

# Preserving electronic publications

Johan F. Steenbakkers

*Director IT & FM of the Koninklijke Bibliotheek, National Library of the Netherlands, P.O. Box 90407,  
NL-2509 LK The Hague, The Netherlands*

*Tel.: +31 70 3140644; Fax: +31 70 3140651; E-mail: johan.steenbakkers@kb.nl*

**Abstract.** Scientific journals are more and more published electronically. Gradually the amount of e-books increases and on Internet a vast amount of digital information is being published. Electronic publishing offers major opportunities for dissemination and for access but at the same time imposes new challenges on national libraries and other deposit institutes in charge of preserving countries' cultural heritage. The challenges that have to be addressed are preservation of the electronic publications.

In the past years national libraries have taken the lead in developing practical solutions for digital preservation. Together with IBM both the Koninklijke Bibliotheek – national library of the Netherlands – and the British Library – national library of England – are developing an OAIS compliant deposit system. The system of the Koninklijke Bibliotheek will be operational at the end of 2002 and will be the core of the Electronic Deposit of the Netherlands.

As a deposit system could not be bought of the shelf, the approach chosen was to team up with a major ICT partner and jointly develop the deposit system. Parallel with the development of the deposit system, the aspects of long-term access were investigated. A proof of concept was done by IBM with 'data preservation' using a Universal Virtual Computer (UVC). The Koninklijke Bibliotheek and IBM jointly will soon publish a series of reports on the long-term preservation aspects of a deposit system.

**Keywords:** Digital preservation, long-term access, national library, deposit system, electronic deposit, electronic publication, OAIS

## 1. Introduction

Traditionally libraries play a key role in preserving the cultural heritage, materialised in manuscripts and publications. This cultural heritage consists of fiction and non-fiction publications, as we find them in bookshops, but also of the published results of scientific research, in scientific journals and books. Especially the publishing of scientific and non-fiction information is these days rapidly turning digital.

The move towards electronic publications offers obvious advantages for the user. At the same time this development creates new chances, but also major challenges, for publishers and libraries. Publishers have to transform their business of publishing in print into a digital analogue. They have to develop new business models adequate for electronic publishing. Stepping into the digital world, libraries have to reconsider and restructure their role, both as broker and as keeper of published information. Therefore new skills, tools and infrastructure for handling and preserving our digital cultural heritage have to be acquired.

In the case of printed publications, e.g., scientific journals, a library as a rule will keep the volumes of previous years for some time in its stacks or, in the case of a deposit library, keep the back-volumes 'forever'. It is obvious that in the case of electronic journals, a library will also have to guarantee their users access to a back-file of the publications.

Guaranteeing perpetual access to electronic publications appears to be an issue that affects both the publishers and the libraries. Until solutions for long-term preservation are in place, publishers will have to continue producing their journals also in print, even if the favoured format used is the electronic one. And libraries have to collect and keep a back-file of these journals in the printed format. For users this

approach has the drawback of loss of the functionality and convenience that is intrinsic to the electronic format.

The necessity of having a digital back-file is acute if the electronic publication is not just an electronic version of a printed publication. Printing for archiving is either not feasible or will result in loss of information. So eventually every library will have to offer its users access to digital back-files. Guaranteeing long-term access to electronic publications is, as will be discussed later, not a trivial matter. To fulfil this task the library can either maintain a back-file of the electronic publications on its own or, make arrangements with trusted third parties, as deposit or national libraries, who's key task is to maintain back-files.

## **2. The requirements**

To understand fully the requirements for preserving electronic publication, it is useful to examine the anatomy of an electronic publication. In essence an electronic publication consists of three components: the bit stream, the logical format in the bit stream, and the functionality needed to decode this logical format [8]. Even complicated publications break down into these three components. What are the main problems when preserving electronic publications and keeping them accessible through time?

In the first place the medium or carrier on which the publication – i.e., the bit stream – is stored, will deteriorate and as a consequence bits will be lost. Also the storage technology used will eventually become obsolete. To solve this problem we can use the functionality ICT companies offer for copying the bits and refreshing the storage medium. If copying is done without loss and the refreshing is done timely, the authentic structure of the bit stream can be saved indefinitely. In the case of deposit libraries the copying techniques should – of course – be appropriate for a large and continuously growing amount of data.

The second problem is, that the logical format will in due time become obsolete. For instance, Word has replaced Word Perfect today by large and there are many more examples of major changes in formats during the past few years. New formats are bound to arise with the increased use of ICT technology in various sectors of business and society. So even if we save the bit stream of electronic publications, at the long-term we will not be able to get to the information because we can no longer or only at great expense decipher the format used. The approach in daily life for this problem is format migration (sometimes referred to as conversion). However format migration, surely if it is applied subsequently, will result in loss of information and is therefore not a suitable solution for digital preservation. In addition to this drawback of format migration, there is the practical problem that deposit libraries will have to deal with a large and continuously growing variety of data formats.

The third problem arises from the fact that we need an interpreter to transform the bit stream into information readable to the human eye. An interpreter is software that provides the functionality for decoding format and data embedded in the bit stream. Software is in essence a bit stream and like the electronic publication it can easily be stored and maintained through time. The problem remains as we can no longer run the interpreter, because the hardware required is no longer available, has become obsolete or has been replaced by new non-compatible technologies.

## **3. The strategy**

In order to achieve preservation of and access to electronic publications for the long-term, three steps are necessary. To understand the necessity of these steps it is useful to again take in mind the analogy of preserving printed publications.

The first step to take is archiving the electronic publications. By archiving I mean assign an identifier, file the electronic publication in a controlled environment, and produce descriptive and technical meta data. This seems pretty basic, but one cannot keep a publication or retrieve it from an archive unless these actions have been executed. Moving the electronic publication from the publishing environment into a controlled preservation environment is also a prerequisite for the second step.

The second step is preserving the digital item. For printed material we take measures in order not to lose (part) of the information. For instance, we bind issues of journals, we fix loose pages and restore books. Analogue actions have to be performed for electronic publications in order to maintain the bit stream complete and in its original structure. As deposit and national libraries act as a last resort, this must be done in a pro-active way. These libraries cannot wait until losses have occurred before taking action. There must be regular procedures to check if the bit stream is endangered and timely copy the bit stream and refresh the storage medium.

The third step is to guarantee long-term access. With a printed publication this is not an issue, because the information can be readily read without any specific tool or functionality. This step is unique for digital preservation. Without taking this into account we may archive and maintain an electronic publication, but we cannot on the long-term guarantee access to the information stored within it.

#### **4. Guidelines and standards**

How can we turn the strategy described above into practice? In the period 1998–2000 a group of national libraries, a national archive and ICT companies have, with the support of three international publishers, investigated the requirements for digital preservation and long-term access. The project NEDLIB (Networked European Deposit Library) was a project co-financed by the European Commission. For more details and results please consult the url [www.kb.nl/nedlib](http://www.kb.nl/nedlib). Amongst the NEDLIB reports produced is a Guideline for Setting up a Deposit System [10]. The guidelines boil down to the following:

- Free the publication from its original carrier or environment, meant for publishing and not for archiving, and store the publication in a controlled archiving environment. In this way it will be possible to handle a large variety and large amounts of electronic publications.
- The controlled archive environment should be constructed in compliance with the OAIS model (Open Archival Information System), a standard initially proposed by NASA [1]. The OAIS model is in essence shown in the Fig. 1.
- The archiving environment – or deposit system – should be a separate unit within the institution's ICT environment, with clear interfaces. It is essential not to take on board all sorts of functionality not related to archiving – i.e., searching, authentication and authorisation, etc. Do not include them in the design of the deposit system. In this way the deposit system will be durable and can be upgraded with new techniques in mass storage, mass handling and preservation of electronic information, without affecting the rest of the ICT infrastructure. An additional, practical reason for this approach is that one unlikely will find a provider to build a very complex and affordable system.

#### **5. Deposit systems under construction**

At the moment the Koninklijke Bibliotheek (KB) and the British Library (BL) are developing a deposit system, in compliance to the guidelines and standards mentioned earlier. Both national libraries have in

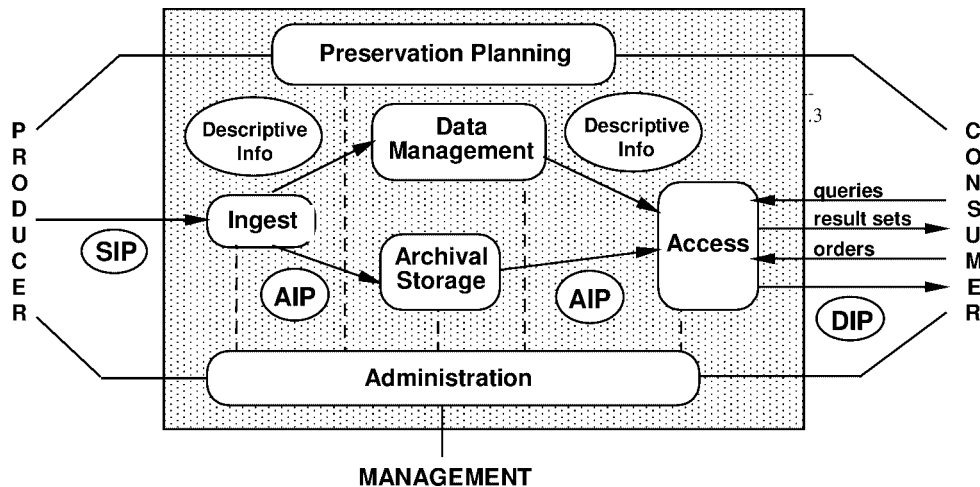


Fig. 1. The Open Archival Information System reference model (OAIS). The functions in the OAIS model are Ingest (for loading), Archival Storage, Data Management, Access and Preservation Planning. The digital data are defined as specific packages, according to the phase of processing. There are a SIP, Submission Information Package, an AIP, Archival Information Package, and a DIP, Dissemination Information Package. The function for preservation has been added to the OAIS model on proposal of NEDLIB. This function is essential for planning, monitoring and processing of data for long-term access. Note that there are only two interfaces, one for the producer (e.g., library staff) and one for the consumer (end user).

2000 teamed up with IBM in their respective countries, to design and build a deposit system. The KB and BL are closely co-operating in the design and implementation of their deposit systems.

To prepare for the development of the deposit system, the Koninklijke Bibliotheek in the period 1994–1999 ran several projects [5,11]. In 1998 a pilot system was implemented as a preliminary digital deposit. This system is still in use for processing and storing electronic publications and digitised documents. The system has two Terabytes storage capacity and contains over thousand electronic journals.

The first step to obtain a definitive deposit system was taken by the KB early in 1999. As a deposit system could not be bought from the shelf, a specific approach was needed. A selection of leading ICT companies was asked to respond to a questionnaire. The aim of the Request for Information was to examine if the ICT market understood and was interested in the problem of digital preservation and long-term access. The response was encouraging because major ICT companies expressed their interest to study digital preservation and develop solutions. As a consequence in summer 1999 the KB took the step to send out a Request for Proposals in a European tender procedure. And in September 2000 the KB could sign a contract with IBM-Netherlands. In the same month the BL signed a contract with IBM-United Kingdom for the development of a deposit system.

According to the contract with the KB, IBM will deliver the deposit system in October 2002. As the project is running on schedule, the expectation is that the system will be ready and operational at the end of this year. At the url [www.kb.nl](http://www.kb.nl) more information can be found about the project DNEP (Deposit Netherlands Electronic Publications).

## 6. Realising the E-Deposit of the Netherlands

The deposit system will offer on the one hand the storage facility and on the other hand the functionality for digital preservation. But the deposit system, nevertheless essential for the Electronic Deposit

(E-deposit), is just one of the building blocks for of the infrastructure of an E-Deposit. To realise its E-Deposit, the KB is actually developing the workflow for archiving the electronic publications and realising the other parts of the infrastructure in which the deposit system has to be imbedded. This infrastructure consists of a variety of functions: for accepting and pre-processing of the electronic publications, for generating and resolving identifier numbers [12], for searching and retrieving publications and for identifying, authenticating and authorising users, etc. At this moment the KB is testing the deposit system in practice and at the same time creating the other functions for operating the E-Deposit. Creating the infrastructure and implementing the deposit system is within the KB organised as a separate programme of activities, parallel to the development of the system. As the KB has for some years already been collecting electronic publications, a part of the implementation programme will also be the migration of these publications from the preliminary system into the new deposit system.

## 7. Solving long-term access

Once the deposit system is in place, the KB can archive and preserve electronic publications and other digital objects, and give access to them. But what is still missing, are provisions for long-term access. As part of the contract with the KB, IBM performs a long-term preservation study. A study and not a solution is requested from IBM as precise requirements for the functionality for long-term preservation cannot yet be defined. A working group guided and supported by IBM is conducting the study. The workgroup consists of KB library staff members, an expert from the RAND Corporation and as observer, the preservation officer of the BL. As part of the study experiments are conducted at the IBM laboratory at San Jose, CA, USA.

Methods to guarantee the rendering of digital information over a longer period of time are as such nothing new. From the moment ICT solutions were applied at large scale in business, collections of digital data have been maintained, sometimes even for a number of decades. Examples are databases of banks, of insurance companies, of oil companies and other branches of industry. The approaches used in these cases, either maintenance of outdated technologies or repeated conversion of the data, are neither generally affordable, nor adequate. To achieve generic and more efficient solutions for long-term access, KB and IBM have started developing specific preservation procedures and techniques.

An essential part of more generic solutions is the description of the rendering environment in technical meta data. KB and IBM have chosen to use for this purpose a layered model as is suggested by OAIS and detailed by NEDLIB [7]. See Fig. 2 for an introduction explanation of the PLM (Preservation Layer Model). All PLM components – except the hardware platform – are digital and as such can also be stored and maintained in the deposit system. The problem that remains to be solved is the obsolescence of the hardware of the rendering environment. To deal with this problem the KB at present investigates two potential solutions.

The first solution was proposed by Rothenberg already some years ago [8]. His solution is in a nutshell: save the bit stream, save the software of the rendering environment, create specifications for a hardware emulator and save the specifications. In the future when the rendering hardware becomes obsolete, the specifications should be used to build an emulator (a program representing the hardware) that can run the rendering software. In this way the original rendering environment can always be recreated. The emulation approach is a valid concept, but need yet to be developed further, before it can be applied [9,10]. More research and practical experiments are necessary. The KB is looking for partners to commission further work so as to turn the emulation approach into a practically applicable technology for long-term access.

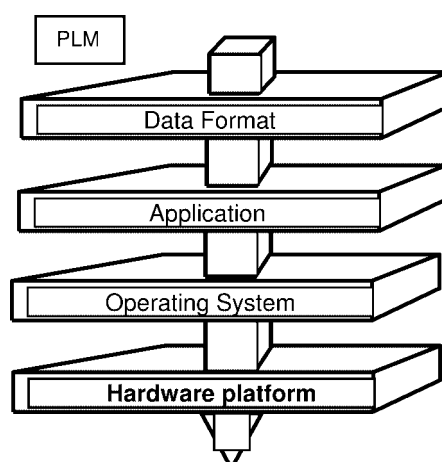


Fig. 2. The Preservation Layer Model (PLM). Information or meta data on the following components, registered through the Preservation Layer Model (PLM), are hardware, operating system, application (e.g., viewer) and data format. These are all the specifications needed – at any moment in time – to create the conditions for accessing the information.

The second solution, the use of ‘data preservation’, has been suggested by Lorie [2,3]. This approach implies the use of a specific format, that should be further developed into an archiving standard, and the use of a basic virtual computer, the UVC (Universal Virtual Computer). In a proof of concept experiment, commissioned by the KB, Lorie has created a logical data description of an article in the PDF format taken from an electronic journal of Elsevier Science and has recreated the article from this description using the UVC. The data preservation approach is applicable to any fixed format, like formats used for documents, music, film, etc. By creating a logical data description of the publication, rendering of the information becomes independent of the original format in which it has been published. Likewise the information can also easily be recreated in a future format and be used in a future ICT environment. However this approach seems only applicable to fixed – simple or complex – formats. In the case of dynamic electronic publications, like a spreadsheet, the only valid solution appears to be the use of the emulation approach mentioned earlier.

The results of the long-term preservation study will be published in a series of reports produced by the KB and IBM jointly. These results have already been taken into account in the design of the deposit system so that additional preservation functionality for long-term access can easily be added.

## 8. Organising digital preservation

In addition to specific equipment, technology and procedures for digital preservation and long-term access, preserving the digital cultural heritage requires political and organisational measures at national and international level. Deposit institutions, like national libraries, archives and museums, will have to adapt their business processes and develop new skills. At national level the development of digital deposits has to be supported and co-ordinated, as these essential deposits are not only expensive, but also difficult to realise and to maintain. Moreover one must keep in mind that achieving digital preservation is in the end not a local but a global problem. [See the UNESCO resolution on Preservation of Digital Heritage at the url [www.unesco.org/images/0012/001239/123975e.pdf](http://www.unesco.org/images/0012/001239/123975e.pdf).] International co-operation is essential to develop joint standards and to achieve interoperability between digital deposits for publications, scientific data, software and other elements of our digital cultural heritage.

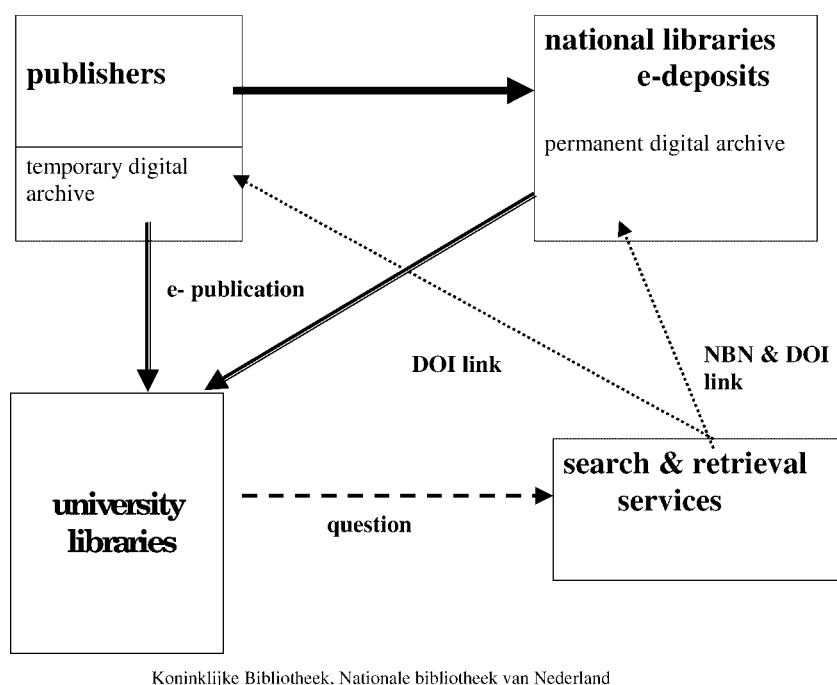


Fig. 3. Organisation of digital archiving in the future. The organisation for digital archiving is build on the presumption that archiving as such, is NOT a core business of a publisher. Actually publishers as a rule do not keep stock for printed publications. The same prevails for academic or university libraries, whose core business is to acquire information for and provide it to its stakeholders. However the core business of deposit libraries – often also national libraries – is to archive the publications and keep this information available and accessible through time. The ‘search & retrieval services’ will for the full text of the electronic publications refer to the e-deposits in national libraries. As linking mechanism both the NBN (National Bibliography Number) and the DOI (Digital Object Identifier) are suggested.

How for instance the future organisation for providing and preserving electronic publications might look like is shown in Fig. 3. I have drafted the scene keeping in mind the core business of the key players in this field: publishers, academic libraries and deposit libraries and their respective roles: publishing, marketing, acquiring and archiving the publications. All these roles require specific organisations, infrastructure, skills and procedures. To achieve and maintain these, great effort and a lot of resources are needed. So I presume that publishers and libraries will very much focus on core activities.

## 9. Conclusion

In many countries and for several years the problem of digital preservation has been studied. Only recently projects have started focusing on the development of practical solutions for this problem. Two national libraries, the Koninklijke Bibliotheek and the British Library, jointly with IBM, have initiated in 2000 the development of a deposit system for electronic publications. This system will be OAIS compliant and as such is not restricted in use for deposit libraries, but can generally be applied for preserving digital publications. It is expected that in 2002 the deposit system will become operational and will provide the functionality for archiving and preserving electronic publications.

However the deposit system version 2002 will only have a provisional function for guaranteeing long-term access. The full functionality for long-term access to the electronic publications is still under devel-

opment. Potential solutions are already in view and are being investigated and tested. Additional effort is needed to develop them further. The KB is looking for partners interested in jointly commissioning practical research on this long-term functionality.

In addition to technological and functional issues development locally and nationally of both infrastructure and policy for digital preservation is necessary. Attention should be given to achieving interoperability between the electronic deposits. Prerequisites to preserve human's digital cultural heritage are therefore, the development of standards, of national and global policy and of co-operation between the archiving institutions.

## References

- [1] CCSDS650.0-R-2: Reference Model for an Open Archival Information System (OAIS), *Red Book 2* (2001). See url: [http://ssdoo.gsfc.nasa.gov/nost/isoas/ref\\_model.html](http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html).
- [2] R. Lorie, Preserving digital information. An alternative to full emulation, *Zeitschrift für Bibliothekswesen und Bibliographie* **48** (2001), 205–209.
- [3] R. Lorie, A project on preservation of digital data, *RLG DigiNews* **5** (2001), 3.
- [4] C. Lupovici and J. Masanès, Metadata for the long term preservation of electronic publications, *NEDLIB Report Series 2* (2000).
- [5] T. Noordermeer, J. Steenbakkens and T. van der Werf-Davelaar, Electronic library developments in the Netherlands, *Liber Quarterly* **8** (1998), 57–80.
- [6] J. Rothenberg, Ensuring the longevity of digital documents, *Scientific American* **272** (1995), 42–47.
- [7] J. Rothenberg, An experiment in using emulation to preserve digital publications, *NEDLIB Report Series 1* (2000).
- [8] J. Rothenberg, Using Emulation to Preserve Digital Documents (2000). Published in print by the Koninklijke Bibliotheek. Also available in pdf at url: [www.kb.nl/kb/pr/fonds/emulation/emulation.html](http://www.kb.nl/kb/pr/fonds/emulation/emulation.html).
- [9] J. Rothenberg, NEDLIB experiment using emulation to preserve digital publications, *Zeitschrift für Bibliothekswesen und Bibliographie* **48** (2001), 200–204.
- [10] J. Steenbakkens, Setting up a deposit system for electronic publications. The NEDLIB guidelines, *NEDLIB Report Series 5* (2001).
- [11] J. Steenbakkens, Developing the depository of Netherlands electronic publications, *ALEXANDRIA* **11** (1999), 93–104.
- [12] T. van der Werf, Identification, location and versioning of Web resources. URI discussion paper, *DONOR Report* (1999).