Long-term preservation and permanent access: How to ensure the long-term reuse value of your digital assets

Hilde van Wijngaarden
is head of the Digital Preservation Department in the Research and Development Division of the National Library of the Netherlands. Her main area of expertise is the development of preservation strategies. She studied history at the University of Amsterdam and wrote her Ph.D. at the University of Groningen (2000). In 1999, she started working as an information consultant for an IT company. Since 2002, she is working at the Koninklijke Bibliotheek where she is head of the Digital Preservation Department since January 2005. She is a member of the EU-project PLANETS Scientific Board and has contributed to several international publications and presentations.

Keywords: *digital preservation, e-Depot, migration, emulation, international projects*

Abstract DAM should focus not only on storage and high-quality retrieval of digital collections, but also on digital preservation. To ensure future access to our digital collections, three actions have to be dealt with: long-term storage, registration of metadata and development of tools for accessibility. Storage media have to be refreshed regularly to safeguard the bits and bytes, and stored digital objects have to be described carefully to enable future retrieval. Apart from the description of the objects themselves, we also have to describe the technical properties of the object. Software is becoming obsolete over the years, leaving objects inaccessible for future users. Preservation strategies, like migration and emulation, have to be developed to deal with this problem. This paper describes how digital preservation has been taken up by libraries and archives with a special focus on the e-Depot at the National Library of the Netherlands (Koninklijke Bibliotheek, KB).

Journal of Digital Asset Management (2007) **3,** 102–109. doi:10.1057/palgrave.dam.3650064

WHAT IS DIGITAL PRESERVATION?

A book can be kept on a bookshelf and still be readable after a hundred years, but a digital object has to be actively preserved and maintained. The basic fact that digital objects are not human-readable, but require a computer program to translate stored bits and bytes, poses many challenges. New systems are succeeding each other much faster. Software developers are focussed on creating more functionalities, faster programs, newer and better platforms. That is in their interest and that is where they make their profits. And the strange thing is that everybody accepts this. Of course, some of us complain at the release of yet another Windows version, but without much problems, the people at home and in professional life accept it, because they want to have newer, faster computers with the newest operating systems. And without much complaint, they seem to accept that old programs do not work anymore, as if this is something that is a natural side effect of using a computer. But more and more, everybody is starting to realize that there is a risk of old files becoming inaccessible. Especially in the most recent years, with the widespread use of digital photography, people start to experience problems in a very direct way and realize this is an issue that has to be solved (Figures 1 and 2).

For libraries and archives, the question of how to preserve digital objects was raised about 15 years ago. These are the organizations that have the responsibility of long-term management of our cultural and scientific heritage, maintaining paper records and

Hilde van Wijngaarden
Digital Preservation Department
Research & Development,
Koninklijke Bibliotheek/
National Library of the Netherlands,
The Netherlands
Tel: + 31 70 3140467/
+ 31 6 24808338
E-mail: hilde.vanwijngaarden@kb.nl
Web: www.kb.nl/e-depot

Figure 1: An old computer with obsolete devices

Figure 2: A damaged tape

publications for centuries. Since the early 1990s, they were confronted with the shift from paper to digital and had to focus on long-term curation of digital objects. For libraries, the start of digital publishing meant that they had to rethink their role. They realized that a new model was needed. With e-journals, a library no longer builds up a collection, but gives access to the journal through licensing agreements. Another important change is that the traditional principle of national libraries that store their countries' publications is not sufficient anymore, as digital articles are not printed in a specific town or country anymore. Publishers realized this change as well and were willing to discuss archiving agreements with libraries, because they realized that long-term preservation of e-journals may not have been their expertise, but was in their best interest as well.[1]

In the archival world, a lot has changed as well. An archive cannot just wait for government organizations to turn over their records 20 years after their creation, because, by then, they may have become inaccessible already. For all records, including judicial records, an active approach toward long-term preservation has to start at the moment of creation of those records, and as soon as possible after that, with a first selection, description and careful storage.

Another group that was among the first to take up the problem is the space research society. Working with large-scale computer systems storing huge quantities of data, NASA was a very early active participant of the digital preservation research community.

Digital preservation is a relatively new problem, but a lot of progress has been made on how to approach this problem. We now know that, for successful digital preservation, three types of actions are required: long-term storage, registration of metadata and development of tools for accessibility.

LONG-TERM STORAGE OF DIGITAL COLLECTIONS

The description and storage of digital documents has received growing attention during this past decade, which has led to the development of metadata standards and standards-based archiving systems. The underlying hardware and carriers have to be refreshed regularly and the choice for media has to be based on longevity. A long-term archiving system is different from just any other server with regard to security and maintenance. The most important aspect of this is that a long-term archive has to be dedicated to storage. Access processes have to be separated from the storage process. During the past ten years, our knowledge of the requirements for a long-term archiving system has enhanced enormously. Of major importance has been the development of an international standard, the Open Archival Information System (OAIS) Reference model. The OAIS model is the result of intensive international cooperation and is now an ISO norm and accepted worldwide as a blueprint of what a digital archiving system should look like (Figure 3).[1]

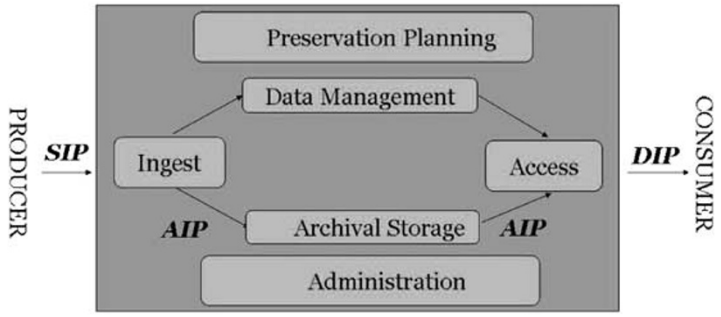

Figure 3: A representation of the OAIS Reference model

The second step is careful registration of metadata. We have to know what we have stored, to be able to find it again. These are the "traditional" descriptive metadata like author, title, etc. But in the digital world, we also have to know the technical requirements of the file itself, to be able to render the object. What is the file format, what program is needed to run the file and even what kind of underlying operating system and hardware is needed to be able to run the program? This kind of information is called representation information, and is also referred to as technical metadata. And then, especially when digital records are concerned, we have to store information on context and authenticity. How do we know the status of a document? In what context has a record been created and what records belong to the same process? In the archival community it is especially this new kind of records management that has received a lot of attention in past years. The question of authenticity is even more difficult. How can we trust what we see? There is a paradox here: in order to protect a record, we are looking at encryption technology. But at the same time, encryption is a major risk for digital preservation. Fortunately, there are other ways to ensure trustworthiness through metadata registration, and this is by registering fixity and provenance information. In this field, as in the field of archival system architecture, international cooperation has resulted in a first version of what could become a standard: the PREMIS preservation metadata data model.[2]

If a digital collection is stored for the long term, the assets themselves, their production settings and format have to conform to certain standards as well. Images should not be compressed, or at least not in a way that could have damaged the original image or in a way that could limit our possibilities of future representation. PDF publications should have embedded font sets and should not include any settings that limit usage or conversion in the future. Tools that check validity and well-formedness of digital assets are not yet readily available. A promising open-source tool for identification and validation, JHove, is currently the only tool available, but it needs to be developed further.[3]

Before the third aspect of digital preservation, how to ensure permanent access, is described, a practical example of long-term DAM at the national library of the Netherlands will illustrate the issues described above.

CASE STUDY: THE KB e-DEPOT

The Koninklijke Bibliotheek is the national library of the Netherlands and a deposit library. Even though there does not exist a deposit law in the Netherlands, based on agreements with publishers, KB receives one copy of every publication that is printed in the Netherlands. In the early 1990s, KB started to receive digital publications. A new model for depositing, archiving and access was needed. In cooperation with publishers, the KB set up an archive for e-journals that grew into an international "safe place" for digital publications.

At first, it was clear that a system for digital preservation could not be purchased "off-the-shelf" and small-scale experiments with applications and systems for managing digital assets publications were started. At that time, it was impossible to have a clear picture of all the

problems, but a practical approach was chosen to gain hands-on experience. In 2000, knowledge of what a full-scale deposit system should look like had grown, and a tender procedure to purchase a system was organized. IBM made the best offer and, together with them, the KB staff developed the e-Depot. This system was taken into production in early 2003. The archiving system is built according to the OAIS Reference model and is dedicated to long-term storage.[4]

The technical heart of the e-Depot is now an IBM product: DIAS (Digital Information Archiving System). DIAS is a solution that consists of several components (like Content Manager and DB2) and is scalable and flexible. At the KB, DIAS is integrated with the library infrastructure and connected to tailormade additions that allow the KB to automatically load and store batches of e-journals 24 hours per day.[5]

Since the system was taken into production, more than eight million articles have been stored. These are articles of international publishers of e-journals. When we started to work on digital preservation, we worked in close cooperation with Elsevier Science, the world's largest publisher in Science, Technology and Medicine (STM) publishing. At first we stored those journals that, in their paper version, were printed in the Netherlands. Very soon this became "would have been" printed in the Netherlands, because more and more, journals were published in digital form only. It became clear that geographical boundaries did not exist anymore in the world of digital publishing. At the same time, Elsevier was very pleased with the way they could store their journals with the KB, so together we set up an agreement to store all their e-journals instead of just their Dutch journals. This agreement was an example for agreements being signed with other publishers, like Kluwer. But these were still publishers that had originally been Dutch publishers. This was different for the next publisher to sign a contract with the KB: BioMed Central, a publisher that did not have roots in the Netherlands. The contract with BioMed was also remarkable in another way: BioMed is an open access publisher. Until then, publications in the e-Depot could only be accessed on the premises of the KB. After all, the e-Depot is there for long-term preservation and not to compete with the publisher's business. BioMed's publications are open access and the copies that are maintained in the e-Depot are also accessible online. Other publishers followed: Blackwell, Oxford University Press, Taylor and Francis, Sage, Springer, Brill Academic Publishers. On the basis of these agreements the e-Depot will eventually hold ten million articles. The annual increase in the number of articles from these publishers will be around 400,000 (Figure 4).[6]

Why is the Dutch e-Depot the international digital archive (Safe Place) for e-journals? The e-Depot holds mainly scientific articles in STM. These are the records of science, which should be preserved and kept accessible for future generations. To set up a "trusted digital repository" to preserve these records, substantial investments are required, not just for building and maintaining the system, but also for continuous research. Technologies are changing constantly and to adjust preservation plans accordingly, an expert team has to support the repository. To ensure permanent access to digital assets, it has to be taken up by organizations that

Figure 4: Tape storage of the e-Depot

have a long-term commitment to preservation. The substantial effort cannot be expected of every library or archive in the world and that is why permanent archiving will be taken care of by a limited number of institutions: Safe Places.[7]

HOW CAN PERMANENT ACCESS BE ENSURED?

When the bits and bytes are stored safely, the question of future representation comes up. All digital records are stored in a certain file format. To be able to render these formats, we need software that can read the bits and bytes. And this software depends on a specific kind of operating system.

The system KB built with IBM did not yet include full-scale preservation functionality. So at the same time the system was taken into production, a research project was started with building up a research and development group to work on preservation strategies. This group develops prototypes, carries out new projects and works closely together with international colleagues. Digital preservation research is not a one-time effort, but requires permanent commitment.

The application of strategies for the future representation of stored objects has also become an issue of growing interest, with tests being carried out with migration and experiments being prepared with emulation. In general, these are the two main ways to approach permanent access. Migration (or conversion) focusses 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. Emulation does not focus on the digital object, but on the environment in which the object is rendered. It aims at (re)creating the old environment on a new platform, so the digital object can be rendered in its original format.

Both strategies have pros and cons. Migration brings the risk of minor errors that can still have an enormous impact. Just think of mathematical formulas: if a letter is placed a little bit higher or lower than it was supposed to, the whole meaning is different. But still, migration will offer the future user an opportunity to read old files in a format that he or she is used to. This is in contrast to emulation: it sounds very appealing to be able to render a file in its original format, but who will remember how to operate an old program. A good example is WordPerfect: we used to memorize the meaning of function keys on the keyboard, but nobody knows about them anymore.

Different types of digital objects bring different challenges. For publications, retaining the appearance is very important. For digital records, the look and feel may be less important than the content and context. For permanent access, no single strategy will be sufficient, but a combination of strategies has to be developed. How and when to apply these strategies is determined through preservation planning, a subject that is the focus of current research.

At KB, both emulation and migration strategies are developed. Already in 1999, KB started to work on emulation for digital preservation with one of the world's experts in this field: Jeff Rothenberg.[8] From 2002 to 2004, KB and IBM developed another approach: the Universal Virtual Computer,[9] and in 2005, a new project on emulation was started with the Dutch National Archive (NA). Emulation is not new. In the gaming community, emulators have been developed for a long time, to allow us to play old games on new computers. The problem with these emulators is that they have to be rebuilt for every new platform and that they have not been built with future accessibility in mind. The emulator that KB and NA are building now is a modular emulator, which is a full software emulation of current hardware. Its modularity is the insurance for future use: only some of the modules will have to be adapted to rerun this emulator on future platforms. In May 2007, a first prototype will be delivered and will be made available to the open source community (Figures 5 and 6).[10]

Migration may seem straightforward to develop. High-quality migration tools that live up to the requirements of digital preservation are, however, much harder to find. In 2006 KB started with a research project on migration, and very soon it became clear that migration tools that were tested for preservation were hard to find.[11] Together with international colleagues, KB is now working on tools to test migration results and on setting requirements for durable migration.

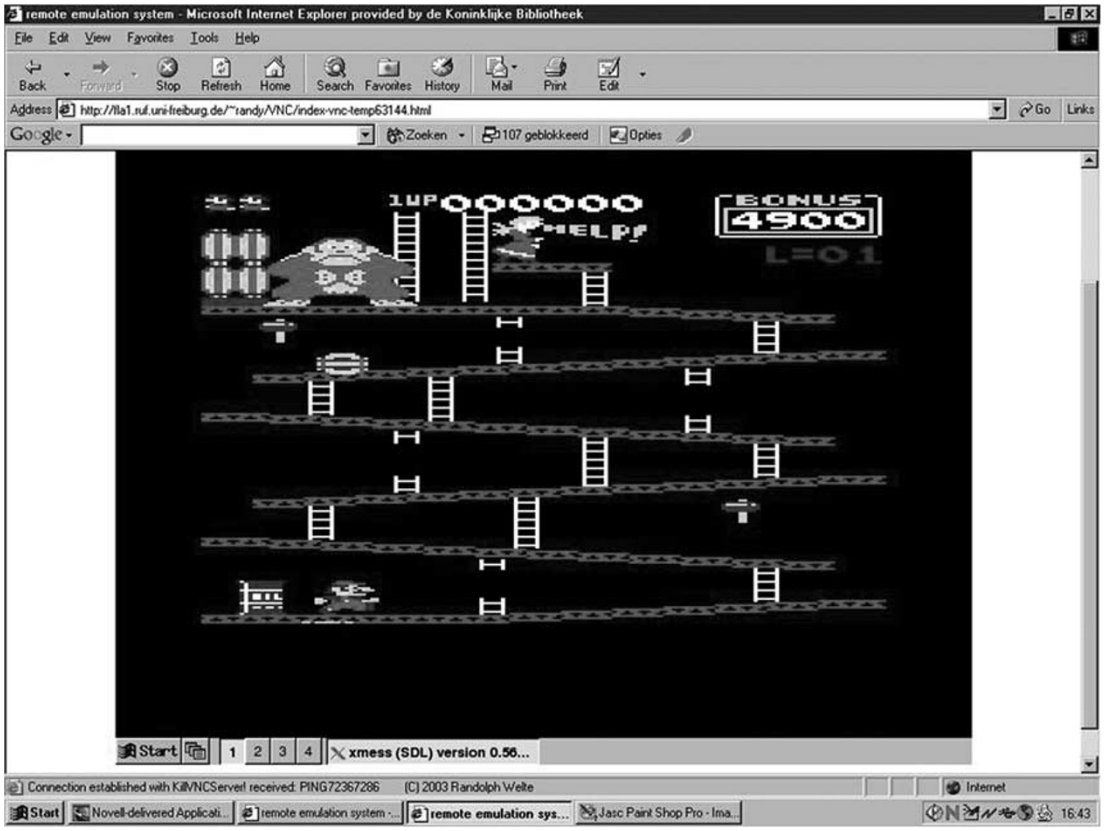

Figure 5: Emulated game on current platform

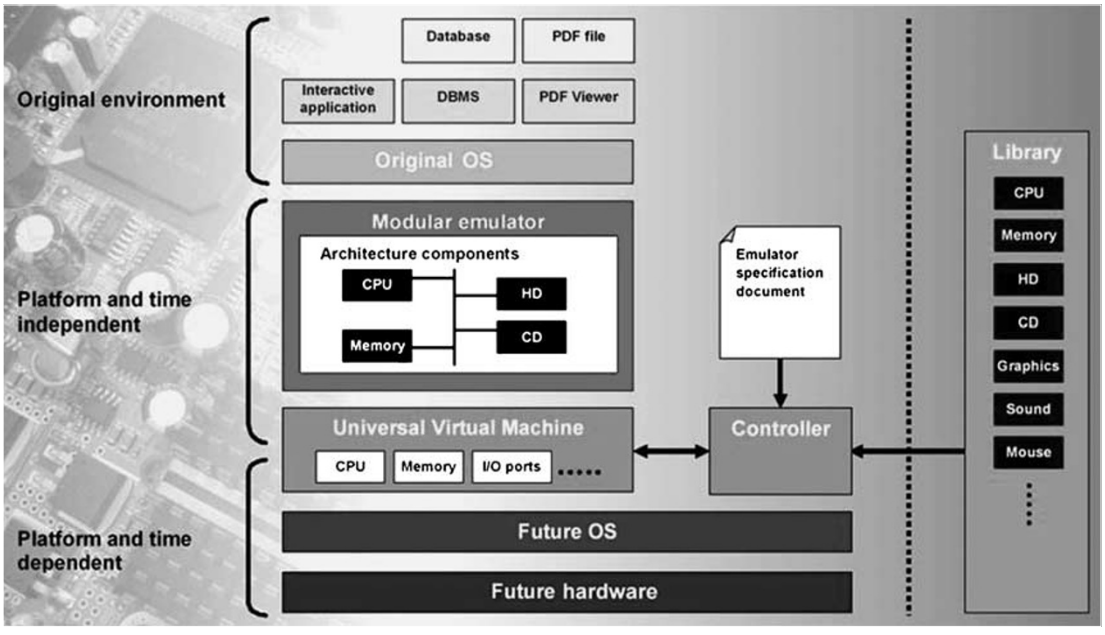

Figure 6: A model of the modular emulator

INTERNATIONAL PROJECTS

Progress in the area of digital preservation largely depends on international cooperation. In the early years, the library and archival community set up informal networks to discuss issues and exchange experiences. Later on, this kind of cooperation was stimulated by the start of specific international projects funded by governments, science foundations or the European Commission. Of course, funding is very important. As more and more official records, including government communication at all levels, are not available on paper anymore but exists only in digital form, the need for archiving procedures is now imminent.

The recent years have seen a broadening group of involved parties. The cultural heritage sector has joined forces with technical data-intensive research institutes. Commercial companies have become involved as well. Stakeholders, like banks, insurance companies and the pharmaceutical industry, have started to work on the problem. The software developers have become interested not only in developing solutions for digital preservation, but also in making new formats and programs more durable.

An important incentive for funding comes from the scientific community. Science depends on access to information, which can only be ensured for the future by taking digital preservation into account. When scientific data are the object of preservation, this topic is often referred to as curation. Data curation deals with the whole process of data management, preservation and future accessibility. To convince the European Commission of the need for funding, a cost calculation was made.[12] EU member countries produce around five billion documents per year, and of this total, around 2 per cent (100 million documents per year) comprise information that is worth archiving. Around two million documents out of this subtotal are held in formats that constitute a long-term preservation risk. The costs incurred in creating these endangered documents have been estimated at around 3bn Euros.
A compelling argument: a relatively small investment in digital preservation research can save billions of Euros…

While the need for digital repositories was recognized years ago, funding has now become available for large-scale projects on digital preservation research. In the US, a growing number of projects were started, stimulated by the Library of Congress's NDIIP program (National Digital Information Infrastructure and Preservation Program.[13] In Europe, three new projects on digital preservation have been started in 2006. CASPAR deals with long-term curation of digital art and scientific data, and DPE (Digital Preservation Europe) is a project to coordinate and disseminate between ongoing research and projects throughout Europe.[14] The recently started PLANETS project will develop systems and tools to support the accessibility and use of digital information.[15] KB initiated this project together with the British Library and 14 European archives, libraries, research institutes and technology partners. The project will deliver tools for the characterization of digital objects and tools for migration and emulation, not just to the PLANETS partners, but to every interested party, through open source technologies. A digital preservation test bed will be set up to test these tools, and also to serve as a controlled testing environment for benchmarking and for testing commercial solutions. The ultimate product of PLANETS will be an automated preservation-planning module that will allow organizations to plan and execute preservation plans.

TO CONCLUDE

With awareness growing, research being funded and digital archives being set up, digital preservation has come to the foreground. The whole notion of software obsolescence and its consequences has, however, not yet caught the full attention and interest of commercial software developers. Their interest lies with the development of new, better and faster programs and that is how it should be. But slowly, new developments have started that will facilitate future preservation (for instance, the development of Open Office, Open Document and PDF7 as an ISO standard). Hopefully in the near future, durability will become an attractive and marketable value as well, and maybe then, DAM systems will offer preservation functionality as one of their basic features as well.

References and Notes

1 Oltmans, E. and Lemmen, A. (2006) 'The e-Depot at the National Library of the Netherlands', *Serials*, Vol. 19, No. 1, pp. 61–67.

2 PREMIS (PREservation Metadata: Implementation Strategies). Working Group,Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, May 2005, United States of America. http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

3 JHove: JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove/.

4 Steenbakkers, J. F. (2005) 'Digital archiving in the 21st century: Practice at the National Library of the Netherlands', *Library Trends*,Vol. 54, No. 1, pp. 33–56. pp. 33–56. (Digital Preservation: Finding Balance).

5 See for a description of the e-Depot at the KB and the DIAS system:. http://www.kb.nl/e-depot and http://www-5.ibm.com/nl/dias/.

6 Oltmans, E. and van Wijngaarden, H. N. (2006) 'The e-Depot at the National Library of the Netherlands', *Library Hi Tech*, Vol. 24, No. 4, pp. 604–613.

7 van Trier, G. (2006) 'Permanent access to the records of science — The international role of the e-Depot at the Koninklijke Bibliotheek, National Library of the Netherlands', *Liber Quarterly* 16, http//liber.library.uu.nl/cgi-bin/pw.cgi/articles/000177/article.pdf

8 Rothenberg, J. and (RAND-Europe) (2000) *An Experiment in Using Emulation to preserve Digital Publications*, NEDLIB Report Series 1, The Hague, 2000. http://nedlib.kb.nl/results/NEDLIB emulation.pdf.

9 Oltmans, E., van Diessen, R. J. and van Wijngaarden, H. N. (2004) 'Preservation Functionality in a Digital Archive', *Proceedings of the Joint Conference on Digital Libraries*, Tucson, Arizona, USA, 11 June 2004.

10 See for the current emulation project:. http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie.html and van der Hoeven, J. R. and van Wijngaarden, H. N. (2005) 'Modular emulation as a long-term preservation strategy for digital objects', *International Web Archiving Workshop (IWAW'05)*, Vienna, Austria, 2005.

11 van Wijk, C. Starting point migration research http://www.kb.nl/hrd/dd/dd_projecten/Starting_Point_Migration_Research.pdf.

12 As estimated for the British Library in preparation of the Planets proposal.

13 NDIIPP.http://www.digitalpreservation.gov/index.html.

14 CASPAR: Cultural,Artistic and Scientific knowledge for Preservation, Access and Retrieval. http://www.casparpreserves.eu/ DPE: Digital Preservation Europe. http://www.digitalpreservationeurope.eu/.

15 PLANETS: Preservation and Long-term Access through Networked Services. http://www.planets-project.eu/.