronic approaches. The Memory of the Netherlands has changed its approach from providing individual educational applications to adding educational metadata to individual objects and collection descriptions. This enables teachers, students and developers to apply the digitized objects any way they see fit.

Despite the relative success of the educational part of the Memory website, the KB will keep a relatively low profile in this kind of activities. The KB aspires to provide information and objects to research communities and the general public. We do not have the necessary expertise to provide context with every piece of information or make content based applications, with exception of a few specific fields, such as book science or actual events. The KB has therefore chosen to focus on providing optimal access to its data by means of the open and simple standardized infrastructure, and to leave creation of context to others. However, the KB can do more to facilitate users and to develop tools that enhance access and integration. One of the ways to do so is to develop standardized web services.

**Webservices**: a future scenario

It is likely that users may wish functionality in interfaces they have seen elsewhere and that is available as a service. To enable users or providers to choose and combine services a new paradigm is needed. If services were to use standard protocols, use standardized data as input, provide standardized data as output and were described semantically in a standardized format, it would be possible to

---

14 Veen, Theo van. *Serving Services in WEB 2.0*. In: Ariadne, April 2006: www.ariadne.ac.uk/issue47/vanveen/
develop a common service infrastructure. It would no longer be of any importance who develops and maintains a service, because it would use standardized metadata as input and would have standardized results as output, which could consequently be used as input for other services. Nowadays it is already possible to use standard metadata elements to trigger a service elsewhere; an element recognised as an ISBN can be used to find the same item in OCLC's Worldcat or 'amazon.com'; a subject heading can trigger thesaurus services from several parties and various knowledge fields. A provider would present different services fitting context and user behaviour. The individual user would then be able to combine different services and sources in a personalised webpage. This would lead to an information infrastructure that gives users complete control over their own working environment by using data and services provided by others.

Concluding remarks

The KB’s physical library department will remain an important service for the foreseeable future, although even in this sector evolutions in the information society will lead to many changes. The main change, however, will affect the library organisation behind the scenes. At the moment, the emphasis lies on the digitization of information and, above all, the development of ways to provide access through internet services and the development of methods and tools to facilitate the use of the digitized information. This requires a continuous adjustment in vision, policy and attitude, next to organizational adjustments. And the limited financial means always require sharp focus. World-wide access to information seems to remain the prerogative of commercial search engines, so to small players like libraries it is essential to focus on their strengths in stead of desiring to compete with these large search engines. These strengths can be found in the complementary activities the large parties leave unexplored. Herein lays the added value to the user. It is essential to make use of open and standardized ways to make information accessible. In this regard, the KB has made an important step forward with the introduction of its new information infrastructure. The logical next step will be to develop a standardized service architecture in co-operation with other parties active in the information area.  

Acknowledgment: I would like to thank Hans Jansen, Theo van Veen and Pauline Bodde for their comments on the draft of this paper.