"Pursuing Digital Continuity & Digital Archiving of Electronic Publications"

Johan Steenbakkers
Director Information Technology & Facility Management & Koninklijke Bibliotheek, National Library of The Netherlands

IATUL 2001, May 29th

Pursuing Digital Continuity & Digital Archiving of Electronic Publications

by Johan Steenbakkers

Director Information Technology & Facility Management
Koninklijke Bibliotheek, *National Library of The Netherlands*

Keywords
Electronic Publications, Deposit System, Digital Continuity, Digital Archiving, Long-Term Preservation, Long-Term Access.

Abstract
At an accelerating speed the results of scientific research are being published in an electronic form. In addition to these 'born digital' publications, more and more printed publications are being digitised by publishers and libraries. A key task of the Koninklijke Bibliotheek (KB) is to maintain a Deposit for Netherlands Publications in order to guarantee long-term access to this information. As more and more of the information is published electronically, the KB has started some years ago a research and development program to realise a Deposit for Netherlands Electronic Publications (DNEP). At this moment a large-scale deposit system is being developed and implemented at the KB in co-operation with IBM-Netherlands. The system is planned to be operational in 2002. An important aspect of the implementation of the deposit system is to achieve digital continuity of the scientific information and to develop the functionality and good practices needed for archiving the electronic publications.

Introduction

National libraries traditionally collect and preserve the information published in their countries. By performing the task of depository library, a national library guarantees access to publications not only for actual users but also for future users. At an accelerating pace information is being published electronically, confronting national libraries with the challenge of archiving electronic publications and keeping them accessible through time, so in other words they provide continuity for the information they gather from and for society. To achieve digital continuity, appropriate procedures and technology are needed. As there were no solutions available of the shelve, the KB started already some years ago several research activities. Recently the KB involved IBM as a technology partner in the development of its deposit of electronic publications. This paper offers a brief introduction to the attempt of the KB to reach digital continuity.

Preliminary steps

From 1995 onwards the KB has conducted research and gained hands-on experience for managing electronic publications. These activities were possible thanks to the early co-operation of some major Dutch publishers – Elsevier Science, Kluwer Academic and SDU Uitgevers. (Later in 1999 a general agreement for depositing electronic publications could be signed with the Dutch publishers jointly.) In 1998 a pilot deposit system was implemented. The system has been expanded since then and nowadays contains 2 TB of storage.

The Deposit System

In 1999 the KB performed a market research for a large scale deposit system. Through an European tender procedure eventually IBM was selected as the supplier for the deposit system and the contract was signed in September 2000. The contract to IBM consists of to parts. The first part is the development and implementation of the deposit system initially with 12 TB of storage but expandable

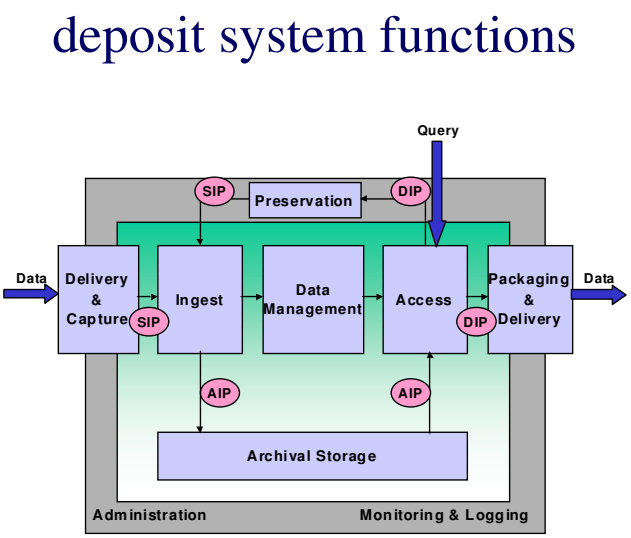
to at least 340 TB. The second part of the contact is the commissioning of research for the long-term preservation function of the deposit system. I will come back to this issue later.

The project DNEP (Deposit of Netherlands Electronic Publications) has started and runs on schedule. Today the functional requirements have been verified, the hardware and system software are installed and the general system design is ready and is being detailed. As the basis for the design we have used the NEDLIB-model.

NEDLIB (Networked European Deposit Library) is a project that was supported by the European Commission. NEDLIB has published a set of reports about setting up a deposit system. The NEDLIB reports can be ordered from the KB free of charge and I have brought copies of one of the reports, *The NEDLIB Guidelines,* for you to take away. (See for more information www.kb.nl/nedlib/ .) Two of the major guidelines of NEDLIB are:

- use the Open Archival Information System reference model (OAIS), an ISO standard for digital archives
- define the archive as a separate, self contained system within its digital environment.

The KB will implement the DNEP system as a black box, but will at the same time fully integrate the system within its digital library infrastructure, using a well developed input- and output interface. In the NEDLIB-model for a deposit system the following functions, using the OAIS terminology, are defined. (See figure from IBM's bid document.) Ingest for receiving, checking and preparing the data to be stored. Archival Storage, a high quality and scalable storage function (not the long-term functionality). Access, the function to retrieve the data for use. Data Management, this function is obvious. And last but not least Preservation, that stands for the long-term preservation function. This function has been added to the OAIS standard according to a proposal of NEDLIB. Delivery &



Koninklijke Bibliotheek, Nationale bibliotheek van Nederland

Capture and Packaging & Delivery stands for the input and output interface of the deposit system. The deposit system to be delivered by IBM to the KB will contain working modules for all the mentioned functionality, except for the preservation functionality. For this functionality a proof of concept will be conducted by IBM's research laboratory and - if possible - a prototype of the preservation module will be constructed and delivered.

Long Term Preservation

For achieving a solution for long term preservation (LTP), three aspects should be addressed by KB and IBM.
In the first place the intellectual preservation, meaning the maintenance of integrity and authenticity. The aim is to define this aspect of preservation and analyse the implications on the LTP function.
In the second place the medium preservation, concerning the preservation of the medium as well as the refreshment of the medium. The aim is here to bring together the current best practices.
And in the third place the technology preservation, dealing with the technology obsolescence. The joint research by KB and IBM especially addresses this aspect of LTP. This includes a Prove of Concept for a possible solution for implementing the LTP function: IBM's UVC emulation approach. For the case study the PDF format has been chosen. (UVC = Universal Virtual Computer)

UVC emulation approach

In IBM Almaden Research Center, California, the senior computer scientist Raymond Lorie is conducting the LTP research commissioned by the KB. Publications of Jeff Rothenberg, a senior consultant of RAND-Europe, about emulation as a possible solution for the long term preservation of digital information, caught some time ago the attention of Raymond Lorie and inspired him to define a practical approach: the UVC emulation. The UVC emulation approach exists of two major steps:

1) save the image of the (PDF) document, page by page
2) save the textual information, exporting it from the actual format so it can be used in a future context.

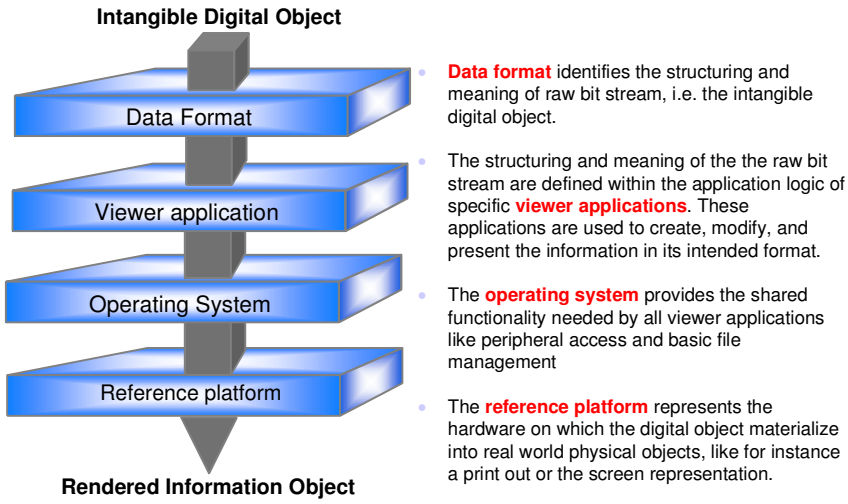
The activities will be performed by a program on a UVC.

To implement the LTP functionality, the PLM concept has been designed. PLM stands for Preservation Layer Model and is an important part of the metadata needed for technical preservation. The PLM is also essential for creating the environment for rendering the information, now and in the future.

Preservation Layer Model

The PLM is a key element in the preservation strategy developed by KB and IBM. A PLM represents in essence four parts (figure by Raymond van Diessen IBM):

On a abstract level a PLM identifies 4 abstraction levels.



Koninklijke Bibliotheek, Nationale bibliotheek van Nederland

1. the reference platform (=hardware environment) on which the digital information is materialised e.g. as a print or on a screen
2. the operating system providing the functionality needed for the rendering application
3. the viewer application used to create, modify and present the information in its intended format
4. the data format identifying the structuring and meaning of the bit stream.

Preservation Strategy

The preservation process using PLMs will consists of the activities shown underneath:

- creation of PLMs
- selection of the appropriate PLM
- association analysis between different PLMs
- defining and registering of the view paths
- critical view path analysis
- migration
- emulation
- reconstruction of technical environment.

Notice that the preservation strategy implies that both migration (sometimes also referred to as conversion) and emulation can be applied. The last step, the reconstruction of the technical environment is not directly related to the LTP strategy, but is more likely to be associated with the function of Delivery & Capture. But this function is also supported by the PLM approach.

Preservation in practice

As first step a PLM has to be created for each reference platform used for a certain type of digital objects. The PLMs are created and stored by the 'PLM administrator'. Be aware that this has to be done only once for each reference platform. So it is obvious to share the preservation metadata and make them generally available to archiving organisations. Also end-users should be able to obtain the PLM needed as an instant 'plug-in' when viewing the preserved information. For this purpose, repositories containing PLMs has to be organised.
Next step is to select and register for a certain document format the appropriate PLMs representing the reference platforms on which the information can eventually be rendered. These we call 'View Paths'. The deposit system will constantly check if at least two of View Paths are valid for a certain document format in order to guarantee access to the information. If this is no longer the case, because for instance the necessary application becomes obsolete, other PLMs should be made available. To do so the 'Preservation officer' has to make a choice to migrate (convert) the document format or to use emulation, either data emulation, application emulation or platform emulation. By doing this new View Paths are opened for rendering the information in its intended format.

Nevertheless KB and IBM are today still working hard to prove that such a long-term preservation approach can be implemented, I think that there is a good chance that within a few years the practice I have just described, will be reality.