Werk

Titel: Preservation of Digital Materials for Libraries

Autor: Mackenzie Owen, John

Ort: Graz

Jahr: 1996

PURL: https://resolver.sub.uni-goettingen.de/purl?514854804_0006|log69

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

European Research Libraries Cooperation:
The LIBER Quarterly, 6 (1996), 435-451.

Preservation of Digital Materials for Libraries

JOHN MACKENZIE OWEN
Senior consultant, NBBI

Introduction

How can we guarantee that information published in digital form will be available to future generations? Many people in the library and archival world are beginning to understand the significance of this question. It is highly likely that within a number of years most information will be published in digital form over networks. It is highly unlikely that the originators of these publications will maintain them on their network servers for centuries or even decades. It is time to think about preserving the digital intellectual record.

Two recent studies have identified the problems of digital preservation and point towards solutions. One such study has been carried out by a Task Force on archiving of Digital Information commissioned by the Commission on Preservation and Access and

the Research Libraries Group in the United States[1]. The other has become known as the ELDEP-study[2], carried out by NBBI in the Netherlands for the European Commission at the suggestion of CoBRA, a group established under the aegis of the Conference of European National Librarians. This paper draws on the findings of these two studies in order to clarify a number of issues which relate tot the future role of libraries in ensuring the continuity of access to digital materials.

Libraries and archiving

It is rapidly becoming common practice to refer to libraries as 'digital archives'. For traditional librarians, this sounds rather strange. They have become used to regarding the library as something entirely different from an archive. From this traditional viewpoint, archiving is concerned with 'unique' documents, i.e. documents of which one or only a small number of copies exist. Archiving is primarily the responsibility of the originator of a document, and its main function is to store and preserve documents after their primary use. Libraries, however, acquire and store publications (of which many copies exist) in anticipation of use, and they have their own responsibility for preservation beyond that of the originator.

It is interesting to see how the concepts of libraries and archives come together in the light of networked publishing. Digital networked resources are unique documents in the sense

1 Preserving digital information: report of the Task Force on archiving of Digital Information commissioned by the CPA and the RLG: final report and recommendations. - May 1, 1996. This report is available at [*http://www-rlg.stanford.edu/ArchTF/*], [*http://www.rlg.org/ArchTF/*] and [*http://ukoln.bath.ac.uk/mirror/archtf/archtf.html*].

2 Mackenzie Owen, J.S. & Walle, J. v.d. - A study of issues faced by national libraries in the field of deposit collections of electronic publications: final report. - Luxembourg: European Commission, 1996.

that they are not distributed in multiple physical copies. In fact, there is often only a single source from where they can be obtained. As we shall see, libraries may only be allowed to acquire and store such documents after a certain period of time, i.e. after their primary use as sources of income for their originators. Whereas in the archival world storage of documents in multiple locations is usually impossible, in the world of networked libraries it is unnecessary. In summary, storage and preservation of resources published over the networks is becoming an archival task.

There is currently some debate as to whether there is any need at all for digital archives. Would it not be better to leave the responsibility for archiving with the originator (e.g. the publisher)? There are several reasons why this is not the case, and why libraries (or other archival bodies) should take on this task:

- Originators have a short-term (economic or other) interest in storing documents for access over the network. When that interest ceases to exist, they will remove them from the network. There is a need for archives which have a specific responsibility for long-term availability and continuity of access, and which have the funding to do so.
- Long-term archiving is correlated with extremely infrequent use (or 'access' in network terms). This is best served by technical and organisational solutions which differ from those required for frequent use. Dedicated archives can provide the most efficient modes of storage and access for this task.
- Archiving is more than storage, it also implies active preservation. As we shall see, preservation of digital materials is a difficult and expensive issue which can only be carried out by specialised centres.

An important mechanism for performing the archiving task, at least in Europe, is the legal deposit mechanism. This implies that digital publications are acquired by a national digital archive (the

deposit library) responsible for preservation and continuity of access. It is our view that the digital deposit library could serve as the archival backbone for future electronic libraries. If national libraries do their job well, other libraries can focus on providing access to their users with little or no concern for creating local collections and for storing and preserving digital publications.

Digital publications and preservation

What kind of digital publications should concern us in the context of long-term preservation? One has to understand that preservation really is a long-term affair, concerned with maintaining the availability of publications over a period of centuries. In this context it does not make sense to pay much attention to short-term issues. It is clear that future digital publications will be dynamic, networked, multimedia digital objects. This means that they will be distributed over networks, will be subject to frequent changes during their economic life-time, and will contain text, images, sound and often a high degree of interactivity. In our view off-line media such as CD-ROM and CD-I are of limited future interest. What can be done on such media can and will be done much better in networked form. Many off-line digital publishers are already considering a move to on-line networks. The key source of publications for the digital archive will therefore be the network, in whatever form it may develop over the years.

In this context, preservation acquires a distinct meaning. It is concerned with:

- maintaining access to the intellectual content and functionality of digital objects, and
- preserving the content structure of the network.

Preserving functionality, in addition to intellectual content, is important. Digital objects as published on the network are increasingly capable of performing integrated functional tasks, e.g.

performing calculations, offering digital simulations, allowing browsing and search tasks, etc. Preserving functionality is a major challenge to preservation, since it is always based on current information technology which may not be available at a later point in time.

Preserving the content *structure* of the network is important because networked objects are becoming more and more interlinked, forming a close web of information where the meaning of an individual object is partly defined by the context provided by a large number of other documents. It is not sufficient to preserve individual objects. One has to preserve them within this structural context.

Issues for digital archiving

As already indicated, the major concern for digital archiving is preservation. We cannot take for granted that that we can guarantee availability of digital objects for future generations just by storing them in a safe place. Before discussing specific aspects of preservation, it is important to point out that the issue of preservation requires an integrative approach, taking into account the entire process from selection to long-term storage. Preservation influences all aspects of digital archiving: decisions taken at earlier stages (selection, acquisition, pre-processing, cataloguing etc.) have implications for the cost-effectiveness of preservation.

This is not a trivial statement. Digital archiving involves many issues which we as yet do not fully understand. The ELDEP-study, for instance, has identified the following issues for deposit libraries:

- **Acquisition**
  - acquisition methods
  - supplier relationships
- **Pre-processing**

  - installation, verification, documentation
  - cataloguing & indexing
    - original cataloguing
    - shared cataloguing
    - embedded metadata
  - standardisation
    - format
    - medium
- **Storage**
  - off-line objects: original form *versus* mass storage; accompanying materials
  - networked objects: level of linking
  - dynamic objects: refreshing / snapshots
- **Access**
  - legal aspects
  - service level
  - continuity of network access

The ELDEP-study discusses these issues in far more detail than can be done here, and points to possible solutions. The point to be made here is that long-term preservation of digital resources is extremely complicated and far beyond the capabilities of most so-called digital libraries.

Digital preservation

Let us now look at digital preservation as such. If we consider a digital archive that has obtained a digital publication, what makes preservation more than storage? The answer lies in the fact that it is not sufficient to store a digital object in its original form as received, if we wish to guarantee continuity of access. It is therefore necessary to convert it, and to continue doing so for as long as it remains in the archive. There are three main reasons why conversion is necessary:

- *Physical deterioration of media.* Digital objects are stored on digital media. Any digital medium currently known to

us has a limited life-span, after which it cannot be used to store the data safely.

- *Obsolescence of the technical environment.* Digital objects, on whatever medium they are stored, depend on specific hardware and software for their use. This combination of hardware (readers, computers) and software (operating system, database software, browsers, viewers etc.) is often referred to as the 'technical environment' under which an object can be accessed. When the technical environment changes chances increase that the object can no longer be used. The current rapid development of information technology means that the life-span of the technical environment is often less than ten years.
- *Economic and management considerations.* It is usually not feasible for a digital archive to handle all types of media, formats etc. on which digital objects are distributed. In view of cost considerations, available skills and the technical resources of the archive, some objects may have to be converted to standardised formats.

The need for conversion leads to the requirement for the digital archive to develop *migration strategies*. This is not an easy task, because we know very little about future technical developments and about the cost involved in migrating large volumes of information. In general one can choose from the following set of strategies:

- *Medium refreshing:* transferring the data to a new medium of the same type.
- *Medium conversion:* transferring the data to another medium with better preservation characteristics.
- *Format conversion / re-standardisation:* transfering the data to another format which is easier for the archive to handle.
- *Migration of the technical environment:* transfering the data to be accessed under new hardware and software.

- *Emulation of the technical environment:* re-creating the original technical environment in a new technical environment, making it unnecessary to migrate large numbers of digital objects.

The uncertain world of digital archiving

The number of uncertainties involved in planning for digital archiving is overwhelming, They include:

- *Cost factors:* what are the important cost factors and how should they be calculated? There is a need for cost models and for the creation of cost data, based on actual experience in order to evaluate these models. Very little work has as yet been done in this area.
- *Cost development:* very little is known about the future cost of obtaining items for archiving and the cost strategies of publishers (e.g. costing per article instead of per journal title). Also, and more importantly in view of migration, it remains unclear whether the current cost of technology (e.g. the cost of storage) will continue to decline at the current rate.
- *Volume of networked publications:* although it is certain that networked publishing will eventually become the normal mode of distribution, it is not clear at what rate the shift from print and off-line publishing to networked publishing will take place.
- *IT-change rate:* will the rate of technological change (leading to the need for migration) continue at the current levels?
- *Emulation technology:* will emulation technology become available and help to avoid the high cost of technological migration?
- *Legal issues:* copyright legislation is evolving rapidly and could have unforeseen consequences for preservation.

This affects both the right to archive and provide services, and the right to convert publications for preservation reasons.

- *The future of the commercial publishing industry:* will the commercial publishing industry survive the move towards digital networks as a publishing channel? Will alternatives (e.g. institutional publishing) allow for effective archiving?
- *Strategic co-operation between libraries and publishers:* will libraries be able to establish satisfactory relationships with publishers to perform their function as digital archives?
- *Inter-library co-operation:* will the library world succeed in setting up a cost-effective way to organise digital archiving? Will the national (deposit) libraries develop a role as an archival backbone?
- *Preservability of dynamic networked digital objects:* finally, there is growing uncertainty as to the intrinsic preservability of dynamic networked digital objects. Perhaps the nature of these objects will make it impossible to preserve them effectively over time.

There are no easy answers to these questions. Digital archiving is a question of trial and error. However, clear thinking and the ability to discard traditional ways of organising information services are required in order to achieve the goal of long-term preservation of the intellectual record.

The cost of digital archiving

One of the most difficult issues for digital archives is the assessment of the future cost of preservation. It has already been mentioned that there is a need to develop -and test cost models. What is clear however, is that the cost of digital archiving and preservation will be high. Too high, in fact, to be carried by

individual libraries. Digital archiving will therefore have to be organised at a higher level, either on a national basis (e.g. by national deposit libraries), or on an international scale through European or global domain-based digital archives.

A large number of cost factors have to be taken into account in order to estimate the full cost of long-term preservation. These include:

- Organisational costs e.g. of co-operative archiving schemes
- The cost of selection, if we concede that it is impossible to archive all materials distributed over the network
- Acquisition costs, i.e. purchase or license costs for digital materials
- Cataloguing and indexing, taking into account the handling of embedded metadata in digital objects
- The cost of conversion to standard media and formats
- Annual storage cost (incl. accompanying materials for off-line materials)
- The cost of maintaining an access infrastructure
- Maintenance cost for preserving continuity of access (including the cost of media preservation and of migration to new technological environments)

Our current understanding of the cost issue can be summarised in just few points:

- Under current conditions, the cost *per access* of digital archiving is roughly the same as that of paper archiving. Electronic publishing may be a cost-effective solution for information distribution compared with print publishing. It certainly is not a cost-effective form of long-term archiving.
- Migration costs could lead to substantially higher per-item costs for digital preservation. The cost of migrating large volumes of digital objects to new technological environments could well prove to be prohibitive.

- Investment in standardisation can lead to future cost savings, because it can lead to significantly lower migration costs.
- The cost of digital preservation compared to print storage depends mainly on the organisational model for the electronic library. Duplicating the long-term storage and preservation of digital objects in many libraries is unnecessary, and therefore unnecessary expensive.
- The most cost-effective approach tot digital preservation is based on a co-operative model for the digital library in which digital archiving is carried out by a limited number of large centres to which other libraries provide access.

The latter point can be illustrated as follows. Under the present system, based on print publishing, many libraries acquire and store the same publications. In the networked world, any object stored in a single location can be accessed over the network; there is no need for multiple storage on a large scale. Libraries could therefore agree to co-ordinate storage, each library archiving a small, mutually exclusive sub-set of published information. Library users would then obtain materials either from the local archive, or from one of many other archives. However, this is not a cost-effective solution. The overhead cost for management and co-ordination, and for maintaining a meta-directory would be extremely high, and it is unlikely that duplication could be avoided to a sufficient extent. Moreover, it would require that a large number of libraries maintain the specialised skills and technical means for long-term preservation. By far the most cost-effective solution would be to set up specialised archival services (e.g. deposit libraries or domain-based archival centres) which maintain the archival collection for the library world.

Co-operation between libraries and publishers

An issue which needs to be discussed here is the relationship between libraries and publishers. In the traditional world of printed publications, the library acquires materials at a certain cost from the publisher, usually as soon as they are published. Publications are held 'in print' by the publisher as long as they generate sufficient revenue, after which publication is discontinued and the title goes 'out of print'. The library stores the publication, often indefinitely, in order to maintain 'continuity of access', i.e. to keep it available for future use, long after it has gone out of print. (cf. figure 1). This 'archival' role is one of the main functions of the traditional library. This is possible because long-term archiving of printed materials is relatively cheap. It is also necessary to a certain extent, since access is more difficult if the item is not available locally. The need for local archiving has, however, diminished over the past decades due to computerised union catalogues, interlibrary lending schemes and document delivery services. Nevertheless, most libraries are still 'collection-based', i.e. archival in nature.

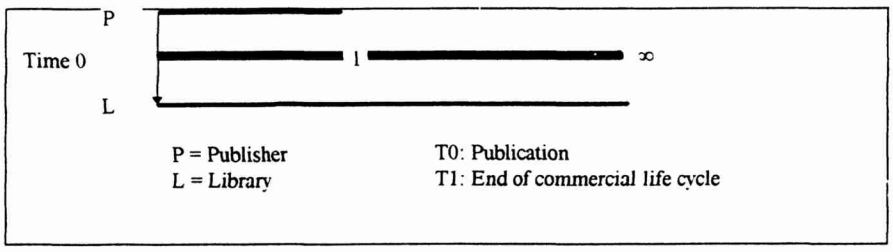

Figure 1

It is now becoming clear that publishers will not use this model for networked distribution of publications in digital form.

In spite of a number of experiments where publishers deliver electronic versions of journals in digital form to libraries, where they are made available to users and stored on digital media - a perfect digital parallel to print distribution - this will not be the way things are done in future. Publishers will store their publications on their own 'digital repositories' rather than distribute them to libraries for local storage and access. License agreements between publishers and libraries will allow users to access the publisher's repositories, either directly or through library-based systems. This brings many advantages to publishers, besides being a far more efficient solution in a world where network access makes large-scale multiple storage a thing of the past.

From a preservation point of view however, this model creates a problem. Publishers can only be expected to store information in their repositories (at least in a way which allows user access) during its 'economic life cycle', i.e. for as long as it is profitable for the publisher to do so. After that, the publication goes 'out of print' , i.e. it becomes inaccessible and may not be stored and preserved for future use. At this point in time, the need for digital archiving arises. What is necessary, therefore, is that publishers and libraries agree that items will then be transferred to a digital archive maintained by the library world in order to guarantee continuity of access for future generations (cf figure 2). Such an agreement would preserve the archival role of the library world, although not necessarily (as argued above) of most individual libraries.

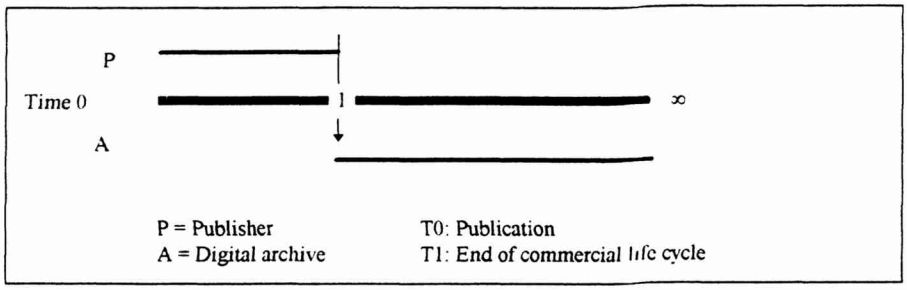

Figure 2

It should be noted that this solution already exists, although in a slightly different form. In most European countries national deposit libraries obtain a copy of digital publications from the publisher under deposit legislation, and national libraries are already setting up mechanisms for handling, storing and preserving the national digital output. Digital archives are therefore already under construction on a national basis. Access to this archive is, however, limited in order to safeguard the publisher's copyright. This means that access is only available on-site under restricted conditions, or at the most over the network to registered users of the national library.

What is needed, therefore is agreement between national libraries and publishers that unrestricted access will be allowed after the economic life cycle of the publication. This could be done in a generic way (e.g. after a certain number of years), or on a per-item basis. Using the deposit library as the future digital archive for published materials would make good use of the existing infrastructure, facilitate negotiations between publishers

and the library world, and offer guarantees for more or less complete coverage of the intellectual record.Although domain-based approaches to digital archiving have certain advantages as well, the ongoing development of digital deposit libraries at the national level in practice could well turn out to be the best solution, especially in the European context. The major challenge for national deposit libraries is to gain acceptance of this important role in the library world at large.

Conclusions for the digital library

We can summarise the issues described in this paper as follows:

- Preservation requires an integrative view of the entire library process.
- Preservation of electronic publications equals preservation of dynamic networked multimedia.
- The large number of uncertainties makes long-term planning of digital preservation extremely difficult.
- The cost of preservation depends primarily on:
  - development of technology (storage cost, need for migration)
  - co-operation between libraries and publishers
  - inter-library co-operation
- Cost reductions can be found through:
  - standardisation
  - emulation as an alternative to migration
  - co-operative organisational models
- Long-term preservation is a specialised task, to be performed by a limited number of digital archives

The conclusions to be drawn from our analysis of the preservation issue and - in a wider context - digital archiving for the future of libraries are easy to understand, but may not be welcomed by librarians used to more traditional ways of thinking.

The first conclusion is that in the long run, the main source of library materials will be the network. To be more precise: users will acquire their information in digital form from networked sources, and libraries will be no more (and no less) than intermediaries, helping the user to identify, locate and access information.

In the relationship with publishers, libraries will manage the access by users to copyright materials stored by the publisher based on license agreements. Libraries will no longer acquire and store current materials from publishers.

Long-term storage and preservation - digital archiving - will no longer be a task for most libraries[3]. This function will be handled by specialised digital archives, either national deposit libraries or domain-based archival centres on an international scale.

The library world is facing major challenges, quite similar in scale to those faced by other organisations in the information chain such as publishers.[4] Digitisation and networking are revolutionising the publishing industry, and will do the same to the library world. One of these challenges is to develop efficient modes for long-term archiving in order to preserve the digital intellectual record. This will involve major changes in the concept of a library. If the library world wishes to serve the future needs of information users, it will have to accept these changes and move towards more centralised archiving. Long-term storage of digital

3 Note that this refers to *long-term* storage. Whether libraries have reasons to store digital publications for current, short-term use is a different matter. The cost-effectiveness of this depends on the expected volume of use. For long-term storage, frequency of use is expected to be sufficiently low to justify networked access to centralised digital archives.

4 An overview of developments towards new models for libraries can be found in: Mackenzie Owen, J.S. and Wiercx, A. - Knowledge models for networked library services. - Luxembourg: European Commission, 1996.

information will not be a normal function of the library of the future.