LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich, Printed in Germany

Cost Models in Digital Archiving: An Overview of Life Cycle Management at the National Library of the Netherlands

by ERIK OLTMANS

ABSTRACT

The number of institutions that are either considering the implementation of a digital archive or already started with an operational digital archive service is increasing. While technological and organisational challenges are well-studied and in few cases even well-defined, the subject of costs often remains untouched, which in many cases prevent organisations from new initiatives. In this paper two leading digital preservation techniques – migration and emulation - will be discussed in terms of life cycle management and associated cost models. Both techniques are studied and considered for implementation at the Koninklijke Bibliotheek, National Library of the Netherlands.

INTRODUCTION

Digital publishing is causing publishers, research institutions and libraries to develop new policies, new infrastructures and techniques, and new business models as well. A major problem is that, at the same rate at which our world is becoming digital, digital information is threatened. New types of hardware, computer applications and file formats supersede each other, making our recorded digital information inaccessible in the long term. The Koninklijke Bibliotheek (KB) has, jointly with IBM, developed and implemented an OAIS-based deposit system: the *e*-Depot. Moreover, the KB signed archiving agreements with major scientific publishers for permanent storage of their digital materials. An important issue in digital archiving is long-term access: how can we guarantee permanent access to digital publications while software and hardware are constantly changing? This issue strongly relates to the object's life cycle management, as wrongful life cycle management might yield unavailability of the digital object in the long run.

In this paper, I'll discuss life cycle management issues at the Koninklijke Bibliotheek. Moreover, I'll discuss two prominent digital archiving models in terms of life cycle management and associated costs. Specifically I'll compare migration and emulation strategies, while discussing their associated life cycles and corresponding cost models.

I'll demonstrate that applying emulation models may be more efficient in terms of life cycle management (and thus costs) compared to migrations models.

THE KB *e*-DEPOT

In 1999 the KB specified the system requirements for a full-scale deposit system, which were based on the ISO standard for digital archives: the Open Archival Information System (OAIS, 2000). As a result of a European tender procedure in 2000, the KB contracted the development of the deposit system to IBM in The Netherlands. In December 2002 the system was delivered to the KB. IBM constructed the system using as many off-the-shelf components as possible, such as WebSphere, DB2, Tivoli Storage Manager, and Content Manager, and branded it under the name Digital Information Archiving System (DIAS). By using DIAS, the KB maintains the service called the *e*-Depot. See Oltmans & Van Wijngaarden (2004) for a complete description and Steenbakkers (2002) for more details about the history of the KB *e*-Depot.

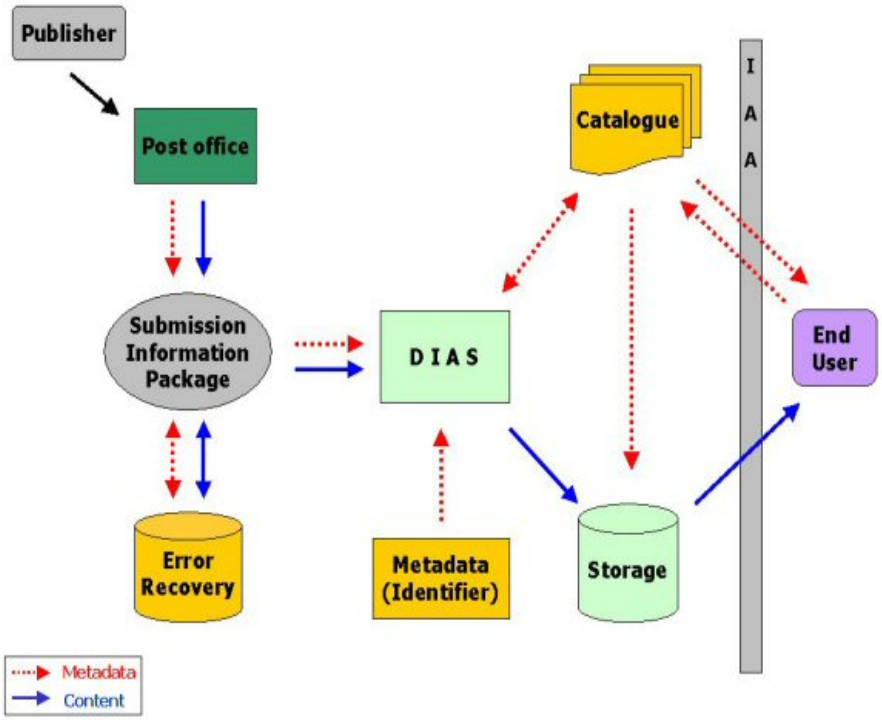
The KB has developed a workflow for archiving electronic publications and has implemented the other parts and interfaces of the infrastructure in which the deposit system is embedded. This infrastructure consists of a variety of functions for:

- validating and pre-processing electronic publications
- generating and resolving unique identifiers
- searching and retrieving publications
- identifying, authenticating and authorising users

The process of loading consists of pre-processing and ingesting the digital content. Two types of electronic publications are stored in the *e*-Depot: offline media such as CD-ROMs (also referred to as "installables ") and online media such as electronic articles (Oltmans, 2003). The ingesting of installables is a time-consuming process. Firstly, the CD-ROM is completely installed on a dedicated Reference Workstation (RWS), including all additionally needed software like image viewers or media players. A snapshot of the installed CD-ROM is then generated into an image, including the operating system on which it is installed. After manual cataloguing, this image is subsequently loaded into the *e*-Depot. If an end user wants to view a particular CD-ROM, the entire image is retrieved from the *e*-Depot, and installed on a RWS. By including the operating system in the stored package, the CD-ROM is guaranteed to work - also in future conditions involving new operating systems, as long as the hardware remains the same.

The second type of publications currently processed is online media. These publications are either sent to the KB on tapes or DVDs, or they are downloaded via FTP. In both cases, publications ready for ingestion end up in an electronic post office in which they are validated. At this stage the content of the submitted publication is validated in regards to its authenticity and well-formedness, based upon earlier agreed to specifications. If the material does not match the checksum (or if other errors occur), the content is passed to a database for error recovery. If the content appears to be valid, content and metadata are combined to form Submission Information Packages (SIPs). These SIPs are then processed by DIAS. See Figure 1 for a complete overview of the data flow.

Figure 1: General e-Depot Data Flow



DIAS ingests both the content and the metadata, converting the publisher's bibliographical descriptions into the KB internal format and adding a National Bibliographic Number. This number functions as the unique identifier of every digital item stored in the system. The content itself is stored in the *e*-Depot, while the metadata

is stored in the KB catalogue. Technical metadata is stored and maintained by using the Preservation Manager, developed in collaboration with IBM (Oltmans, van Diessen & Van Wijngaarden, 2004). End-users may query the online catalogue and retrieve the full text of the publications. In the case where access restrictions are imposed by the publisher on the content, this may occur only after a process of identification, authentication and authorisation (IAA). The *e*-Depot itself cannot be accessed directly, but passes relevant publications to the end-user after verification.

Five major publishers have signed unique archiving agreements with the KB on long-term digital archiving of their electronic publications:

- Elsevier Science
- Kluwer Academic
- BioMed Central
- Blackwell Publishers
- Taylor & Francis

Agreements with more publishers will be signed soon. At this moment the digital publications of these publishers are being loaded into the *e*-Depot, involving more than 2,500 journals, containing over 6 million articles. In case of publications that are processed both in digital and printed form, the KB has decided to process only the digital manifestation of the publication. With respect to version management of electronic publications, the *e*-Depot is able to deal with updates, renewals, and withdrawals or retractions. Updates of electronic publications are sent to the KB with different time stamps compared to the original submissions. These authentic publications are not discarded from the system, but the original metadata is temporarily withdrawn, so that only the updated material will be found in the central catalogue. This way, the KB does preserve the complete records of science, but allows publishers to ask for temporarily withdrawal of specific articles. Once an article is in the system, it will never be discarded or deleted (Steenbakkers, 2003).

New acquisition methods are necessary in order to obtain the electronic publications, like extensions of the OAI harvesting protocol (Lagoze, Van de Sompel, Nelson, Warner, 2002 and Jerez, Liu, Hochstenbach and Van de Sompel, 2004). Another issue is that there are lots of publications on the Internet that will require some form of web harvesting. A complicating matter in this respect is the lack of automatic delivery of bibliographic descriptions of the harvest.

LONG-TERM PRESERVATION AT THE KB

New types of hardware, computer applications and file formats are constantly being developed, making digital information inaccessible. Even if the hardware or the carrier-media does not deteriorate, the technology to access the information will inevitably become obsolete. Preservation or permanent availability of digital information is one of the processes which is dramatically affected by the change to an all digital world.

In general there are two main digital preservation approaches. 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 on 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. Both models are considered for implementation at the KB, and will be discussed in brief.

1. Migration

When choosing for migration or conversion, file formats will be converted into new formats as soon as the original formats run the risk of getting obsolete. For example, if technology scans indicate that PDF version 1.1 will soon be out of date, all files in the digital archive of format PDF 1.1 have to be converted into for example the format PDF 1.4. This way, the digital publications will be prepared for rendering for another period of time, until the format PDF 1.4 runs the risk of getting obsolete itself. In that case another migration programme has to be carried out.

An advantage of migration as a digital preservation strategy is that electronic publications will be always available in the form that is mostly used, e.g. PDF, and that hardware and software will be able to render these formats without serious problems. Older documents that are properly migrated will be available for some time, and their electronic content can be used for copy and reuse. A major drawback might be that while converting documents from one form to another, lay out and – worse - data might get lost. If preserving the original 'look & feel' of the document is important, then migration might not be the best way to do this. Migration is necessary for every single document in the collection, and should preferably be carried out each time a serious update of the file format is available. It may be straightforward to convert from version A to version B, but converting from version A to version C or D might be a complicated matter (Caplan, 2004). This may not be necessarily true, but we will never know for sure. Therefore, we have to constantly study conversion programmes and execute them when possible.

2. Emulation

Emulation, on the other hand, does preserve the authentic document, and provides the user with an emulation tool, so that 'old' software and 'old' viewer programs can be used to render this original document. An emulation tool generates an authentic view by launching the original viewer in the context of the original platform. It is the emulation tool, which makes the original viewer and the original platform work in future environments.

An advantage of emulation is the fact that the original 'look & feel' of the publication will be preserved. Like with preserving books, the authentic instantiation will be there to be rendered, in contrast to migration in which possible other instances are used than the original one. However, a serious drawback is the complexity of developing and maintaining an emulation tool. In the future, we have to maintain several emulation tools, and it cannot be proven that these will always work on future computer platforms.

The need for both emulation and migration

In the next paragraph emulation and migration techniques will be compared in terms of life cycle management and associated costs. In general, we will see that costs arguments will be in favour of emulation techniques. However, the KB does not favour one particular strategy over another. There are arguments both for preserving the original 'look & feel', as well as for converting documents into 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. Therefore, we promise publishers to keep the original bit stream they send to the KB. Emulation tools are needed in order to render these publications in the same way as they were published. Secondly, for end users who want to access publications according to the 'original look & feel' as they were published, we also need emulation tools (Van Diessen & Van der Werf, 2002).

On the other hand, there is also a specific need for converting documents into the most actual standard. For future end users who want to have access to publications according to the standards and functionalities of that time, migration might be needed so as to allow end users to copy and reuse data.

In short, the KB does not prefer one strategy to another. In fact, both are studied and considered for implementation. However, emulation as a digital preservation technique requires more substantial research compared to migration. This is why the most advanced developments at the KB in this respect relate to emulation. The first prototype of an emulation tool based on the Universal Virtual Computer concept is now available. For more information about this project, we refer to Lorie (2002) and Van Wijngaarden & Oltmans (2004).

LONG-TERM PRESERVATION AND ASSOCIATED COSTS

Any particular digital preservation strategy strongly determines the life cycle management of digital publications and thus the associated costs as well. In order to specify the long term costs of a digital archive, we therefore need to understand the implications of choosing for a particular preservation strategy. On the other hand, costs arguments may in turn determine (or limit) the choice for certain preservation strategies.

Emulation models require more initial investments compared to migration models. Emulation tools have to be developed, and this requires serious R&D, including technical skills to implement the concepts. What is more, emulation tools have to be maintained over time, which also requires investments in both researchers and implementations. However, emulation tools can be shared among institutions which makes it possible to share the costs of R&D investments as well (I owe this argument to Raf de Keyzer).

Migration on the other hand is relatively cheap in the sense that many conversion tools are available, and executing a conversion programme is a relatively straightforward task. However, migration by definition applies to the entire collection repetitively: each and every single object in the digital archive has to be converted, again and again. This means that the bigger the archive gets, the more expensive migration will be. This is in contrast with emulation: emulation tools apply to the collection as a whole, and no special action is needed in case a digital object is rendered.

In order to specify this difference exactly, I will use the cost model formulae as presented by Shenton (2003). Shenton specified for instance the costs of preserving a non-digital monograph over time as follows:

$$K(t) = s + a + c + pl + hl + p(t) + h(t)$$

Where K(t) is the total cost of holding an item for a period of t years, where s=selection, a=acquisition processing, c=cataloguing, pl=initial preservation, hl=initial handling, p(t)=longer-term preservation, h(t)=storage.

By applying this formula, it can be easily calculated what the long-term costs are of preserving a monograph. Moreover, this formula makes it possible to calculate the downstream costs if for instance, an additional 100,000 Euro would be available on acquiring monographs.

It would be of significant interest if such a formula would also be available for electronic publications. Obviously, not all the variables in Shenton's formula are applicable to digital objects. A fundamental part of every formula that applies to electronic materials would consist of:

K(t,a) = s + ing + h(t,a)

Where K(t,a) is the total cost of holding a objects for a period of t years, where s=selection, i=ingestion, and h=storage.

The selection process is quite obvious: it consists of acquiring the objects and preparing them for further processing. The ingestion process consists of the automatic processing of the digital objects by some sort of software program. It should among others convert the associated metadata into a usable format, and store the digital objects on some sort of storage system. The storage costs itself are for purchasing the storage media, the media refreshment, and maintaining some sort of database management system. There is a direct relation between the overall costs and both the number of items and the number of years they are preserved. More items will cost more, and longer storage will cost more as well.

A part of the formula above will be fundamental for two other formulae that I propose. However, I will not calculate the costs for selection and ingestion, for two reasons: First of all, the costs for selection and ingestion will be the same for both emulation and migration; in other words they will not influence the relative difference. Secondly, it is quite difficult to estimate these costs, as it depends on the archiving agreements with the publishers (selection) and the type of software that is in place (ingestion). Both will differ considerably, depending on the circumstances. Therefore I focus on storage costs and the dedicated costs for either migration or emulation.

The first formula, for migration, is as follows:

K(t,a) = h(t,a) + m(t,a)

Where K(t) is the total cost of holding a objects for a period of t years, where h=storage, m=migration.

A new variable is introduced that expresses the costs of migrating an object. The costs of migrating digital objects is dependent of time t (the longer we preserve the objects, the more often we have to convert them) and of the number of objects a (the more objects in the archive, the more conversion actions have to be executed).

The formula for calculating the emulation costs is as follows:

K(t,a) = h(t,a) + E + e(t)

Where K(t) is the total cost of holding a objects for a period of t years, where h=storage, E=setting up initial emulation tool, and e(t)=emulation over time.

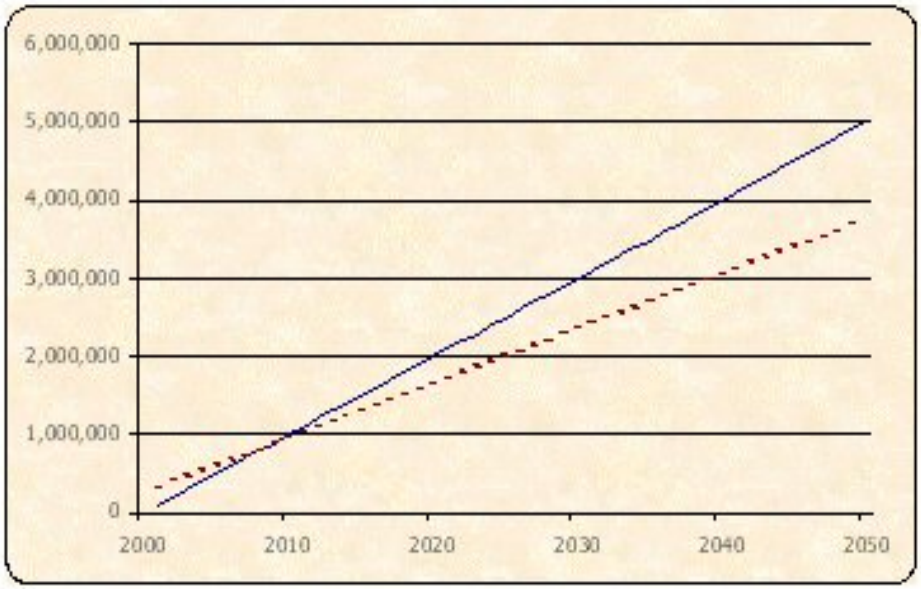
Two new variables are introduced: The costs for developing an emulation tool are expressed by E, while the yearly maintenance is expressed by e. The maintenance is not dependent of the number of objects: emulation tools apply to the entire collection, and no special action is needed when rendering an object in the digital archive. However the emulation tools need to be maintained over time, which makes the maintenance costs dependent on the number of years.

Having the first, primitive, formulae in place, we can now associate specific variables to costs. In the next table all variables and their corresponding costs are shown:

h	Storage	0,05 per digital object
m	Migration	0,05 per digital object
E	Emulation tool	250.000
e	Emulation	20.000

The costs for storage and migration are based on figures from the literature (Fox, 2002: storage costs), but may vary. A complicating matter in this respect is that both figures usually express costs per Megabyte or Gigabyte. Considering these figures, the costs of preserving 1,000,000 objects for a period of 50 years can now be calculated, both when applying migration and emulation as the leading digital preservation technique. The graph below demonstrates these costs:

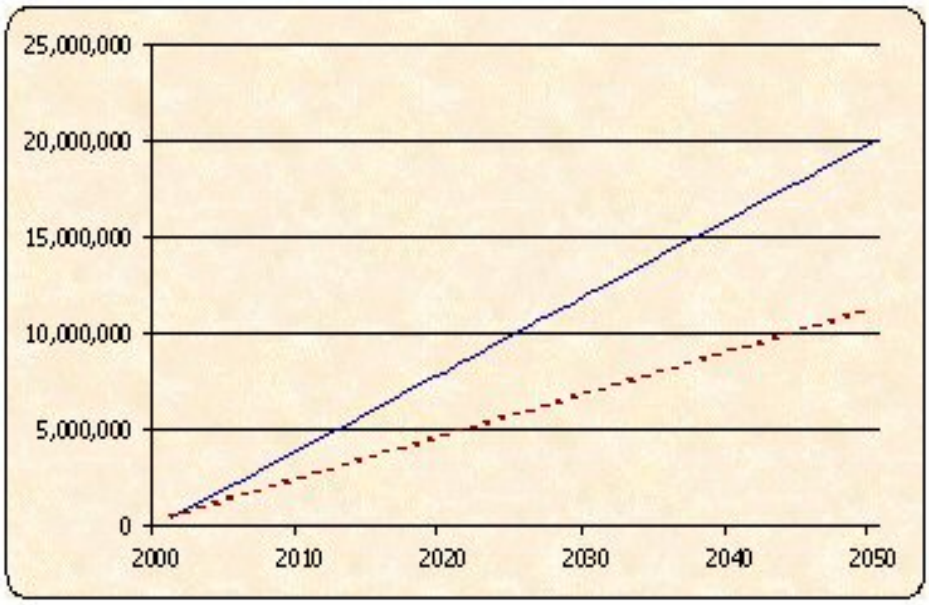
Figure 2: Costs for Migration and Emulation (dotted line) for maintaining an archive of 1,000,000 digital objects over a period of 50 years.



The initial investments of setting up an emulation tool yield high costs in the first 10 years. But soon after that, the migration costs start to increase faster than the emulation costs. In 50 years, the migration costs are approximately 25% higher than the emulation costs.

The difference between emulation and migration is clear, and can be completely explained by the difference in costs coverage of the two techniques. As the size of the collection directly affects the migration costs, it will be clear that the bigger the archive gets, the higher the migration costs will be. This effect is demonstrated in the second graph. It covers the same period of years, but the size of the collection now is 4.000.000 rather than 1.000.000.

Figure 3: Costs for Migration and Emulation (dotted line) for maintaining an archive of 4,000,000 digital objects over a period of 50 years.



Compared to the graph in Figure 2, the size of the collection is four times as big, while the migration costs are now approximately twice as high as the costs for emulation.

CONCLUSIONS

In this paper I discussed life cycle issues in the context of long-term preservation of digital objects. At the KB a fully operational digital archive is in place, and this archive, the *e*-Depot, provides the context for studies to a number of digital preservation techniques. Both emulation and migration are discussed, as the KB needs to provide access to the original publication as sent by the publisher, while at the same time the KB wants end users to access converted materials according to the most recent standard and functionalities.

Emulation and migration are inherently different in terms of life cycle management that causes a serious difference in costs. While migration applies to all objects in the collection repetitively, emulation applies to the entire collection as a whole. This makes emulation most cost-effective in cases of large collections, despite the relatively high

initial costs for developing an emulation tool. When considering the fact that only small fragments of digital archives need to be rendered in the long run, it may turn out that from a financial perspective emulation techniques will be more appropriate for maintaining larger archives.

In this overview, I deliberately neglected a number of important issues. First of all, I did not consider the fact that migration may get less expensive in costs if the number of objects to be converted gets considerably high (economies of scale). What is more, I calculated the costs of migration in terms of the number of objects, while it would also make sense to calculate the costs in terms of Gigabytes. The problem here is that it is not clear how many objects there are in a Gigabyte of PDF files. Therefore it is clear that the issue of costs in digital archiving needs more practical experience, and more study. The results presented here are a first step for determining life cycle issues in digital archiving, and may serve advanced studies that reach for a complete understanding of cost models in long-term preservation.

Acknowledgements

This overview could not have been written without the valuable input of Johan Steenbakkers, the initiator of the *e*-Depot. I also want to thank Hilde van Wijngaarden for her substantial contribution.

REFERENCES

Caplan, Priscilla: "Building a digital preservation archive: Tales from the front ". *VINE*, 34(2004)1, 38-42.

Van Diessen, R.J. and T. Van der Werf-Davelaar: *Authenticity in a Digital Environment.* IBM/KB Long-Term Preservation Study Report Series number 2. Amsterdam : IBM Netherlands, 2002.

Fox, Peter: "Archiving of electronic publications - some thoughts on cost ". *Learned Publishing*, 15(2002) 1, 3-5.

Lorie, R.: *The UVC: a method for preserving digital documents - proof ofcConcept.* IBM/KB Long-Term Preservation Study Report Series number 4. Amsterdam : IBM Netherlands, 2002. http://www.kb.nl/kb/hrd/dd/dd_onderzoek/reports/4-uvc.pdf

Jerez, H.N., X. Liu, P. Hochtenbach and H. Van de Sompel: "The Multi-faceted Use of the OAI-PMH in the LANL Repository ". Proceedings of the Joint Conference on Digital Libraries. Tucson, Arizona, June 7-11, 2004.

Lagoze, C., H. Van de Sompel, M.Nelson and S. Warner: The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html

OAIS 2002: *Reference Model for an Open Archival Information System (OAIS)*. Blue Book, Issue 1, Consultative Committee for Space Data Systems . January 2002. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

Oltmans, E. and H. Van Wijngaarden: "Digital Preservation in Practice: The e-Depot at the Koninklijke Bibliotheek ". *VINE*, 34(2004)1, 21-26

Oltmans, E.: "Legal Deposit of Digital Materials ". *Liber Quarterly*, 13(2003)3/4, 281-289. http://liber.library.uu.nl/publish/articles/000031/index.html

Oltmans, E., R.J. Van Diessen and H. Van Wijngaarden: "Preservation Functionality in a Digital Archive ". Proceedings of the Joint Conference on Digital Libraries, Tucson, Arizona, June 7-11, 2004.

Shenton, Helen: "Life Cycle Collection Management ". *Liber Quarterly*, 13(2003)3/4, 254-272. http://liber.library.uu.nl/publish/articles/000033/index.html

Steenbakkers, J.F.: *The Road to e-Deposit at the Koninklijke Bibliotheek.* The Hague : Koninklijke Bibliotheek, 2002. http://www.kb.nl/kb/hrd/dd/dd_links_en_publicaties/publicaties/eskb-roadtoe-deposit3_20aug02.pdf

Steenbakkers, J.F.: "Permanent Archiving of Electronic Publications ". *Serials*, 16(2003)1, 33-36

Van Wijngaarden, H. and E. Oltmans: "Digital Preservation and Permanent Access: The UVC for Images ". Proceedings of the Imaging Science & Technology Archiving Conference, San Antonio, USA, April 23rd 2004. Also available at: http://www.kb.nl/kb/hrd/dd/dd_links_en_publicaties/publicaties/uvc-ist.pdf

WEB SITES REFERRED TO IN THE TEXT

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

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

Koninklijke Bibliotheek. Nationale Bibliotheek van Nederland. http://www.kb.nl/