67th IFLA Council and General Conference
August 16-25, 2001

Code Number:	**185-168-E**
Division Number:	**VI**
Professional Group:	Preservation and Conservation
Joint Meeting with:	Information Technology
Meeting Number:	168
Simultaneous Interpretation:	**-**

Preserving Digital collections

Titia van der Werf-Davelaar
Royal Library
The Hague, Netherlands

Abstract:
In this paper Titia van der Werf discusses progress made in the newly emerging discipline of digital preservation. The discussion takes a deposit library perspective and demonstrates how specific application domains and specific material types have specific preservation management requirements. The author shows that, by contrast, IT-based technologies are not domain specific and that, when applying such technologies it is best to take a generic approach. She illustrates this with work done in the context of OAIS and NEDLIB. However she observes that digital preservation technologies are not mature enough yet to be implemented as standard solutions - but goes on to argue that this should the general aim of all communities involved and that the IT-industry should be triggered into this direction.

Introduction

Digital library collections rapidly grow to contain an ever increasing diversity of material - published and unpublished, commercially available, collected as institutional and personal donations, born-digital and digitised, distributed via offline media, hosted by third parties and captured from the web.
Does management of these collections require special effort and what issues are involved? In particular, what issues are involved when archiving and preserving digital collections? I have been asked to give you a review of the most current thinking on this subject, in as global a fashion as possible.

In this paper I will discuss progress made in the newly emerging discipline of digital preservation, drawing from the experience and insights gained during the NEDLIB project [1] and the local implementation of a digital deposit system at the Koninklijke Bibliotheek in the Netherlands [2]. I will also draw thankfully from fruitful interactions I have had in recent years with colleague institutions such as the British Library, the National Library of Australia, the Library of Congress, the Yale University Library, the National Archive of the Netherlands and others.

Digital preservation

Let me set the scene by clarifying the terminology used in this paper. Although the proposed title of my presentation was originally termed 'Archiving digital collections', I prefer to use the title 'Preserving digital collections'. The library and archiving communities often use these terms interchangeably to cover all aspects relating to long-term digital storage and preservation. The term archiving, however, as it originates from the computer sciences, is related to storage backup and maintenance processes without any real long-term preservation perspectives.

The subject matter of digital preservation is not mature enough yet, still it has become the talk of the day. Terms and concepts such as 'archival repositories' and 'digital libraries' proliferate. They are applied to any digital collection, regardless of the nature of the custodial institution. As different communities increasingly use the same terms in different ways, effective communication becomes difficult. Cross sector and cross-disciplinary approaches and the trend towards generalisation, do not always promote a better understanding of issues. Discussion of the value of the archival perspective for the digital library has led to academic speculation about concepts such as the 'authenticity' and 'reliability' of digital objects. Discussion of preservation issues concerning born-digital and digitised material is another potential pitfall. While born-digital content is depicted as fragile and prone to 'digital death', digitisation is heralded as a means to preserve content on paper and analogue media. The different management requirements for digitised collections and for born-digital collections contrast with their seemingly similar nature. A little untangling is needed here. I propose to do this by taking examples from the library perspective. I will demonstrate how specific application domains and specific material types have specific preservation management requirements. It is important to understand that, by contrast, IT-based technologies are not domain specific and that, when applying such technologies it is best to take a generic approach. I will illustrate this with work done in the context of OAIS [3] and NEDLIB. Finally it is important to realise that digital preservation technologies are not mature enough yet to be implemented as standard solutions - but that this should be our aim and that we should trigger the IT-industry into this direction.

Cross sector and cross disciplinary approaches

The technologies used for digital activities such as imaging, archiving and preservation are applicable to all sectors and disciplines, regardless of the specific nature of the application area. The technicalities of a scanning device, for example, its performance in terms of image resolution and compression, are equally relevant to digitisation programmes of libraries, archives and museums. Sharing expertise and knowledge at this level is rewarding, as can be witnessed by the success of handbooks such as Anne Kenney and Oya Rieger's work, 'Moving Theory into Practice: Digital Imaging for Libraries and Archives' [4]. This holds true for technological solutions proposed in the area of digital preservation as well, such as migration, conversion, emulation and encapsulation. The need to understand technical options, the advantage of implementing generic solutions and interoperable information systems and the prohibitive costs of implementing and managing IT- infrastructures are major drivers for collaborative efforts across all sectors and disciplines.

The need to investigate technical options and to integrate these with digital collection management practice has in turn triggered attempts to develop cross-disciplinary consensus on issues relating to digital collection management. Such issues span a wide range of perspectives such as appraisal strategies and selection criteria, document authenticity and intellectual integrity, collection description and finding aids, access rights and authorisation, etc. They focus on technological, legal, economic and organisational aspects. They are of concern to governmental, academic, commercial and cultural heritage sectors. They involve a wide array of stakeholders including libraries, archives, museums, historians, scholarly researchers, artists, authors, musicians, entertainment, news, audio/visual media producers, publishers, etc. As a result cross-sector, cross-disciplinary national consensus frameworks are being developed in order to enable institutions to implement digital technologies that apply as broadly as possible across all relevant disciplines and perspectives.

A good example is the 'little blue book', published by the UK National Preservation Office, entitled: Digital Culture: maximising the nation's investment [5]. Looking at the larger picture of managing digital preservation as a whole the booklet proposes to address the challenge by setting up organisational structures. It provides guidance and recommendations to identify the different stakeholders and their responsibilities, to determine preservation costs and funding strategies, to design an integrated policy framework that enables concerted action and a national implementation strategy.

Another example is the Preservation Management of Digital Materials Workbook, compiled by Maggie Jones and Neil Beagrie [6]. It gives a generalised overview of issues relating to digital collection management – targeted to a very broad audience from all sectors and disciplines. The issues associated with digital preservation are grouped in three broad categories: technological, organisational and legal.

Specific requirements for specific application domains

It remains the case that the specificity of historically grown disciplines and sector-bound practices defies attempts to formulate broad consensus frameworks and clear statements on issues surrounding digital collection management. For memory organisations with similar functions, drawing on each other's theory and practice may be rewarding, as advocated by Helen Tibbo in her short article, published in the digital libraries issue of *Communications of the ACM*: 'Archival perspectives on the emerging digital library' [7]. Yet, the difference in theoretical background and the divergence in practice underscore the distinctive features of similar organisations. To take Tibbo's example in the field of digital archiving, both archives and libraries have a task to select, describe, store and preserve digital material. These two different types of organisations have quite different perspectives on the kind of information that concern them, the forms in which this information is created and embodied, the uses to which this information is put, and consequently what is required to preserve it. Whereas archives collect unpublished and unique materials of institutional or personal origin with a focus on the records that document the creation, lifecycle and organisational context of the materials, libraries mainly collect published material, with a focus on wide scale availability of publications. These differences in perspective across the archival and library communities manifest themselves, for example, in different ideas of what it means for preserved information to be 'authentic' and 'valid' for its intended use. Within the archival world significant work, under the umbrella of the INTERPARES project [8], is underway to map evidentiary requirements and principles to electronic records and to explore digital technologies that can support such requirements. In the library world no such discussion has taken place yet, because traditionally authenticity of published material is no real issue. However, the relative ease with which it is possible to alter informational content in a digital environment, compared to the paper environment, has introduced the concepts of authenticity and intellectual integrity in the library community, as well.

Preservation starting points for deposit libraries

Deposit libraries with a task to guarantee last-resort access to all published material produced in their own country, are challenged by possibilities and restrictions of preservation technologies. Computer scientists and IT-consultants encourage them to make their preservation requirements and authenticity principles more explicit. In his interactions with deposit librarians, Jeff Rothenberg has distilled the following authenticity principles [9], which still need verification within the community:

- a preserved publication should be as much like its original published form as possible,
- a preserved publication should retain content, behaviour, functionality and look-and-feel of the original publication as much as possible,

- applying preservation techniques should not lead to re-publishing or reformatting of the original publication (no creation process).

The long-term preservation study being carried out by IBM-Netherlands as part of the implementation of the deposit system for electronic publications at the Koninklijke Bibliotheek [3], looks at authenticity principles as well.

A starting point that is felt to be important in this respect is that the preserved deposit copy needs only to be viewed in a document reader environment as opposed to edited and re-used in a document processing environment. This principle conforms to the way in which publishers make their publications available. If publishers do provide processing functionality, in exceptional cases, the deposit library should also try to make future re-use of data possible. This general starting point reduces the required functionality of a preserved publication significantly, thereby simplifying technical preservation solutions. It should be noted that this authenticity principle is very specific for the deposit regime of electronic publications, and does not necessarily apply to other memory institutions such as data archives – where reusability of content is a top one requirement. It shows how authenticity principles can vary according to the institutional mission and how dangerous it is to try and generalise preservation and authenticity requirements into broadly applicable consensus frameworks across all sectors and disciplines.

Another starting point, related to the previous one, is that the publisher's dissemination environment, with features such as graphic design, branding and advanced searching capabilities, is not considered to form intrinsic part of the deposited publication. In other words the deposit copy is considered as an autonomous published entity that should be definable and identifiable outside its dissemination context as well. This allows deposited publications to be archived in a separate deposit environment with its own searching functionality supporting deposit collection uses. The usefulness of the distinction between content and functionality, enabling the separate development of value-added functionality for dissemination purposes and for archival purposes, was also highlighted by the Yale University Library and Elsevier Science, in their proposed approach for a collaborative project funded by the Andrew W. Mellon Foundation [10]. This project is a showcase for close publisher-library interaction with a view to foster long-term access to digital publications. It also shows that institutional roles and responsibilities in the digital environment are not converging or merging. Publishers and libraries are reassessing their complementary capabilities and expertise and in doing so they reaffirm their traditional institutional mission. They do not interfere with but respect each other's institutional imperatives. This approach differs somewhat from ideas developed in the archival community where it is expected that digital preservation requirements need to be integrated with the creation process, in other words preservation decisions have impact on the way in which the material is created, described and stored. This may be a very legitimate standpoint from an archival perspective, but for deposit libraries it is unthinkable to impose their requirements on the business process of publishers.

Distinguishing between dissemination and preservation environments raises the issue what to do about web archiving. Web archiving has grown to mean harvesting and preserving web pages from the Internet, with the aim to safeguard the web and its history for future generations. In how far can and should web pages be preserved in their dissemination context? How can we define and delimit web publications for preservation purposes? Are web sites to be considered as publications in their own right or are they publisher dissemination environments? Should a web archive support hyperlinks across web publications, should it provide web search engine functionality? Should it reflect the functionality of the web as it develops over time?

A starting point taken by the Koninklijke Bibliotheek on these issues, again as part of the long-term preservation study carried out by IBM-Netherlands, is that web archiving has a different aim than the deposit of electronic publications. While web publications can and should be part of the deposit collection, web archiving is out of scope because it also aims to keep the dissemination environment of publishers. The deposit library needs to agree with web publishers what publications fall under the deposit regime and procedures for deposit. This standpoint is strengthened by several observations [11]:

- 80% of the web content originates from 20% of the total number of web sites
- increasingly, web content cannot be harvested because of dynamic database publishing practices
- institutions that do web archiving have been confronted with legal complications such as copyright and access restrictions imposed by web publishers

These observations suggest strongly that it is maybe feasible and wiser to come to deposit agreements with web publishers directly - as has been the practice with conventional publishers.

Preserving born-digital and digitised material

In the little blue book from the NPO mentioned earlier, it is stressed that massive investments are made in digitisation and 'that a great deal of money can be wasted if digitisation projects are undertaken without due regard to the long-term preservation of the digital files.' This standpoint can be found in much of the literature on digitisation. As I have stated earlier in the book review column of *Alexandria [12]*, this standpoint does not reflect the complexities involved in the decision making process concerning investments in digitisation versus investments in long-term preservation. Both are costly but mostly incompatible undertakings. By tackling both the preservation of digitised heritage and the preservation of born-digital heritage in one, this approach confuses the different issues at stake. This was also one of the conclusions of the NEDLIB project, which from the start scoped its attention to the long-term preservation of born-digital material only.

Firstly digitisation usually concerns material that exists in paper-based form or whatever other physical form, whereby the digital version is not the object of preservation but just a surrogate that is more appropriate for dissemination. Only in the few instances where the original object is in a far state of deterioration, is digitisation used as a means of preservation.

Secondly the issue of authenticity is not as relevant for digital versions of physical originals as it is for born-digital originals. Whilst one would want to safeguard the born-digital original, one would prefer to upgrade the quality of the scanned versions of physical objects if it were possible. Digitisation is primarily a means of making physical and site-bound material more accessible to a wider audience. In view of the fast changing imaging technology with ever-increasing resolution quality, compression capabilities and automatic pattern recognition possibilities, the aim to preserve digitised files for the long-term seems futile. Digitisation is a welcome means of exploiting the physical collections of memory organisations. It seems however not very appropriate to create large high-resolution archive files consuming terabytes of storage capacity where more economic sized presentation files will do the job. A more thorough argumentation and discussion of these very important issues are missing in current studies on digitisation.

Transition from specific user requirements to generic IT-solutions

In the previous passages I have stressed the importance to specify preservation requirements according to institutional aims and activity areas. It is important to understand that, by contrast, IT-based technologies are not domain specific and that, when applying such technologies it is best to take a generic approach and to look for globally applicable solutions. I will illustrate this with work done in the context of OAIS and NEDLIB. In this context a model for the implementation of an archive system has been developed, based on the general requirement to realise a storage- and access system for digital collections. The data model and functions are designed with a view to accommodate all types of material and to perform all basic processes required by an archive. In this sense the model can be implemented for any application area, be it spatial data archives, deposit libraries or national archives. The OAIS concept of self-contained information packages (SIP, AIP, DIP) are at the base of the information model and are handled in a generic way by the system, regardless of the content they carry. The functions Ingest, Data-Management, Archival Storage, Administration, Preservation and Access are basic and cater to a fully functional system, regardless of its operational environment. It will be at the system configuration level that specific implementations cater to specific user requirements. At this level the deposit library or the spatial agency will need to define what the content information unit of the archival information package will be, for

example, a deposit copy of an electronic publication or a data set from a space observation mission. During the European tendering procedure soliciting IT-vendors to make an offer for an OAIS-based deposit system, the Koninklijke Bibliotheek concluded that the realisation of such a system by use of existing technologies and standards was feasible, but that the preservation functionality request was too premature. There are no ready to implement technologies and solutions for digital preservation.
Johan Steenbakkers envisions a ubiquitous preservation function which he has depicted, during the NEDLIB workshop in December 2000, as follows [13]: "Suppose you need a digital article or picture produced in the year 2002. The article or picture was stored as a bitstream at the time, using a special software utility called *omega.* You click on the omega file containing the necessary preservation metadata. Via the omega metadata the correct decoding mechanism for the original bitstream is provided by the archive as a plug-in and there you are: within seconds you can read the original article or view the original picture."
He has challenged the IT-industry to work towards such a standard ubiquitous solution. It seems the IT-industry is willing to take up the challenge, but it will require a lot more fine tuning of preservation requirements and refined discussion, research and experimentation with preservation techniques. Because behind the vision of the magic omega button lies a whole world of rights and access management, of technical, administrative and preservation metadata exchanges, of migration and recreation techniques - which we have only begun to sketch in bold lines. Welcome to the fascinating world of digital preservation!

Notes

1. NEDLIB was a very successful European deposit library collaborative effort. See NEDLIB project home page: www.kb.nl/nedlib/
2. The Koninklijke Bibliotheek is implementing a deposit system for electronic publications, together with IBM, Netherlands. See DNEP project home page: www.kb.nl/dnep-project/
3. Reference Model for an Open Archival Information System (OAIS). See http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
4. Anne Kenney and Oya Rieger, 'Moving Theory into Practice: Digital Imaging for Libraries and Archives'. RLG (California, 2000).
5. Ed. Mary Feeney, 'Digital Culture: maximising the nation's investment. A synthesis of JISC/NPO studies on the preservation of electronic materials'. NPO, British Library Board, 1999.
6. Maggie Jones and Neil Beagrie, 'Preservation Management of Digital Materials Workbook'. Pre-publication draft, October 2000. http://www.jisc.ac.uk/dner/preservation/workbook/.
7. Helen Tibbo , 'Archival perspectives on the emerging digital library' in: Communications of the ACM. May 20001.
8. International project on Permanent Authentic Records in Electronic Systems (INTERPARES); www.interpares.org.
9. Taken from a presentation given by Jeff Rothenberg during a Digital Preservation conference organised by the Helsinki University Library on April 23 rd, 2001.
10. See Proposal for a digital preservation collaboration between the Yale University Library and Elsevier Science. Version 4. Date: 30 September 2000.
11. These observations are based on an analysis of NEDLIB Harvester testing results from different national library sites.
12. See Alexandria, March 2001.
13. Johan Steenbakkers, 'NEDLIB Guidelines: setting up a deposit system for electronic publications', presentation held at the NEDLIB workshop. December 2000. www.kb.nl/nedlib/workshop/