LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich. Printed in Germany

The Dutch Deposit of Electronic Publications (DNEP) - 1995-2000

by LEX SIJTSMA

1 INTRODUCTION

In 1993 the Internet took off with the introduction of HTML and the first browser (Mosaic[1]). Two years later, in 1995, the Koninklijke Bibliotheek decided to start a series of experiments and projects which would lead to a deposit system for Dutch Electronic Publications. In the same year the Koninklijke Bibliotheek made a policy decision to include electronic material into its deposit.

That marked the start of the Dutch Deposit for Electronic Publications (DNEP[2]). Both as an operational service and at the same time as a test-bed for research into digital archiving.

The Koninklijke Bibliotheek has a staff of 254.5 FTE[3]. The ICT-department has 15 FTE (about 6%).The ICT-department is responsible for the systems management of the operational systems, for the support of the end-users and for research and development. Apart from the R&D done in the ICT-department the Koninklijke Bibliotheek also has a department of library research (see the website[4] of the Koninklijke Bibliotheek for more information).

In the first few years a lot of experiments were done. Various hardware and software was tested and research was done on issues such as metadata, the number of electronic publications available, how to process them in the library etc. At the end of 1998 the Koninklijke Bibliotheek decided that the time was ripe to make the next step. This was the implementation of the DNEP on a large scale and as part of the normal workflow inside the library departments.

Early 1999 a Request For Information (RFI) was sent out to selected companies. This was done to establish whether the functionalities the KB deemed

neccessary for a digital deposit were available in the marketplace. On the basis of the positive outcome of this Request a European Tender was started. The Koninklijke Bibliotheek developed a process model for a digital deposit as well as detailed functional requirements for such a system. A supplier, IBM, was selected. At the moment (early June 2000) talks about the implementation are in progress. The project will start in the summer of start of autumn of 2000 and will take 24 months. The Dutch government has acknowledged that the establishment of the Dutch electronic bibliography and the DNEP itself are indeed tasks of the Koninklijke Bibliotheek. We expect that appropriate structural funding to supports these tasks will become available. In this paper an overview is given of what the KB has done since 1995 up till now and how this has led to the implementation-project that is about to start.

2 FIRST STEPS

In 1996 the KB came into contact with AT&T. This company earlier was selected as a result of a tender procedure to become one of the suppliers for the new local area network (LAN) of the Koninklijke Bibliotheek.

A strategic cooperation was formed for the Advanced Information Workplace (AIW) concept of the Koninklijke Bibliotheek. One of the results of this cooperation was the installation of the AT&T product called *Rightpages* at the Koninklijke Bibliotheek.

This was a system developed by one of AT&T's research labs to maintain subscriptions on magazines. These magazines were scanned, stored and made available through *Rightpages* to all the employees of the laboratory.

The Koninklijke Bibliotheek also had contacts with two major Dutch publishers: Elsevier Science and Kluwer Academic Publishers. Both agreed to take part in the experiment and supply content in the form of electronic versions of scientific journals. In return they were kept very well informed on the status of the project.

The KB together with AT&T developed the technical infrastructure, procedures, software, organisation neccessary for the processing of the material from the publishers. The publisher in return got feedback from the KB. With this feedback they could enhance in-house procedures. This kind of close cooperation with publishers has worked very well and still forms an important part of the workflow of the DNEP. In a test the material was also being made available to a limited set of users.

From the outset on more than just the ICT-department was involved in this project. The Central Catalog Department of the Koninklijke Bibliotheek was responsible for the loading of all the articles. The ICT-department delivered the technical expertise needed to keep everything running, but the content expertise was brought into the project by the relevant departments.

Later on the Delft Technical University signed a contract with Elsevier Science allowing them to search and view articles from journals for which they had a subscription. The articles were stored in the DNEP. The Koninklijke Bibliotheek functioned as hosting provider for the information. This was a new role for the library.

The cooperation with AT&T stopped when the part of AT&T the Koninklijke Bibliotheek dealt with, seized to exist. It was taken over by Lucent Technologies. However the DNEP itself continued to evolve.

3 Other Issues

In the beginning the focus had been primarily the technical side of the digital deposit. However, we also wanted to know more about things like:

- What metadata you need to describe the content of an electronic publication?
- What is the workflow of an electronic publication?
- What skills do I need as a cataloguer?
- What kind of organisation do I need to process electronic publications?
- What kind of production can we expect from the Dutch publishers in the coming years?

To investigate these issues more thoroughly the Koninklijke Bibliotheek started a project, with financial support of IWI[5]. This project was called DNEP-IWI[6].

In the project the following things were done:

- In 1997 a market survey was done by a Dutch bureau (NBBI). It turned out that already in 1997 there were quite a number of publications that should be added to the deposit.
- A report detailing selection criteria was made.

- In 1997 research has been done to the identification and description of electronic publications.
- In 1997 the costs and possibilities of long-term storage were investigated.
- A testbed of 100 electronic publications was followed on their way through the organisation: from the mailroom into the archive. Of these 100 documents, 50 were so-called off-line, that is: cd-rom's. The other 50 were on-line documents, they were selected by librarians and harvested from the Internet.
- In february 1998 an information session for publishers was organised. This was very well received.
- Research has been done into a number of legal aspects concerning digital documents such as copyright and all kinds of arrangements and contracts with publishers and organisations of publishers.

4 TOOLS UNDER INVESTIGATION

The Koninklijke Bibliotheek wanted to continue what it had started with *Rightpages*, but it needed a new tool. One reason for this was the termination of the co-operation with AT&T, another, even more important one, was that *Rightpages* did not have enough potential to support a deposit with tens of thousands of journal titles. After all it was designed to make available the subscribed journals in a research lab, not to run a complete large scale digital deposit.

4.1 Essentials for a Deposit

For the Koninklijke Bibliotheek two things have been essential. We believe they hold for any archive, be it digital or not. They are:

1. High quality long-term storage. The act of just storing a file throughout a large number of years is not a trivial undertaking. Apart from maintaining the integrity of a file, files also need to be transferred from one medium to another because physical media deteriorate quite fast or are replaced by better and cheaper ones. Another important issue here is the management of large numbers of files itself.

2. Preservation of the contents. Access to the document is essential for the archiving itself. One might say that without access archiving is not rele-

vant. A problem with preservation is that at the present time there exists no standard method of doing this.

Another important thing we realised through the years is the scope of the deposit. More and more we trimmed down the functionality of the deposit to that comparable of a warehouse. After all, we already had various other systems in place for acquisition, cataloguing, searching and retrieving the information. The deposit had to be able to exchange information with all of these systems and should also be extensible to new ones.

4.2 Digital Library

The Koninklijke Bibliotheek looked around and tried some tools but we were not satisfied with them. We needed a tool that had the core functionality for a digital deposit but at the same time was flexible enough to adapt it to our (changing) needs and current systems. In the end the Koninklijke Bibliotheek chose for *Digital Library*[7], a product from IBM.

Digital Library has a toolbox approach to the problem of long-term archiving. With it you get a robust and scalable high quality storage system for digital assets. Because of the toolbox-approach you can adapt this to your own situation.

The Koninklijke Bibliotheek has built on top of *Digital Library* loader programs for the journals of Elsevier and Kluwer (see figure 1) and a search and retrieve-tool (see figure 2).

This combination of tools developed by the Koninklijke Bibliotheek itself and the use of the IBM-product constitutes the digital deposit of the Koninklijke Bibliotheek and hence is called ‚KB Digital Library'.

Working with a product like *Digital Library* has a steep learning curve. It offers a lot of flexibility but it takes quite a while to master the product. Because of the flexibility it is quite complex. The Koninklijke Bibliotheek co-operated with IBM very much in this process.

We have the *KB Digital Library* now operational since 1998. The Koninklijke Bibliotheek has both a production-environment and a test-environment. The production-environment consists of the following configuration (see table 1).

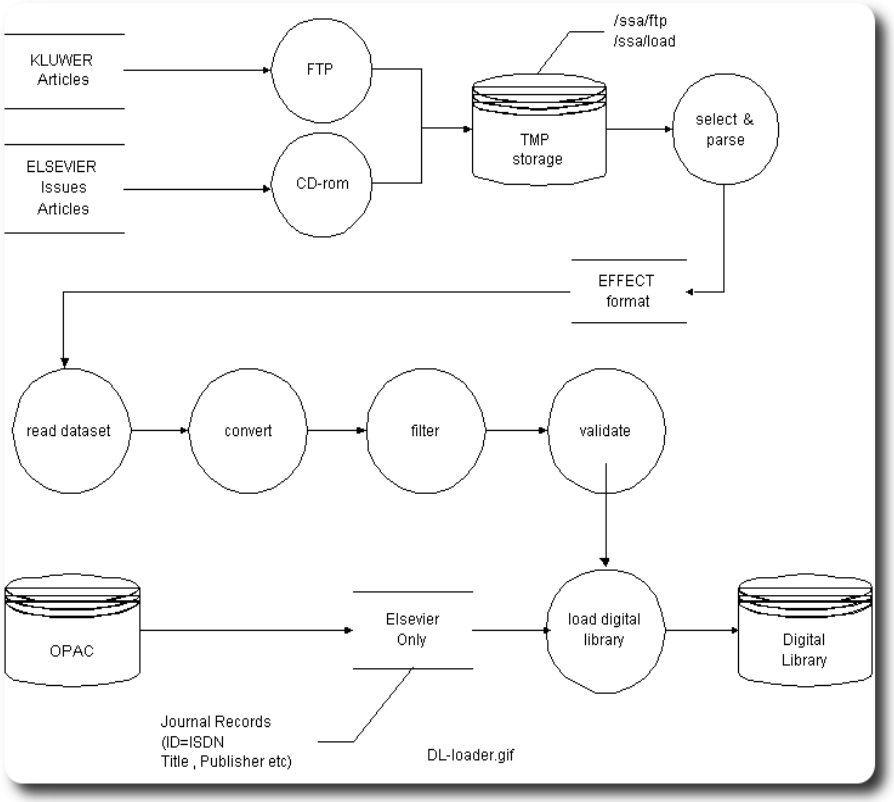

Figure 1: Flowchart for loading articles into DNEP

Computer	IBM RS/6000 SP 3 nodes with 2 proc. each
Internal memory	1280 Mb
Magnetic Disk Storage	108 Gb
Optical Disk Storage	450 Gb
Tape-unit	110 Gb
Operating System	IBM AIX
Storage Manager	IBM Tivoli
Deposit Program	IBM Digital Library + KB-additions

Table 1: Configuration of DNEP

In the production environment on June 1st 2000 the following number of articles were stored (see table 2). In May 2000 (one month) the deposit had grown with more than 10,000 articles. This is a growing figure.

Publisher	Journals	Articles
Elsevier Science	351	323,454
Kluwer Academic Publ.	516	67,439

Table 2: Journals in DNEP at June 2000

The *KB Digital Library* is also used to store other electronic items such as digitised books and images, harvested webpages etc. The Koninklijke Bibliotheek is planning to do experiments in harvesting on a regular basis the Dutch part of the Internet. But this is outside of the scope of this paper.

What is essential in this is that the Koninklijke Bibliotheek wants to have one technical infrastructure that can support all these functions. We do not want separate systems for the DNEP, the harvested webpages etc. We can do this because of the tight scoping we have applied to the deposit (see 4.1). Its prime function is high quality storage (and preservation) of digital objects. All functions that operate on these objects, such as search and retrieve, document delivery, title description etc. are done outside of the deposit (see also the process model in figure 6).

At the moment work is under way to add another major publisher to the DNEP. This is SDU, the former Dutch State Printing Office. Because we have developed a generic loader program, it is relatively easy to add more publishers without much work.

Off-line electronic publications are not stored in the *Digital Library* system. For the time being they are stored in our warehouse. We plan to add them into the next version of the deposit (see 6). The number of stored off-line publications was about 3,000 at the end of 1999.

The bibliographical metadata are stored in our current library system from Pica[8]. The DNEP-identification is added to the title-description. This constitutes the link between the library system and the deposit.

Access to the deposit is through the regular OPAC, KB-CAT (the search-interface of the Koninklijke Bibliotheek itself) and the interface for journals (see figure 2).

Figure 2: Searching for Articles in DNEP

5 REQUEST FOR INFORMATION

At the end of 1998 the Koninklijke Bibliotheek decided that it had enough experience gathered in the previous years to successfully set up a digital deposit on a production scale.

This meant that the deposit had to operate within stricter constraints:

- It should be able to process not only deposited electronic publications but also material that is being kept for the host-function, web archives of the Dutch part of the Internet and digitised material of the library itself. This was different than all previous experiments that were geared to only one function. Support for multiple functions means added complexity.
- Process and store a large and growing number of documents. Storage space must be able to expand indefinitely. We expected the following numbers for the next 3 years (see table 3). Storage should also be

managed transparently. That is you should be able to mix various types of media and view them as one large virtual storage space.

Function	2000	2001	2002
Deposit	3	12	55
Host	1	10	45
Web Archive	1	4	20
Digitising	6	45	200
Other	1	4	20

Table 3: Estimated Storage Needs (Terabytes)

- Disasterproof. Since the archive also has a last-resort function, it is absolutely essential that information does not get lost. Either because of an accident or deliberate.
- Digital Preservation. Not only should the deposit provide high quality storage facilities but also have functionalities to keep documents readable in the most broadest terms. Storage and availability are according to the Koninklijke Bibliotheek essential prerequisites for any deposit (be it electronic or not).
- The handling of digital publications should be made part of the regular work of the relevant departments (acquisition, cataloguing, public services etc.) at the KB. An important design criterium was that the Koninklijke Bibliotheek did not want to set up a separate department for the handling of electronic publications. Instead the current departments for acquisition and cataloguing should modify their procedures to incorporate electronic publications. This was a very important principle.
- The deposit should be more integrated into already existing systems at the Koninklijke Bibliotheek. This includes the Pica-system that is used for a.o. cataloguing.

However we were not at all sure if this all was possible, that is: if the needed technology was available. We felt pretty sure that in a matter of years the problem of long-term preservation of data would make itself felt outside ALM[9]-institutions, but we were not sure at all how things were at the present (early 1999).

We started off by writing a request for information (RFI[10]). This is a document in which you state your intentions and ask a number of selected suppliers (12 in our case) if they have a solution available.

The writing of such a document has the added benefit of forcing you to think really hard about the problems at hand and to be quite exact. It also functions as a means of starting discussions within the organisation.

In the RFI we described the current situation at the KB, the workflow of the DNEP and a number of specific questions to the suppliers. To make it sure that we were free to do as we like after the RFI, we payed all the suppliers that sent in reports a fee of EURO 500. With this we made the production of the reports just a business transaction and not an - implicit- part of the sale to be. All suppliers were made very clear that they still had to take part in the official tender procedure[11] and that they had no special privileges. Of the 12 suppliers we invited to respond, 5 did so by writing a report with detailed answers. They also made use of the possibility to give a presentation. The overall conclusion was that the state of the art was sufficient to implement a digital deposit. One exception was made for the aspect of preservation. There are no production-ready solutions available at the moment. The costs of building such a system would be somewhere between 3 million and 5.5 million EURO's.

During the whole process we tried very hard to make sure that the suppliers did have a firm and clear understanding of the problem. This is very important because otherwise you might end up with answers on questions you did not ask. We also tried to make the whole process as transparent as possible.

We did this by giving presentations in which we stated our view of the problem and what we were looking for in a solution. We held Q&A sessions and also organised a tour of the library. We finally created a supplier-oriented website at <http://www.kb.nl/dea> (see also figure 3). All the information plus all kinds of background-papers etc. were available from this site. On a password protected page we put up the names and addresses of all the suppliers. They were in a position to contact each other and discuss things.

This approach has proven to be very successful. We got very positive feedback from the suppliers. We continued doing this during the tender procedure proper (see figure 4).

The website is frozen at the moment but it will be rejuvenated at the start of the implementation. At that time, however, the focus will no longer be the suppliers, instead we will focus on a more general audience, for instance employees of the Koninklijke Bibliotheek, other stakeholders (such as publishers, other libraries) and interested outsiders.

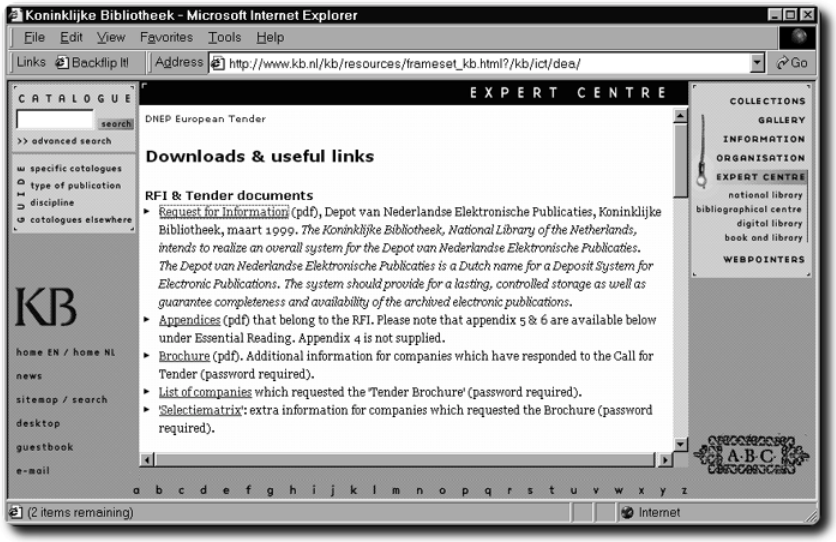

Figure 3: Downloads on DEA website

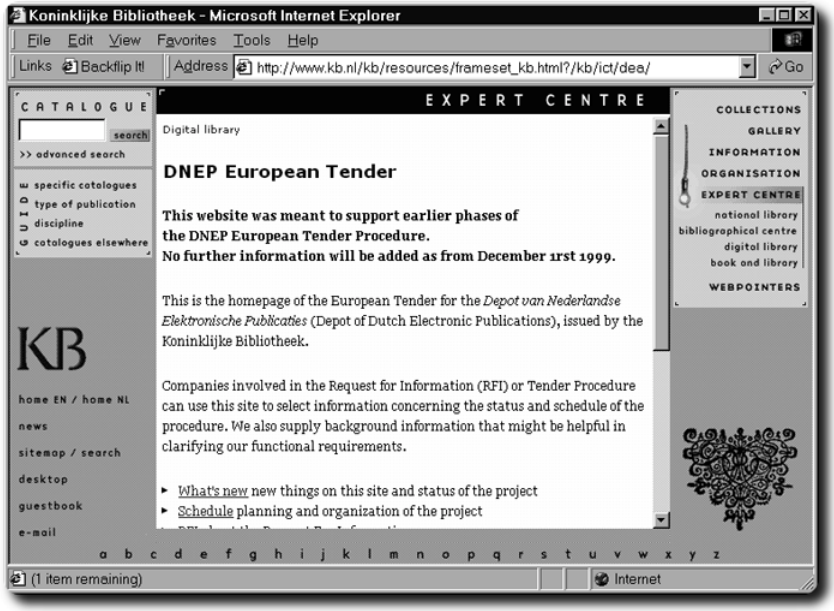

Figure 4: Homepage Tender Project

6 EUROPEAN TENDER

The results of the RFI made it clear to the Koninklijke Bibliotheek that the creation of a large digital deposit could be done. Although for the preservation part no solution could be provided at the present time, we felt sure that it would be the best thing to already start working.

We compared the lack of solutions for digital preservation with the situation on preservation for printed documents. Also in that area there exists no full-proof method that can be applied to all documents. Instead, progress is made step by step. We expect the same thing to happen with digital documents.

6.1 The Tender Procedure in General

The European Tender procedure is required for public organisations in the EU for all purchases of goods and services that exceed an amount of approx. EURO 200,000.

The European Tender procedure consists of a number of steps for selecting one or more suppliers. It takes about 6-9 months to go through all the steps. Some people think such procedures are only a hassle, but we have very positive experiences with it, because it forces you to think really hard about what it is you actually want and how you want to achieve that. You also have to rationalize your decisions very thoroughly. By doing things step by step it also functions more or less as a checklist, so you can be sure that you do not forget important things.

Although the Koninklijke Bibliotheek had done some considerable investments in software and hardware from IBM (see 4), this did not automatically mean that IBM had an advantage in the tender. We started the tender with an open view and were prepared to stick to whatever supplier had the most favourable offer. We think this is an essential attitude for a successful tender.

There are a number of variants for the procedure. We used a so-called closed procedure in which the Koninklijke Bibliotheek did a pre-selection of the suppliers.

The procedure itself is public and very transparent. For instance all the criteria you use for selecting or rejecting a supplier have to be known and published beforehand. All decisions you make during the process are subject to an audit afterwards. This forces you to do things very precisely.

First we published our intentions for tendering and asked companies to react. From the ones who did, we selected 5 and asked them to make us an offer. 4 of them did so. Of these 4, 1 was selected.

On September 1st, 1999 we started the procedure by sending a request to the EU. This request was published by the EU on September 11th ,1999. Out of the first selection came 5 suppliers who were asked to make a formal bid. The letter was sent out at December 3rd, 1999. We wanted to have a response at February 14th, 2000.

For this bid we had made a long list of functional requirements of the new system based on the model in figure 6. For each of the identified processes we created a list of yes/no questions to be filled out by the supplier. We created this list with the help of all the relevant departments inside the library (see figure 5 for an example of how the Call For Tender document looks).

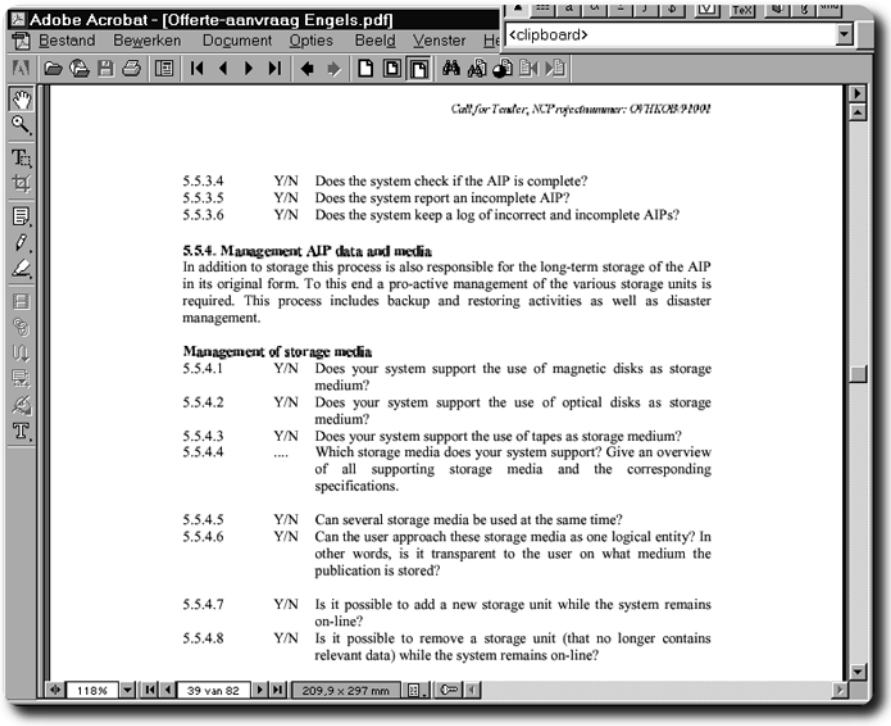

Figure 5: DNEP Call For Tender

We had made public a list of criteria we would use to grade the bids. These criteria were sent to the companies as part of the bid request. The bids were graded by a group of about 15 people from all relevant departments of the library. This was the same approach we used for the RFI.

On the basis of this list of criteria, we concluded IBM had made the best offer. After this decision discussions with IBM about the final contract were started. The Koninklijke Bibliotheek has called in support by a lawyer specialised in these kinds of contracts.

We expect to finalise the contact in June 2000. The actual implementation of the project will probably start after the summer and take a maximum of 24 months.

6.2 The NEDLIB Model

For the functional model work done in the NEDLIB project (see A.5) was used (see the picture in figure 6). This model is in turn based on an ISO-standard that is currently being developed by NASA[12]. This is a reference model for an Open Archival Information System (OAIS). More detailed information about this model can be found at the website[13] of the NEDLIB project.

In this model (see figure 6) a distinction is made between the deposit proper and various systems that work around the deposit, such as systems for cataloguing, search and retrieve etc. The deposit proper is in the oval called DSEP (Deposit System for Electronic Publications), while the surrounding systems are in the box marked DLS (Digital Library System). The other systems that are not inside the box do not belong to the DLS.

The model of a DSEP consists of the following top-level processes that we find essential for a digital deposit. These are augmented with two processes that take care of the connection to the outside world.

Delivery & Capture
This process functions as an interface between the deposit and systems on the outside for matters related with the loading of material into the deposit. Because the processes in the DSEP can only handle electronic publications in a specific format[14], there has to be a way to inspect whether publications are already in the prescribed format, and if not, to convert them to that.

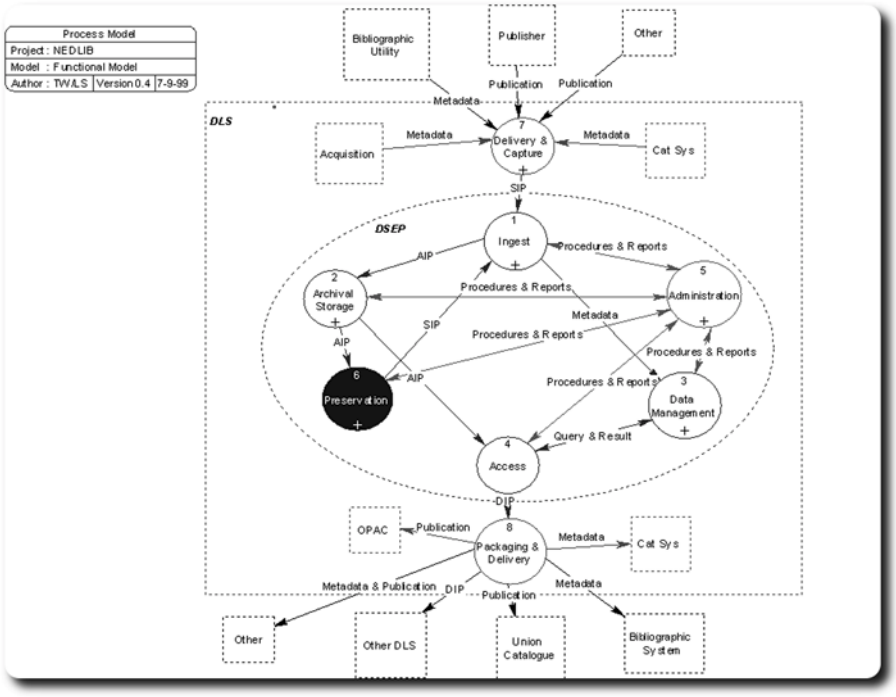

Figure 6: NEDLIB Functional Model Deposit System

Ingest
The Ingest process does a quality check on the received material. It adds metadata, establishes authenticity and integrity of the electronic publication and prepares it for archival storage.

Archival Storage
This process takes care of long-term storage of streams of bits. It does not have any knowledge about the contents of the document but it only knows how ‚to keep the bits healthy', that is that the integrity of the stored files is not compromised in any way. Archival Storage should be set up in such a way that files could be in there for hundreds of years. Archival Storage provides an abstract notion of a storage space that supports the use of multiple media (tape, disk, optical) and migration of files between these media because of decaying carriers.

Data Management
This is the central repository for all the metadata in the DSEP. This comprise both the metadata for the publications and the metadata to keep the DSEP operational.

Access
This process supports querying the metadata in Data Management and the retrieval of the document from Archival Storage in a format suitable for viewing at the user's terminal[15].

Administration
The day-to-day operation of the DSEP, quality management etc. is handled by this process. All other processes report to Administration.

Preservation
This process is not part of the original OAIS-model. Because the viewing of an electronic publication involves hard- and software you either have to find a way to preserve these also (e.g. with emulation) or you have to convert the contents of the original document (transformation). The preservation process does not prescribe a particular method of preservation. Instead it describes a number of steps that have to be taken to implement it in a comprehensive way.

Packaging & Delivery
This process functions as an interface between the deposit and systems on the outside for matters related with the making available of material. As such it is the counterpart of *Delivery & Capture*.

A RELATED PROJECTS

During the past 5 years the Koninklijke Bibliotheek also participated in a number of projects that also dealt with digital deposits. Below we mention just a few of them. Please note that some of them have already ended.

A.1 AIW

The Advanced Information Workplace is an integrated multimedia system which can be used to navigate various networks on which quality information is available and to process the information found. It is installed on some 80 public workstations inside the Koninklijke Bibliotheek. Parts of it are also

available on the Internet. It was - and still is - developed by the Koninklijke Bibliotheek with financial support from NWO[16].

This project will be finalised in 2000.

A.2 Biblink

Project BIBLINK[17] is funded by DG XIII/E-4 under the Telematics Application Programme of the European Union Fourth Framework Programme. It aims to establish an electronic link between national bibliographic agencies and publishers of electronic material, in order to establish authoritative bibliographic information that will benefit both sectors.

This project is already finished.

A.3 Cerberus

The research project CERBERUS[18] is concerned with authenticity and integrity of electronic documents in digital libraries with a deposit task. How can the authenticity and integrity of electronic publications be guaranteed when the publications were treated with the objective to provide access on the very long term? Documents are migrated to other carriers, converted to other formats or operating systems and old hardware and software is emulated. What is the impact on the electronic document? Is lay out and information lost - and if yes - how much information or lay out is lost?

This project is already finished.

A.4 Donor

On May 1st, 1998 the project Directory Of Netherlands Online Resources (DONOR[19]) has started. The objective of DONOR is to create an enabling infrastructure for information management and retrieval on SURFnet, the national research network of the Netherlands. DONOR will provide a coordinated approach to document and metadata management on the Web.

This project is already finished.

A.5 NEDLIB

Project NEDLIB[20] - Networked European Deposit Library - was launched on January 1st, 1998 with funding from the European Commission's Telematics Application Programme. The project ends December 31st, 2000 (see figure 7).

NEDLIB is a collaborative project of European national libraries. The Koninklijke Bibliotheek is project leader. It aims to construct the basic infrastructure upon which a networked European deposit library can be built. The objectives of NEDLIB concur with the mission of national deposit libraries to ensure that electronic publications of the present can be used now and in the future.

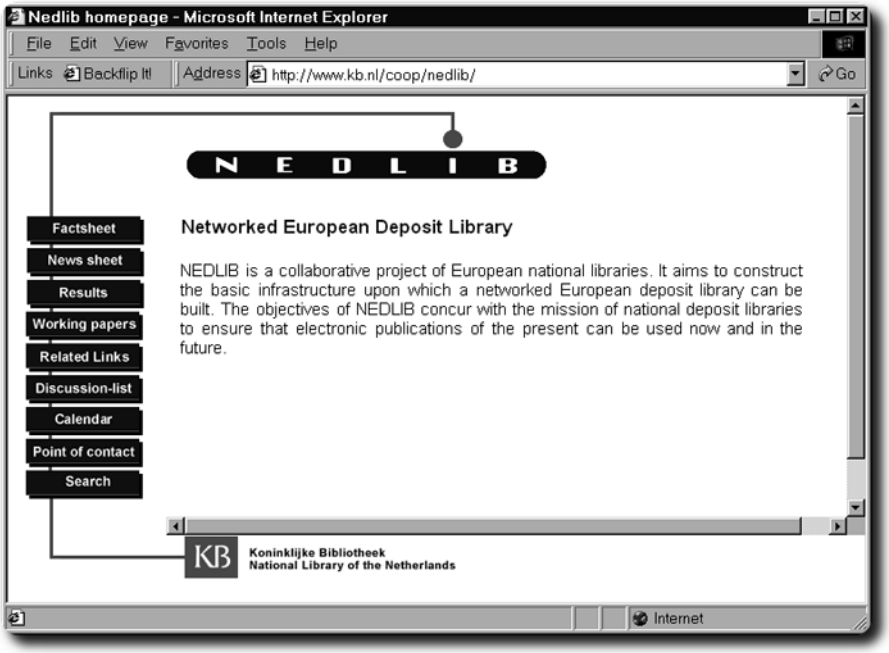

Figure 7: Homepage NEDLIB project

NEDLIB is an international project with partners from Portugal, Italy, France, Switzerland, Germany, The Netherlands, Finland and Norway. The partners are libraries and one state archive (of the Netherlands).

Within this project a number of products have been developed. These range from deposit guidelines, various software tools, a process model, a glossary of terms, research into metadata and preservation to descriptions of standards that can be used.

Within the project a demonstrator has been built as a proof of concept for the functional model of Nedlib (see figure 6). This demonstrator is operational until the end of the project at the site of the Koninklijke Bibliotheek.

In the last 6 months of 2000 also an experiment will be done in harvesting a part of the national Internet domain of the partners. For the project a harvester is developed that will be used for this. Webpages will be harvested for a fixed period of time. From these results interesting facts like the number of harvested pages, the number of different file formats, the number of web-servers etc. will be computed.

A.5.1 Research into preservation

Within the project research has been done into preservation of digital documents. This research is done based on the emulation-theory of Jeff Rothenberg, a computer scientist at the Rand Corporation who holds a keen interest into emulation as a means of digital preservation.

Emulation as such has been around for quite some time. However its applicability to digital preservation has never been proven. Jeff Rothenberg has been asked to set up and carry out a number of experiments designed to test whether emulation can be used for this and whether emulation can be used in a real-life production environment. That is: with a large number of documents of a great number of types and formats and a variety of hardware platforms. The experiment is set up as a number of iterations where with each iteration the complexity increases. Parts of these tests are done within NEDLIB, other parts are done for the Koninklijke Bibliotheek directly. The description of the set-up of the experiments and the results of the first iteration are available at <http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf>.

B REFERENCES

Here are some sites and documents not already mentioned in the text that you might consider relevant.

<*http://www.nla.gov.au/padi*>
Preserving Access to Digital Information (PADI). A subject gateway to digital preservation resources. A very good starting point. Also provides a discussion list.

<*http://www.cordis.lu/libraries*>
This site provides core information on the work carried out by the European Commission in the libraries field under the Third and Fourth Framework Programmes for Research and Technological Development, Telematics for Libraries, from 1990 to 1998.

<*http://www.cordis.lu/fp5/home.html*>
The Fifth Framework Programme (FP5) defines the European Union's strategic priorities for Research, Technological Development and Demonstration activities for the period 1998-2002.

<*http://www.leeds.ac.uk/cedars*>
Cedars stands for „CURL exemplars in digital archives" and the main objective of the project is to address strategic, methodical and practical issues and provide guidance in best practice for digital preservation.

<*http://www.kb.nl/kb/ict/dea/download/dig-info-paper.rothenberg.pdf*>
Ensuring the longevity of digital information, Jeff Rothenberg, february 1999. Note: this paper is an expanded version of the article *Ensuring the Longevity of Digital Documents* that appeared in the January 1995 edition of Scientific American (Vol. 272, Number 1, pp. 42-7).

<*http://www.clir.org/pubs/reports/rothenberg/contents.html*>
Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, Jeff Rothenberg, January 1998.

<*http://www.dlib.org/dlib/january00/01hodge.html*>
Best Practices for Digital Archiving, An Information Life Cycle Approach, Gail M. Hodge, January 2000. In: D-Lib Magazine, Volume 6 Number 1.

<*http://www.dlib.org/dlib/april99/bearman/04bearman.html*>
Reality and Chimeras in the Preservation of Electronic Records, David Bearman, April 1999. Reaction on the quicksand-article by Jeff Rothenberg above. In: D-Lib Magazine, Volume 5 Number 4.

<*http://www.w3.org*>
Homepage of the World Wide Web Consortium. The World Wide Web Consortium was created in October 1994 to lead the World Wide Web to its

full potential by developing common protocols that promote its evolution and ensure its interoperability.

C EXAMPLE ELECTRONIC DOCUMENT

This paper deals about long-term storage and preservation of electronic documents. As a real-life example, this document will be used to give examples of

- Bibliographical description of electronic documents with Dublin Core;
- Identification of electronic documents with a PURL;
- Integrity and authenticity control with electronic signatures.

These are just three things that are important when dealing with electronic documents.

This paper itself is an electronic document. It is created with a typesetting system called LaTeX. The output of LaTeX is medium/format-independent. When the document was created the focus was on the contents of the document, not how it should look. This document is available in more than one format. One of those is the traditional printed format, but this document is also available in electronic form as a PDF[21]-file. It would be fairly easy to also create, for instance, an HTML-version of the document.

All these different formats also have different functionalities. The printed version can only be read, but the PDF-file can, for example, be searched, annotated and contains links to the Internet that can be added to the document by the reader.

In the future publishers will probably be creating documents more and more this way. Instead of creating a product with only the printed version in mind, they will produce products in a more medium- and format-independent way. This document is just one example of how that can be done.

C.1 Bibliographical Description

Dublin Core[22] describes a limited set of elements for the bibliographical description of an electronic document. For a typical digital deposit you would need much more elements, for instance for administrative, rights, preservation

and technical metadata. At the moment there are no definitive standards for any of these classes of metadata. Dublin Core seems like a good starting point.

As an example the Dublin Core Version 1.1 metadata for the PDF-version of the document is listed in table 4.

Element	Value
Title:	The Dutch Deposit of Electronic Publications (DNEP)
Creator:	A.V. Sijtsma, Koninklijke Bibliotheek
Subject:	Digital deposits, archiving
Description:	Overview of activities undertaken by the Koninklijke Bibliotheek since 1995 with respect to archiving of digital material.
Date:	2000-06-15
Type:	text
Format:	application/pdf
Identifier:	purl.oclc.org/NET/liber2000 (PURL)
Language:	en
Coverage:	The Netherlands

Table 4: Dublin Core V1.1 Metadata for this document

C.2 Identification

Note that the identifier in table 4 is a PURL, a *persistent* URL. A PURL is a URN-like structure that has the advantage that it identifies the document as such (the entity) as opposed to an URL, that identifies a specific instance on a specific location. However, multiple versions of the document in multiple formats on multiple sites in different parts of the world can all be linked to one PURL.

If you go to the site <http://purl.oclc.org> and search for ‚/NET/liber2000' you'll find all the URL's of the document. If an URL is changed, only the administration at the PURL-server has to be changed to reflect the new situation. If on the other hand, you only have access to an URL you will get a ‚Error 404 - NOT FOUND' error and you will not be able to retrieve the document.

PURL is a free service of OCLC[23]. It is currently not in a state which allows implementation on a large scale or use in a production environment, but it is a nice - actually working - example of what a good identification mechanism

should be capable of doing and how this should work. On Monday June 19th, 2000, more than 564,110 PURL's were registered on the mentioned server.

C.3 Integrity and Authenticity

The PDF-version of this document is electronically signed by the author. This can be used to verify the integrity and establish the authenticity of the document. This is important because it is very easy to make changes to electronic documents (see figure 8).

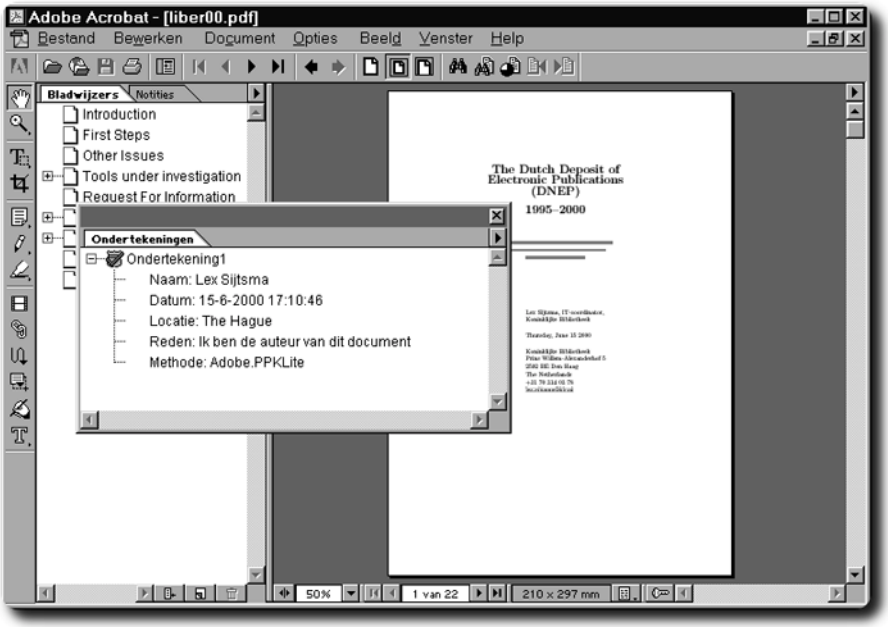

Figure 8: Electronic Signature

REFERENCES

1 Still available at <http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html>.

2 In Dutch: Depot voor Nederlandse Electronische Publicaties, hence DNEP.

3 FTE = Full Time Equivalents.

4 <http://www.kb.nl>.

5 See <http://www.surf.nl/iwi/portal.htm>. A three-year (1996-1998) program geared toward innovation of the scientific information in the Netherlands. IWI stands for Innovatie Wetenschappelijke Informatievoorziening.

6 <http://www.kb.nl/kb/menu/ken-dig-en.html>.

7 <http://www.ibm.com/software/is/dig-lib/>.

8 <http://www.pica.nl>.

9 Archives, Libraries and Museums, a.k.a. Memory Institutions.

10 <http://www.kb.nl/dea/rfi-dnep.pdf>.

11 In the European Tender Procedure there is no RFI-stage. The whole RFI-process is not part of the Tender Procedure.

12 <http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html>.

13 At <http://www.kb.nl/nedlib>, look for deliverable D1.4 *The Deposit System for Electronic Publications (DSEP), A process model.*

14 These are called Information Packages (IP).

15 This could be for instance a regular pc, a hand-held device or any other device.

16 See <http://www.nwo.nl>. The Netherlands Organization for Scientific Research (NWO) is the central Dutch organization in the field of fundamental and strategic scientific research. NWO encompasses all fields of scholarship.

17 <http://hosted.ukoln.ac.uk/biblink/>.

18 <http://www.kb.nl/kb/sbo/dnep/cerberus-en.html>.

19 <http://www.kb.nl/coop/donor/index-en.html>.

20 <http://www.kb.nl/nedlib>.

21 PDF: Portable Document Format, from <http://www.adobe.com>.

22 <http://purl.oclc.org/dc>.

23 <http://www.oclc.org>.