LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich, printed in Germany

Legal Deposit of Digital Materials

by ERIK OLTMANS

LIBRARY PROCESS REDESIGN

Results of scientific research are nowadays as a rule published digitally. It is almost inconceivable that any scientist would not present the results of his or her work in digital form, even if it is in parallel to a physical copy delivered by a publisher. This is certainly the case for Science, Technology and Medicine, the research fields that produce over 80% of all scientific publications. Traditionally academic libraries have contributed to the keeping of the output of science through time. For deposit libraries - mostly the national libraries - the maintenance of the Record of Science is even a key task. The maintenance of collections by libraries is under pressure as more publications turn digitally. One reason is the change of policy of publishers who preferably do no longer sell but rather license their publications, offering them through sophisticated search and retrieval services. Another reason is that handling and maintaining digital publications requires new skills and a different infrastructure than for printed publications. And complicating matters even further is the fact that great research and development efforts are required, as no best practices on the issues are yet available.

Acquiring the infrastructure and skills

In 1994 the Koninklijke Bibliotheek (KB) decided to include electronic publications into its deposit collection. To facilitate this, specific infrastructure, organisation and skills had to be developed, not only for handling electronic publications but also for guaranteeing long-term availability. From 1995 onwards the KB experimented with electronic publications in co-operation with some publishers. In 1999 the Dutch Publishers Association signed an arrangement to deposit all electronic publications at the KB. The arrangement covers offline and online publications and prescribes restricted access conditions. Readers may only view the publications on-site. The intention is to develop in the near future new business models between national libraries and publishers for the access to the digital information.

DNEP - Developing the deposit system

In 1999 the KB started the project 'Deposit for Netherlands Electronic Publications' (DNEP) to acquire a full-scale deposit system. Preparing for the selection procedure, the library specified the system requirements, which were based on the ISO standard for digital archives: Reference Model for an Open Archival Information System (OAIS).

The deposit system had to offer large-scale, high quality storage and digital preservation functionality. Also its design should be future-proof and provide for scalability, extensibility and flexibility. As a result of a European tender procedure, the KB contracted in 2000 the development of the deposit system to IBM The Netherlands - Global Services. The development of the system started in October 2000 and was planned to take a two years period.

IBM developed the deposit system on site on the KB premises. In October 2002 the system was delivered to the KB. Meanwhile the library had created a workflow for electronic publications and had designed interfaces to the catalogue and other digital library functions. The KB deposit system initially has a storage capacity of 12 TeraBytes and is scalable over 500 TeraBytes. The system was constructed using as much as possible off-the-shelf components, like DB2, TSM, WebSphere and Content Manager. As the system is a generic archival system, IBM has branded it with the name Digital Information Archiving System, or DIAS, and will maintain it as a product. By using this product, the KB maintains the service branded as *e*-Depot. Parallel to the development of the system KB and IBM have jointly studied and tested Long-Term Preservation issues. The *e*-Depot will be extended further, technically and functionally, including realising specific preservation technologies.

The e-Depot realised

The deposit system DIAS is the technical heart of the KB's *e*-Depot. The contribution of several members of the KB staff, defining the requirements and testing the system during the development phase, has been a critical success factor. But after the system was delivered the real job started. The processing department had to implement the new workflow and acquire the skills needed for processing the electronic publications. The IT department had to run a system larger and more complicated than they ever had done before. The challenge offered by the electronic publications has a major impact on the KB organisation and on the tasks performed by its staff. But next to being a challenge, this development also offers an exciting opportunity to position the KB in the digital area.

Archival agreements with publishers

In the first half of 2003 Elsevier Science, Kluwer Academic Publishers, and SDU have signed a unique agreement with the KB on long-term digital archiving of all their electronic publications. Agreements with more publishers will be signed soon. At this moment the digital publications of these publishers are being loaded in the *e*-Depot, involving more than 2,200 journals, containing over 5 million articles. In case of publications that are processed both in digital and printed form, the KB has decided to process only the digital manifestation of the publication, while at the same time reducing the number of digital publications on carriers that are manually processed. With respect to version management of electronic publications, the *e*-Depot is able to deal with

updates, renewals, and withdrawals or retractions. Flexibility is thus a crucial property. New acquisition methods are necessary in order to obtain the publications. Another issue are publications on the Internet that will require some form of web harvesting. A complicating matter in this respect is the lack of automatic delivery of (qualitative) bibliographic descriptions of the harvest.

THE *e*-DEPOT – IT'S WORKING

The *e*-Depot offers on the one hand the storage facility and on the other hand the functionality for digital preservation. The KB has developed the workflow for archiving the electronic publications and has realised the other parts of the infrastructure in which the deposit system is embedded. This infrastructure consists of a variety of functions: for accepting and pre-processing the electronic publications, for generating and resolving identifier numbers, for searching and retrieving publications and for identifying, authenticating and authorising users, etc.

The stage of processing and ingesting the digital content is called loading. Two types of electronic publications are to be stored in the *e*-Depot: offline media like CD-ROMs, also referred to as installables, and online media like the high-volume of electronic articles sent by the publisher. Ingest of the installable is a time-consuming process that is performed manually. First, the CD-ROM should be installed completely, including all additionally needed software like viewers or media players. All files from the CD-ROM are subsequently copied to a Reference Workstation (RWS), so that the CD-ROM can be viewed stand-alone. A snapshot of the entire workstation is then generated into an image, including its operating system. After the bibliographic description is generated manually, the image is subsequently loaded into the *e*-Depot. If a customer wants to view a particular CD-ROM, the entire image is retrieved from the *e*-Depot, and installed on a dedicated RWS. By including the operating system in the stored package, the CD-ROM is guaranteed to work –also in future conditions involving new operating systems.

The second type of publications that are currently processed are online media. These publications are either send to the KB on tapes (for processing the backfiles) or by means of FTP. In both cases, publications ready for ingest end up in an electronic post office in which they are validated. In this stage the content of the submission is checked on well-formedness, based on specifications agreed upon earlier. If the material does not fulfil the checksum, or if other errors occur, the content is passed to a database for error recovery (BER). In fact, inspection of this database is currently the only manual effort involved. If the content occurs to be valid, content and metadata are put together as a Publisher Submission Package (PSP), and this PSP is then processed by a part of DIAS

called the *batch builder*. In fact, the batch builder itself consists of a series of applications, like Content Manager, and Tivoli Storage Manager.

The batch builder ingests both the content and the metadata and converts the bibliographical descriptions from the publisher into the KB's internal format, including the addition of a National Bibliographic Number (NBN). After conversion the content itself is stored into the *e*-Depot, while the metadata is stored into the KB catalogue. Clients may query the online catalogue and retrieve the full text of the publications, in the case of restrictions imposed by the publisher only after a process of identification, authentication and authorisation (IAA). The *e*-Depot itself cannot be accessed directly, but passes relevant documents to the client after clarification. See Figure 1 for a complete overview of the data flow.

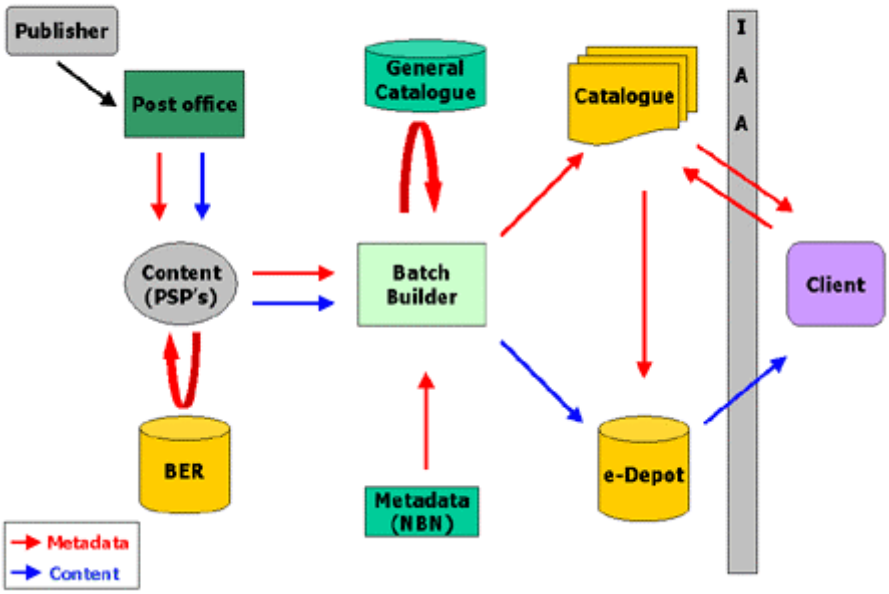

Figure 1: Data flow *e*-Depot

Fully automatic loading is done for large quantities of electronic publications, in which hundreds of thousands (in the end: millions) of articles are loaded during batch processes. An important requirement for this type of loading is that the publisher sends extensive and well-specified metadata along with the publications. This metadata, often according to proprietary standards, is subsequently converted into the KB preferred

format (Dublin Core-like in XML). By using the publisher's metadata, an important labour-intensive task is by-passed.

Conceptually, there is little difference between manual loading and fully automatic loading. The process of loading and storing is performed by DIAS. The DIAS solution provides a flexible and scalable open deposit library solution for storing and retrieving massive amounts of electronic documents and multimedia files. It conforms with the ISO Reference OAIS standard and supports physical and logical digital preservation. Once the asset is successfully stored it will be maintained and preserved. Stored assets can be accessed either via a web-based interface (for assets having standard file types) or via a specific work environment on a Reference Workstation.

LONG-TERM PRESERVATION AT THE KB

At the same rate at which our world is becoming digital, our information is threatened. New types of hardware, computer applications and file formats supersede each other, making digital information inaccessible. Even if the hardware or the carrier-media does not deteriorate within the time frame considered, the technology to access the information will inevitably become obsolete. Information technology is developing constantly and rapidly offering us new and appealing applications while at the same time making existing hardware and software obsolete. Preservation or permanent availability of the record of science is one of the processes which is dramatically affected by the change to an all digital world. In recent years, the digital preservation challenge has been recognised by people outside the traditional memory institutions (libraries, museums, etc). The problem of digital preservation is broad and society as a whole is having to deal with it. This has resulted in the issue of digital preservation being widely discussed today.

Towards persistent digital information

In essence digital information consists of two things: a formatted bit stream and the functionality needed to decode this logical format and render the information to the user. Even for apparently complicated digital items these are the real vital aspects. Therefore the key actions for digital preservation are:

- preserving the (formatted) bit stream, also called the 'digital object'
- ensuring accessibility over time to the information embedded in the digital object

For the problem of preserving digital objects, the European project NEDLIB has consolidated in a nutshell a wide range of internationally acquired research results in the

'Guidelines for setting up a Deposit System for Electronic Publications'. A strategic choice by NEDLIB was to transfer the electronic publications from the publishing environment to an archiving environment, nowadays often referred to as a 'Safe Place'. Note that the word 'place' in 'Safe Place' should not be taken literally, but rather as a concept, indicating an institution that is committed to digital preservation and possesses appropriate infrastructure, resources and skills for the task.

The essence of the data preservation concept is to extract the data from the format it has been published in order to render it in a different IT environment, also in future times. In order to describe the subsequent parts of the current IT environment needed to render the object today, the KB makes use of *Viewpaths*. Viewpaths are instantiations of an abstract model called the *Preservation Layer Models* (or PLM).

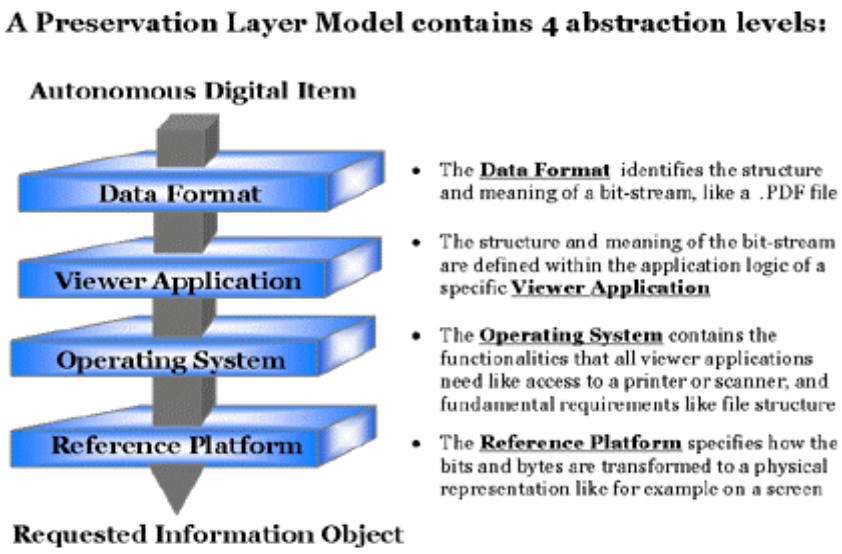

Figure 2: Preservation Layer Model

Every level in the PLM can be instantiated, thus generating a specific Viewpath, specifying which software and hardware is necessary to render the digital item. A PDF-file can be rendered in numerous ways: for instance by using Acrobat Reader that runs on a Intel machine with Windows 95 as an operating system. This PDF can also be viewed on an IBM RS/6000, running AIX 4.2 including an AIX PDF viewer. Both are

examples of specific view paths, and by administering the valid viewpaths, we get an overview of all possible ways to render a specific file type.

In order to manage all file types and their corresponding viewpaths, the KB and IBM have jointly developed the LTP Preservation Manager, which enables us to address the management of long-term preservation aspects of the stored digital items. It will keep track of stored file formats, manage the technical metadata and signal endangerment of rendering functionality. So if Windows 95 appears to become obsolete, this can easily be specified in the LTP Preservation Manager. It automatically determines the view paths that are Windows 95 depended, and marks them invalid. If a stored document, due to obsolescence of any software or hardware, is in danger of not being viewable anymore, actions have to be performed to 'save' the document (such as conversion, migration, emulation). Implementation of these actions are currently being developed by the KB and IBM, resulting in a proto-type of what is called the Universal Virtual Computer (UVC). This system is expected to render digital items based on a logical data view, independent of any future software or hardware environment. For more information about the development of the UVC the reader is referred to the KB and IBM websites.

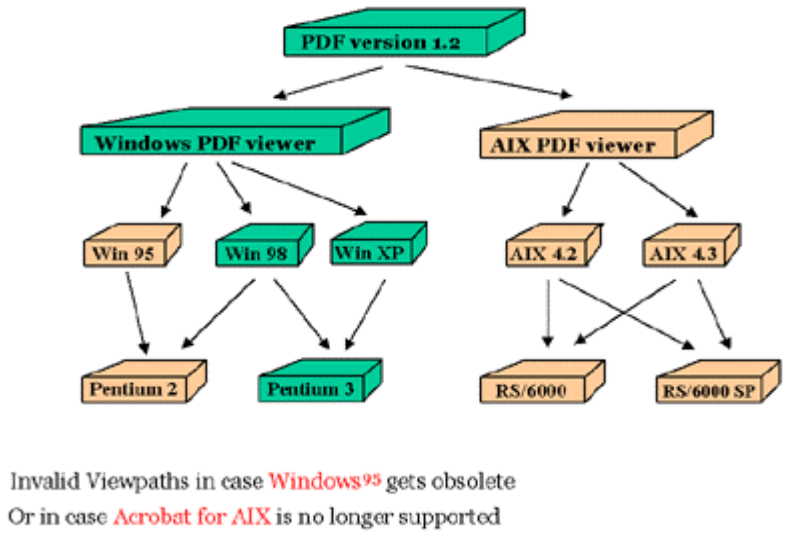

Figure 3: Specific Viewpaths get out of date

CONCLUSIONS & FUTURE RESEARCH: THE NEXT STEPS TOWARDS PERMANENT ACCESS

Implementing the LTP Preservation Subsystem is the first step towards long-term preservation; guaranteed rendering of that information is the second part. In addition to the safe keeping of the digital objects, access to the information in the objects has to be permanently provided. Tools, techniques and procedures are needed to provide access to the stored objects now as well as in the future. Research on tools and procedures for permanent access has started, but is still in its infancy. IT companies are only recently becoming aware of the problem of relatively short-term accessibility of digital objects. The standardised archival system developed by KB and IBM is designed to preserve and control digital information for the long term. Still to be resolved however is how to guarantee permanent access to the stored information. How can we render digital information for users in the future? The problem of permanent access has up to now been addressed by several, scattered and small-scale initiatives. To accelerate this development national libraries, archives, universities, research institutions and IT companies should collaborate in order to create tools for permanent access. In this context, the KB and IBM are constantly looking forward to set up new consortiums in order to join forces for further research and development on this important topic.

Acknowledgements

This overview could not have been written without the valuable input of Johan Steenbakkers, the initiator of the *e*-Depot. I also want to thank Hilde van Wijngaarden, Chris Bellekom, and Anne-Katrien Amse for their substantial contributions.

NOTES

1. The KB and IBM jointly performed research on long-term preservation issues. The reports are available through http://www.kb.nl/kb/ict/dea/ltp/reports/reports.html

WEB SITES REFERRED TO IN THE TEXT

Deposit of Dutch Electronic Publications, Koninklijke Bibliotheek . http://www.kb.nl/kb/menu/ken-arch-en.html

Digital Information Archiving System (DIAS): http://www.ibm.com/nl/dias/

Dutch Publishers Association = Nederlands Uitgeversverbond. http://www.nuv.nl/

Koninklijke Bibliotheek, National Library of the Netherlands. http://www.kb.nl/

NEDLIB. http://www.kb.nl/coop/nedlib/

Reference Model for an Open Archival Information System (OAIS). http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf