The e-Depot at the National Library of the Netherlands

Based on a paper presented at the UKSG seminar 'Strictly Legal: Toeing the Copyright Line' British Library, London, 7 June 2005

ERIK OLTMANS
Head, Acquisitions & Processing Division

ADRIAAN LEMMEN
Account Manager, Acquisitions & Processing Division

National Library of the Netherlands

Electronic journals have come to dominate the field of academic literature, and it is of great importance to the international scientific community that this electronic intellectual output is preserved well and that it remains accessible in perpetuity. The traditional principles for the archiving of printed academic literature no longer suffice in the digital world. These are based on national frontiers: each national deposit library preserves its own national academic heritage. Regarding electronic publications, however, the geographical criterion is not very useful and the geographical provenance of material is irrelevant, since electronic data can exist independently of a geographic location. Most current journals of multinational publishers no longer have a fatherland that can be easily identified. New ways of co-operation in the field of long-term digital archiving of electronic publications and its metadata will emerge. This paper looks at the policy and ambitions of the National Library of the Netherlands (KB) regarding digital archiving of electronic publications.

Introduction

Electronic journals have come to dominate the field of academic literature, and it is of great importance to the international scientific community that this electronic intellectual output is preserved well and that it remains accessible in perpetuity. The traditional principles for the archiving of printed academic literature no longer suffice in the digital world. These are based on national frontiers: each national deposit library preserves its own national academic heritage. Regarding electronic publications, however, the geographical criterion is not very useful and the geographical provenance of material is irrelevant, since electronic data can exist independently of a geographic location. Most current journals of multinational publishers no longer have a fatherland that can be easily identified. New ways of co-operation in the field of universal bibliographic control together with a structured worldwide long-term digital archiving of electronic publications and its metadata will emerge. In this paper we will discuss the policy and ambitions of the National Library of the Netherlands (Koninklijke Bibliotheek/KB) regarding digital archiving of electronic publications.

KB Policy

Virtually every country has a national (legal) deposit of printed publications, and in most cases these collections are housed in the national libraries. Gradually, more national deposit libraries will also build electronic deposits for long-term preservation and access. The KB, as the national library of the Netherlands, also has a duty to collect and preserve all publications printed in the Netherlands, and its e-Depot is a logical extension of this task into the digital world. For several reasons it is therefore uncertain whether the traditional model, based on national deposits and geographical frontiers, will be able to guarantee the long-term safety of the international scientific output in a digital form. Since academic literature in electronic form no longer has a fatherland that

can be easily identified, it therefore has no obvious guardian. However, although many journals (articles) do not have a clear place of publication, they do need a place in which to be archived safely. Moreover, publishers will not deposit their material in an infinite number of digital archives. They probably want to sign archiving contracts with a limited number of e-depots – partly to spread their risks and partly for geopolitical reasons – and to deposit their material at a number of national institutions around the world. Hence, in the traditional model there is a huge risk of records of science being lost forever. Such a risk is unacceptable and a systematic and more concentrated approach is needed: the 'safe place network'.

This model is directly derived from the requirements for permanent archives. Permanent archives presuppose permanent commitment. A permanent archive should provide reasonable guarantee for continuity. Furthermore, permanent archiving calls for a substantial investment, not just financially, but also in the form of building up the necessary specific skills and expertise. Moreover, the preservation function will require an unremitting research and development commitment (see Van Drimmelen[1]).

From these requirements it follows that permanent archiving should be taken care of by a limited number of institutions, dedicated to this task. Permanent archiving should be prominent in their mission. Not every library should try to establish its own permanent archiving system. In the case of international scholarly journals, a handful of permanent archives, wisely spread around the globe, will suffice. The economies of scale that can be achieved provide a key incentive for developing this safe place model. The initial investments that will be required, in terms of financial resources and staffing, are very high. But once these investments have been made, expanding such an operation into an international service will clearly reduce the cost per unit of stored information.

Many of the arguments above come from the perspective of a national library, and a deposit practice. The recently issued statements formulated by the Andrew W Mellon Foundation and endorsed by the Association of Research Libraries, stress other perspectives as well[2]. They say that libraries and associated academic institutions must recognize that preservation of electronic journals is a kind of *insurance* against permanent loss, and that research and academic libraries may collaborate in the form of an insurance collective. Preservation is a way of managing the risk against the permanent loss of electronic journals, and against having journal access disrupted for a protracted period following a publisher failure.

In order to address these risk factors and to provide insurance against loss, qualified preservation archives should provide a minimal set of well-defined services, storing electronic journal files in trusted archives outside the control of the publisher. Archives must receive files that constitute a journal publication in a standard form, either from a participating library or directly from the publisher, and must store the files in non-proprietary formats. Moreover, archives should use a standard means of verifying the integrity of ingoing and outgoing files, and provide continuing integrity checks for files stored. They must also limit the processing of files in order to keep costs down, but provide sufficient processing so that the archives could locate and adequately render files for participating libraries in the event of loss. And finally, archives must restrict access by the participating libraries to archived files that are under copyright, in order to protect the publisher's business interests, except when the publisher goes out of business or is otherwise unable to provide consistent access.

Whether the concept is called 'safe place network' or 'insurance collective', the implication is clear: dedicated institutions are needed to take up the responsibility.

The KB aims to play a prominent role within the international safe place network. It meets the requirements mentioned above: the KB possesses a sound technical and organizational infrastructure and specialist skills and expertise, and has committed itself to an ongoing research and development effort. These assets provide a firm foundation on which to expand the e-Depot's international role. This international role generates substantial economies of scale, since it enables the investments necessary for the national e-Depot to be used even more efficiently.

Governance, funding and organizational structure

The KB was founded in 1798 and since 1993 has been an autonomous administrative body financed by the Ministry of Education, Culture, and Science (OCW). It receives an annual grant from this Ministry, amounting to €34m in 2005. The KB also

has some self-earned income (library passes, document supply and interest), which amounts to less than 10% of the annual budget. The Library may apply for additional funds to support special projects or investments in the infrastructure.

As for the e-Depot, the KB has reallocated funding within its own budget for several years. In addition, since 2003 the KB receives an ear-marked grant of €1.1m per year from the Ministry for system maintenance and the staff handling the operations of the e-Depot. The system maintenance is outsourced to IBM. The associated research and development budget was an additional €200,000 for staff. In 2005 this annual grant went up to €0.9m, exclusively dedicated to research into digital preservation. These funds are expected to increase further in 2006 and 2007, subject to approval by the Cabinet.

The e-Depot system falls under the Acquisitions & Processing Division, whereas the Research & Development Division includes the department for digital preservation research. The IT Division is responsible for technical maintenance, together with IBM. Totalling up all staff employed in handling the system, incorporating the publications, research projects, and management, more than 15 full-time equivalents are involved.

Agreements with publishers

In 1993 the KB decided to build a deposit collection of electronic publications, which was a logical extension of the deposit collection of printed publications already in place. General policy lines were formulated and in 1995 the KB started experimenting on a small scale with facilities for automatic handling of e-publications.

With this extension of tasks, the KB was confronted with the dilemma of electronic media: its short life expectancy. Digital material has a brief life-span, because of the limited longevity of information carriers and the software and hardware that make the stored information accessible to users. Therefore, since 1994 research and development on long-term digital preservation has been a topic of growing importance for the KB.

In 1996 the KB and the Dutch Publishers Association agreed on an arrangement for the voluntary deposit of offline electronic publications. At the same time discussions were initiated with Elsevier Science aimed at acquiring the content of Elsevier e-journals with Dutch imprint, and the first experimental bilateral archiving agreement was signed. Soon afterwards a similar archiving experiment was agreed with Kluwer Academic (see Steenbakkers[3]). The Dutch Publishers Association agreed on a new arrangement in 1999, which covered offline as well as online electronic publications with Dutch imprint (updated again in 2005).

A landmark electronic archiving agreement was drawn up with Elsevier Science in 2002: the experimental agreement of 1996 was expanded to cover the entire set of Elsevier journals. In total, the agreement defined the responsibility for preserving approximately 1,300 journals covering all areas of science, technology and medicine. The agreement also covers journals digitized as a part of Elsevier's retrospective digitization project. This arrangement turned the KB into the first official digital archive in the world for journals published by an international scientific publisher. In 2003 an official archiving agreement with Kluwer Academic followed. The early and successful implementation of the e-Depot and the commitment of Elsevier and Kluwer Academic, based on trust and commercial interest, put the KB in a natural position to assume an international role. After the agreements with Elsevier and Kluwer, the KB concluded similar agreements with, amongst others:

- BioMed Central (2003)
- Blackwell Publishing (2004)
- Oxford University Press (2004)
- Taylor & Francis (2004)
- SAGE Publications (2005)
- Springer (2005)
- Brill Academic Publishers (2005).

The third publisher the KB entered into an agreement with was BioMed Central. This contract signified an important step in two ways. Firstly, it underlined the international role of the national deposit system. BioMed has no Dutch origin. Furthermore, BioMed was established as an *open access* publisher right from the start. This also was new to the KB. Thus, the BioMed agreement represented a major strategic step. As the list of publishers makes clear, the KB does not discriminate between the places of origin, the publisher's business model, marketing strategy or any other features.

Designated community

There is a minimum set of conditions to be fulfilled if the KB enters into an archiving agreement. Publishers must deposit their publications free of charge. On the other hand, the KB has to accept restrictions on access, avoiding interference with the publisher's commercial interests. But there is a minimum: the KB provides permanent access to the journals on site to all authorized library users, including availability for inter-library document supply within the Netherlands, and including remote access if allowed by the publishers. For example, the archiving agreement with BioMed Central secures free (remote) access to over 100 open access journals covering all areas of biology and medicine. In addition, should there be a catastrophic disaster such that the publisher is inoperable for a long period of time, the KB would be part of the interim service system. The official archive thus serves as a guarantee to all licensees worldwide, by safeguarding the access that licensees have paid for. Finally, should the publisher or a successor cease to make these journals available, the KB could open access to all on a walk-in or remote basis.

Content characteristics

The e-Depot's content is predominantly driven by the archiving agreements. At present the e-Depot is receiving two types of electronic publication: offline media (CD ROMs that are fully installed before they are loaded into the e-Depot, including operating systems and additionally required software) and online media such as the electronic articles deposited by publishers. By early 2006, the e-Depot will contain approximately 5 million digital objects, corresponding to a little more than 5 terabytes of storage space. The total number of e-journal titles will be over 3,500. Full implementation of all current archiving agreements will result in an electronic archive containing more than 9 million digital publications. The annual increase in the number of articles from these publishers will be around 400,000.

The aim of the KB for the coming years is twofold. The KB will actively try to conclude archiving agreements with more of the major international scientific publishers. As can be seen in Table 1, the 20 largest publishing companies cover around 90% of the total world production of electronic STM literature. The KB would like to reach that level of coverage in the e-Depot. The KB will also try to obtain the most cited scientific journals in its e-Depot, irrespective of the publisher.

Next to this active strategy, the KB will accept STM literature from any other publisher who wishes to deposit material with the e-Depot, provided that the publisher is able to deliver the material in the preferred format and with the necessary metadata, and provided the publisher complies with the minimum set of access conditions as stated earlier.

	Publisher's name	Number of e-journals	Share	Total
1	Elsevier	1.313	25.7%	25.7%
2	Springer Verlag KG	885	17.3%	43.0%
3	Blackwell Publishing	433	8.5%	51.5%
4	Taylor & Francis Group	428	8.4%	59.9%
5	John Wiley & Sons	324	6.3%	66.2%
6	Lippincott Williams & Wilkins	185	3.6%	69.9%
7	BioMed Central Ltd	158	3.1%	73.0%
8	Thieme Verlagsgruppe	102	2.0%	75.0%
9	SAGE Publications	99	1.9%	76.9%
10	Oxford University Press	83	1.6%	78.5%
11	S Karger AG	79	1.5%	80.1%
12	World Scientific Publishing Co	74	1.4%	81.5%
13	Nature Publishing Group	71	1.4%	82.9%
14	Cambridge University Press	70	1.4%	84.3%
15	Haworth Press Inc	69	1.4%	85.6%
16	Mary Ann Liebert	56	1.1%	86.7%
17	IOS Press	53	1.0%	87.8%
18	Institute of Electrical and Electronics	48	0.9%	88.7%
19	Institute of Physics Publishing	44	0.9%	89.6%
20	Bentham Science Publishers Ltd	42	0.8%	90.4%
21	Emerald Group Publishing Limited	39	0.8%	91.1%
22	Adis International Limited	35	0.7%	91.8%
23	Koninklijke Brill NV	35	0.7%	92.5%
24	ACS American Chemical Society	34	0.7%	93.2%
25	American Institute of Physics	30	0.6%	93.8%
26	OECD	30	0.6%	94.4%
27	Urban und Fischer Verlag	30	0.6%	94.9%
28	Maney Publishing	29	0.6%	95.5%

Table 1. Overview of the largest publishers in the field of electronic scientific, technical and medical (STM) literature. The figures are derived from internal research based on Swets and EBSCO lists. Only e-journals and only STM journals were investigated.[4] The highlighted lines indicate publishers that have signed archiving agreements with KB.

Technical architecture and work flow

The first experimental deposit system was based on AT&T Right Pages. When Right Pages was withdrawn from the market in 1996, IBM Digital Library was selected to replace the AT&T software. It was recognized that IBM Digital Library was only a temporary solution because it did not have the functionality needed for a full-scale deposit system. In 2000, after a European tender procedure, IBM was selected to develop a new system together with KB staff. In this project the expertise of the KB and the technical knowledge and research forces of IBM were combined, resulting in DIAS: Digital Information and Archiving System. In late 2002, DIAS was delivered and embedded, resulting in the current e-Depot system. It is now fully operational and incorporated into the KB organization, as a department within the Acquisitions & Processing Division.

The infrastructure of the e-Depot consists of both components that were specifically developed for processing, archiving and maintaining e-publications, and typical digital library functions. According to the NEDLIB Guidelines, the deposit system should be a separate, dedicated entity within the library's digital infrastructure. For the traditional library processes (such as cataloguing, search and retrieval, and user registration and authentication) the KB uses the provisions already in place, thus avoiding duplicating these functions within the deposit system. This approach allows both the e-Depot system and the traditional library systems to evolve at their own pace.

The installable CD/DVD-based publications are first completely installed on a reference work station including all additionally required software such as image viewers and media players. A snapshot of the fully installed publication – together with the operating system on which it is installed – is then generated into a disk image. For these electronic publications, it is the disk image which is incorporated into the e-Depot, and customer use requires retrieving the disk image and completely installing it onto a work station.[5]

Most electronic publications and their associated files are obtained via digital tape or are acquired via FTP. Most publications arrive in PDF format, in what is called the 'Electronic Post Office'. The files are validated first, and then batched for further processing, while corrupt content is recognized automatically and is dealt with according to error handling procedures. The processing incorporates both the content files and the metadata. It converts the publisher's bibliographic data into the KB's standard format and adds a National Bibliographic Number (NBN) which is later used as the unique identifier of the stored item. There are functions for search, retrieval and delivery: the local overall catalogue database is freely available, whereas the content itself is only available after a procedure for the identification, authentication, and authorization (IAA) of end-users.

The functional design of DIAS is based on the open archival information system reference model.[6] The system is designed to be durable, and provides for scalability and flexibility. In 2003 an international Task Force on Digital Repository Certification was initiated. This task force has developed an audit instrument which is now being tested. Three digital archives have been chosen as pilots for the test-audit, among which is the KB e-Depot. The test-audits take place in February 2006 and will involve the investigation of the e-Depot's organizational and technical infrastructure and processes. The purpose of auditing a digital archive is to determine the degree of certainty the archive provides for the long-term availability and the functionality of the digital resources that are stored. The audit should ultimately result in a certified system.

Approaches to digital preservation

Providing permanent access to electronic material is a complex problem. Digital material is often unstable and has a brief life-span, because of the limited duration of information carriers and the software and hardware that make the stored information accessible to users. Safeguarding the integrity and authenticity of the material is therefore a key challenge when dealing with long-term preservation. Regardless of the chosen strategy, permanent access calls for continuous attention and action. The rapid pace of technological change means that the techniques and procedures for long-term storage and accessibility requirements need to be adjusted and improved constantly. A permanent R&D effort is therefore indispensable.

In general there are two main ways to approach digital preservation. The first one focuses on the digital object itself and aims at changing the object in such a way that software and hardware

developments will not affect its availability. By changing or updating the format of an object, it is made available on new software and hardware. The digital object will be adjusted to changes in the environment, which makes it possible to render objects by using current systems. The second approach does not focus on the digital object, but on the environment in which the object is rendered. It aims at (re)creating an environment in which the digital item can be rendered in its authentic form.

The first approach (changing the object) is known as migration or conversion. The second approach (changing the environment) is known as emulation. There are arguments for preserving the original 'look and feel' as well as for converting documents to new standards. The main reason for preserving the authentic form is that the KB digital archive serves as a safe place for original materials from publishers. The KB promises to keep the original bit stream of the received document. In the future, emulation tools will be needed in order to render these publications in the same way as they were published originally. Secondly, authenticity of a publication may be of importance for end-users who want to access publications and experience the original 'look and feel.' For these reasons, emulation tools are needed (see Lorie[7]).

On the other hand, there is a specific need for converting documents into the most current standard as well. For future end-users who want to have access to publications according to the standards and functionalities of that time, migration will be needed in order to copy and reuse data. For these reasons, both models are studied and considered for implementation at the KB, including cost issues.[8] The preservation planning module in DIAS allows for implementing both strategies. International collaboration (IBM, ICABS, PREMIS, PLANETS) will result in widely acknowledged technologies, preferably along distributed environments.

Whatever the chosen strategy, it will always imply repeated actions. There is uncertainty, however, as to exactly what these actions will have to be, since future technology is unknown. It is to be foreseen that new technologies will provide new solutions in the future. Therefore the KB will continue its permanent R&D effort, focused on the full range of available preservation techniques. The KB has developed its own R&D programme for the coming years and is a prominent partner in the European project Permanent Long-term Access through NETworked Services (PLANETS), aimed at generating a distributed framework for the development and application of instruments for preservation planning, preservation actions (tools) and content characterization.

Summary

The KB's policy and ambitions regarding permanent archiving of electronic publications can be summarized as follows:

- There is a growing volume of electronic publications without a natural fatherland, being crucial for academic research.
- These publications must be preserved for the long term, by organizations who take up the responsibility, and who are dedicated and equipped for this task (safe places).
- The KB has the ambition to be one of these safe places, and has had an electronic deposit system in place for nearly three years; its policy is acknowledged by the government.
- The KB looks forward to concluding archiving agreements with more international publishers.
- Two prominent methods for permanent preservations are studied and implemented, in close collaboration with international partners.
- The KB seeks constantly for collaboration, and wishes the e-Depot to be audited by an independent organization, preferably according to ISO-certification procedures.

References

1. Van Drimmelen, W., Universal Access through Time: Archiving Strategies for Digital Publications. In: *Libri, International Journal of Libraries and Information Services,* 54(2), 2004.
2. ARL Announcement, October 31, 2005. http://www.arl.org/arl/pr/presvejrnloct05.html
3. Steenbakkers, J.F., Developing the Depository of Netherlands Electronic Publications. *Alexandria – The Journal of National & International Library and Information Issues*, 11(2), 1999.
4. Voorbij, H., *Overview of electronic STM journals*, Internal KB publication, 2005.
5. Oltmans, E. and Van Wijngaarden, H.N., Digital Preservation in Practice: The *e*-Depot at the

Koninklijke Bibliotheek. In: *VINE, The Journal of Information and Knowledge Management Systems,* 34(1), 2004.

6. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, Blue Book, January 2002.
7. Lorie, R., A Project on Preservation of Digital Data. In: *RLG DigiNews*, 5(3), 2001.
8. Oltmans, E. and Kol, N., *A Comparison Between Migration and Emulation in Terms of Costs.* In: RLG DigiNews, 9(2), 2005.

■ **Erik Oltmans**
Head, Acquisitions & Processing Division
Koninklijke Bibliotheek
PO Box 90407, The Hague, NL
Tel: +31 70 3140 243
E-mail: erik.oltmans@kb.nl

Adriaan Lemmen
Account Manager, Acquisitions & Processing Division
Koninklijke Bibliotheek
PO Box 90407, The Hague, NL
Tel: +31 70 3140 468
E-mail: adriaan.lemmen@kb.nl