The KB e-Depot digital archiving policy

Erik Oltmans and Hilde van Wijngaarden
Koninklijke Bibliotheek, The Hague, The Netherlands

Abstract

Purpose: Electronic journals dominate the field of academic literature, and it is of great importance to the international scientific community that this electronic intellectual output remains accessible in perpetuity. In this paper we discuss the policy and ambitions of the National Library of the Netherlands (Koninklijke Bibliotheek, KB) regarding digital archiving of electronic publications.

Design/methodology/approach: We discuss three possible threats against permanent access, and we propose a coordinated and systematic approach to address these risks: the Safe Places Network.

Findings: This paper also includes a comprehensive overview of the e-Depot system and the KB approaches to digital preservation. The KB e-Depot has been operational for more than three years, and fulfils the most important requirements.

Research limitations/implications: The KB focuses on both migration and emulation as preservation strategies

Originality/value: This paper fulfils an identified need for collaboration

Keywords: Digital storage, Archiving, National libraries, The Netherlands

Paper type: Case study

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 permanent access. It is uncertain, however, whether the traditional model, based on national deposits and geographical frontiers, will be able to guarantee the long-term safety of the international academic output in a digital form. Academic literature is produced by multinational publishers, and has often no longer a country of origin that can be easily identified and thus no obvious guardian. Hence, in the traditional model there is a huge risk of academic records being lost forever. A systematic and more concentrated approach is needed to address this unacceptable risk.

Another threat is disrupted journal access for a certain period following a publisher failure, or publishers that stop making journals available for commercial reasons. As the prevailing model in digital publishing is licensing rather than archiving, libraries should know where they can go in case of loss, so as to guarantee continuous access for end users.

Finally, the last risk that we want to address here, is technological obsolescence. Digital material is often unstable and has a brief lifespan, because of the limited longevity of information carriers and of the software and hardware that make the stored information accessible to users. Although currently we can still render most file formats, we need to be prepared for objects appearing to be damaged or impossible to render. If we fail to pay attention and fail to continue our research efforts, this situation will be inevitable. The three major types of risks are depicted in Figure 1.

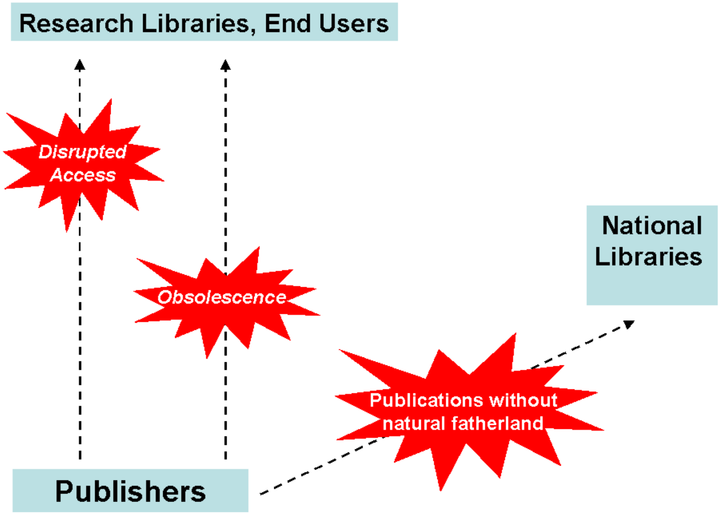
Research Libraries, End Users
Disrupted Access
Obsolescence
National Libraries
Publications without natural fatherland
Publishers

Figure 1: Risks threatening permanent access

The model that the KB proposes in order to control these risks, is called the Safe Places Network. It is based on systematic cooperation with publishers, who deposit their materials at a limited number of Safe Places. We argued that many journals do not have a clear place of publication, but they *do* need a place to be archived safely. Publishers are not likely to deposit their material at an infinite number of digital archives. They probably want to sign archiving contracts with a limited number of institutions around the world to deposit their materials –partly to spread their risks and partly for geopolitical reasons.

These institutions, serving as permanent archives, require permanent commitment. A permanent archive should provide a reasonable guarantee of continuity. Furthermore, permanent archiving calls for 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. 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 (cf. Van Drimmelen, 2004).

Many of the arguments above come from the perspective of a national library practising electronic deposit. The recently issued statements formulated by the Andrew W. Mellon Foundation, endorsed by the Association of Research Libraries (ARL), stress other perspectives as well (ARL, 2005). 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 the access of 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

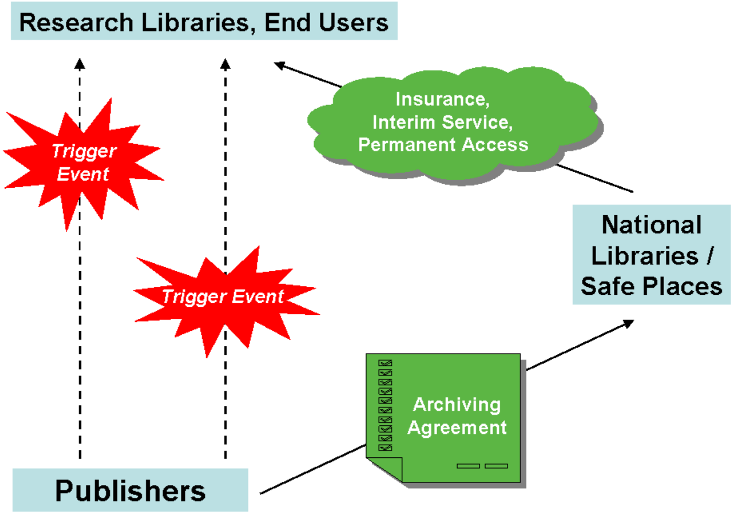
Research Libraries, End Users
Trigger Event
Insurance, Interim Service, Permanent Access
National Libraries / Safe Places
Trigger Event
Archiving Agreement
Publishers

Figure 2: Addressing the risks against permanent loss

access**.** These trigger events would be the main exception allowing widened access to third parties like other libraries or end users (see Figure 2).

Whether the concept is called *Safe Places Network* or *Insurance Collective*, the implication is clear: dedicated institutions are needed to take on the responsibility. The KB aims to play a prominent role within the international Safe Places Network. It has implemented all the requirements mentioned above: with the operational digital archive e-Depot, the KB possesses a sound technical and organisational 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, generating 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. The KB receives an annual grant from this Ministry, amounting to €40m in 2006. The KB also has some internally-generated income (library passes, document supply and interest), which amounts to less than 10% of the annual budget. The KB may apply for additional funds to support special projects or investments in the infrastructure.

As for the e-Depot, the KB has re-allocated funding within its own budget for several years. In addition, since 2003 the KB receives an earmarked grant of €1.1m per year from the Ministry for system maintenance and for 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. The total number of staff handling the system, ingesting the publications, research projects, and management, is equivalent to more than 15 full time posts.

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 lifespan, because of the limited longevity of information carriers and of 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 aiming 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 on with Kluwer Academic (cf. Steenbakkers, 1999). 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 nearly all Elsevier journals, also covering journals digitized as part of Elsevier's retrospective digitization project (estimated at a total number of approximately 7 million articles). 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 Kluwer and Elsevier, 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:

- 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. Secondly, 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 before the KB enters into an archiving agreement. Publishers must deposit their publications free of charge. However, the KB has to accept restrictions on access, avoiding interference with the publisher's commercial interests. But there is a minimum level of provision: the KB provides permanent access to the journals on site to all authorized library users, including availability for interlibrary 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 event, 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. In this way, the KB secures permanent access to both libraries and end users, without threatening a publisher's business interests.

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 publications: 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. In March 2006, the e-Depot contained over 5.8 million digital objects, corresponding to a little more than 6 terabytes of storage space. The total number of e-journal titles is 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. The twenty largest publishing companies cover almost 90% of the total world production of electronic STM (Science-Technology-Medicine) literature and 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 for its e-Depot, irrespective of the publisher. Alongside this active strategy, the KB will accept electronic 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.

Apart from archiving scientific digital publications, setting up a successful archiving workflow and infrastructure has also opened up opportunities for the long-term storage of other kinds of digital material. Projects have started to develop functionality and models for the storage of digitised material and websites. The KB is also working together with the Dutch university libraries to store their scientific output for the long-term in the project DARE (Digital Academic Repositories) [1].

Technical architecture and workflow

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 embedded in the KB organization, as a department within the Acquisitions & Processing Division. As well as at the KB, the DIAS system is also in use at Die Deutsche Bibliothek (DDB), the German national library. The current users of the DIAS system meet twice a year to exchange experiences and to work together on improving the functionality.

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 (Networked European Deposit Library), 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 firstly completely installed on a Reference Workstation, 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 ingested into the e-Depot, and patron use requires retrieving the disk image and completely installing it onto a workstation (Oltmans and Van Wijngaarden, 2004).

Most electronic publications and their associated files are obtained via digital tape or are acquired via File Transfer Protocol (FTP). 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 ingests 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 Identification, Authentication, and Authorization (IAA).

The functional design of DIAS is based on the Open Archival Information System Reference Model (cf. Consultative Committee for Space Data Systems, 2002). 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 by the Research Libraries Group (RLG) and the National Archives and Records Administration (NARA) [2], which has developed an audit tool which is now being tested. Its purpose is to produce certification requirements for establishing and selecting reliable digital information repositories. Three digital archives have been selected as pilots for the test-audit, of which the KB e-Depot is one. The test-audits are taking place in February-April 2006 and will involve the investigation of the e-

Depot's organisational 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.

The KB approaches to digital preservation

Providing permanent access to electronic material is a complex problem. As has been said, digital material is often unstable and has a brief lifespan, because of the limited longevity 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.

There are two main approaches to digital preservation. The first one is migration and focuses on the digital object itself. It 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 using current systems. The second approach is emulation, which does not focus on the digital object itself, 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 the same form as upon delivery to the archive.

There are arguments for preserving the original *look and feel*, as well as for converting documents to new standards. Both models are therefore being studied and considered for implementation at the KB, taking cost issues into account (cf. Oltmans and Kol, 2005).

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 do its utmost to safeguard the integrity of the articles that are deposited in the e-Depot. If at all possible, KB wants to save an article 'as is'. In digital preservation the contradiction is that changing less, implies doing more. In the long term, emulation tools will be needed to render these publications in the same way as they were published originally, and this kind of emulation tool does not exist yet. The development of preservation-based emulation is also important for those end users who want to access publications and experience the original look and feel. In the shorter term, migration can be performed without changing the content too much, especially when considering plain text articles. But if migrated articles have to be migrated again, stacked errors may occur, which damage the integrity over time. On top of this, new publishing formats offer the opportunity of adding moving images, interactive models or spreadsheets, challenging current migration techniques. In this situation, emulation becomes a necessity rather than a choice. This is why KB has started a project to develop a modular emulator for digital preservation together with the Nationaal Archief of the Netherlands (Van der Hoeven and Van Wijngaarden, 2005). In 2007 the emulator will be delivered for implementation in the e-Depot infrastructure.

However, emulation is certainly not our only strategy. Migration can be a good alternative in the short term, as is said above. It can also be developed further with the aim of offering future access to publications according to the standards and functionalities of that time. Migration will be needed to enable future browsing, copying and reusing data. Therefore, in 2006 KB began a project to do research on the quality of existing migration tools and on the way these tools can be integrated in the KB workflow. Gaps in the availability of viable migration tools will be identified, resulting in new plans to fill these gaps.

In order to execute preservation plans we also need structured information about the technical properties of the stored file formats. Together with IBM, an application for the storage of technical metadata and rendering information has been developed, which is called the *Preservation Manager* (Oltmans *et al.*, 2004). In 2006, KB will also implement the JSTOR/Harvard Object Validation Environment (JHOVE), an application developed by Harvard University Library and JSTOR to extract technical information from delivered publications [3].

The development of preservation planning capability requires a 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 acknowledges the need for international collaboration which will result in widely accepted technologies, preferably in distributed environments. An important step towards joining forces in this area is the European project Planets (*Preservation and Long Term Access through Networked Services*). Planets was proposed to the European Commission in the 6th European Framework Programme for Research and Technological Development and is currently under negotiation to start in May 2006. Planets is coordinated by the British Library and brings together a diverse but expert group of partners. Stakeholders in finding solutions for preservation planning and permanent access are the national archives of Great Britain, The Netherlands and Switzerland and the national libraries of Austria, Denmark, Great Britain and The Netherlands. Research institutes are working on the problem from a more scientific point of view and will work together with the stakeholders on building solutions. Partners are the University of Cologne, University of Glasgow, University of Freiberg and the Technical University of Vienna, which all have experience and expertise in the field. Technology vendors have joined our group to build tools and the technical infrastructure that will allow us to set up work together in a networked environment. These technical partners include the Austrian Research Center, IBM, Microsoft and Tessella. The participation of these commercial partners facilitates the take-up and dissemination of research results. The result of the project will be a distributed preservation framework for the development and application of tools for preservation planning, preservation actions (tools) and content characterisation. It will also include a decision support system, which will help institutions to decide which preservation strategy suits their situation best.

Summary

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

1. There is a growing volume of electronic publications without a natural country of origin which are crucial for academic research.
2. These publications must be preserved for the long term, by organizations who take on the responsibility, and who are dedicated and equipped for this task (safe places).
3. 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.
4. The KB looks forward to concluding archiving agreements with more international publishers.
5. Two prominent methods for permanent preservation are being studied and implemented, in close collaboration with international partners.
6. The KB is constantly seeking opportunities for collaboration, and would like the e-Depot to be audited by an independent organization, preferably according to ISO-certification procedures.

Notes

1. www.darenet.nl
2. www.rlg.org/en/page.php?Page_ID=367
3. http://hul.harvard.edu/jhove/

References

ARL Announcement (October 31, 2005**),** "ARL endorses call for action to preserve e-journals", available at: www.arl.org/arl/pr/presvejrnloct05.html (accessed 19 April 2006).

Consultative Committee for Space Data Systems (January 2002), *Reference Model for an Open Archival Information System (OAIS)*, Blue Book, available at: http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed 24 April 2006).

van Drimmelen, W. (2004), "Universal access through time: archiving strategies for digital publications", *Libri, International Journal of Libraries and Information Services* ,Vol. 54, No. 2, pp. 98-103.

van der Hoeven, J.R. and Van Wijngaarden, H.N (2005), "Modular emulation as a long-term preservation strategy for digital objects", in: *Proceedings of the 5th International Web Archiving Workshop (IWAW05) held in conjunction with the 8th European Conference on Research and Advanced Technologies for Digital*

Libraries (ECDL 2005), September 22-23 2005, Vienna, Austria, available at: www.iwaw.net/05/papers/iwaw05-hoeven.pdf (accessed 19 April 2006).

Oltmans, E. and Kol, N. (2005), “A comparison between migration and emulation in terms of costs”, *RLG DigiNews*, Vol. 9, No. 2, available at: www.rlg.org/en/page.php?Page_ID=20571#article0 (accessed 19 April 2006).

Oltmans, E. and van Wijngaarden, H.N. (2004), “Digital preservation in practice: the *e*-Depot at the Koninklijke Bibliotheek”, *VINE, The Journal of Information and Knowledge Management Systems*, Vol. 34, No. 1, pp. 21-26.

Oltmans, E., van Diessen, R.J. and van Wijngaarden, H.N. (2004), “Preservation functionality in a digital archive”, in: *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* , Tucson, Arizona, USA, June 11 2004, ACM Press, New York, pp. 279-286.

Steenbakkers, J.F. (1999), “Developing the Netherlands depository of electronic publications”, *Alexandria – The Journal of National & International Library and Information Issues*, Vol. 11, No. 2, pp. 93 et sqq.