Libri, 2004, vol. 54, pp. 98–103
Printed in Germany · All rights reserved

Libri
ISSN 0024-2667

Universal Access Through Time: Archiving Strategies for Digital Publications

WIM VAN DRIMMELEN
Koninklijke Bibliotheek, Den Haag, Nederland

The Royal Library in The Hague has been a leader in developing and exploring approaches to long-term preservation of digital objects. *Libri* invited its distinguished Board Member, Wim van Drimmlen, to share some of his views on this important issue with our readers. What follows is based on a presentation made in May 2003 at the STM Conference in Amsterdam. We hope its appearance will encourage more papers on this important topic.

It was a pleasure to be invited as speaker for this conference. When I learned about the theme of your conference, I was just simply delighted. Universal access, that's a librarian's dream! On second thoughts, however, I wondered whether I had missed something. Had publishers suddenly turned into philanthropists? How could I have missed that on the News? This brought me back to reality again. I have recovered now. It's a pleasure to participate in a realistic discussion on universal access.

As you all are aware, librarians are fond of order and precision. They love cataloguing and classifying. I won't let you down. I'm happy to comply with the image of my profession. My first point is on terminology.

There is a series of terms that may easily lead to confusion. I am referring to terms like: archive, permanent archive, dark archive, open archive, depository, electronic deposit, repository, institutional repository, and there may be more like these. It's not my ambition to make an end to this confusion. I know my proper place. However, the least I can do is to try to clarify what I mean with the terminology I am using. Therefore I will start with a characterisation of the concept *'permanent archive'*, or in short *'archive'* or *'electronic deposit'*.

(Permanent) archive = e-deposit

The primary goal of a permanent archive is long term preservation, safeguarding future availability and accessibility. A permanent archive also protects the authenticity and integrity of the publications it has taken charge of. That's a logical component of its primary goal. In order to reach this goal, the institution that manages the archive must commit itself to the permanent development of an ever-changing preservation toolbox.

Open access is not the primary goal of an archive. So restrictions on access are accepted, if publishers demand so. The archiving agreement between Elsevier and Koninklijke Bibliotheek provides an example. It restricts access to on-site only and allows for a restricted interlibrary document supply, on a print basis and as last resort only. Of course, publisher and archive can agree otherwise.

An archive does not necessarily provide the same functionality as the publisher's own site. It provides a basic search and retrieval system. Ad-

Wim van Drimmelen is Head of Koninklijke Bibliotheek, Postbus 90407, 2509 LK The Hague, Netherlands. Tel: 070-3140911, Fax: 070-3140450, Telex: 34402 KB NL, E-mail: Wim.vanDrimmelen@kb.nl

Based on a paper presented at the STM-Conference, Amsterdam, May 2003.

ditional functionalities would only increase the burden of the preservation efforts. Last, *but not least*, publishers are to deposit their electronic publications free of charge.

Yet, there is another pair of terms that easily causes misunderstanding: OAI and OAIS, the Open Archives Initiative and The Open Archives Information System. They are look-alikes, yet they cannot stand in for each other because they address different challenges. OAIS is a so-called "archival reference model", a kind of standard, meant to promote interoperability. In October last year our new e-deposit system, developed jointly with IBM, became operational. The system is fully OAIS compliant, but access to its content can either be restricted or unrestricted, depending on whatever is agreed with the publisher. In the Open Archives Initiative, however, unrestricted access is basic. It stems from the fundamental belief that the intellectual output of mankind should be accessible for all.

I hope this introduction will prove to be helpful. Let me now address the main theme of my paper. When exploring models for digital archiving, it might be helpful to look back into the world of printed paper.

Archiving in the paper world

Archiving in the paper world is characterised by a dual model. On the one hand there are the official deposit libraries, usually national libraries. Preserving our heritage is a prominent part of our mission. Deposit collections are comprehensive and systematic collections that are built on a national and geographical basis. It's a worldwide system, providing a clear assignment of responsibilities. It is based on global arrangements supported by IFLA and UNESCO.

On the other hand, there are numerous collections in individual libraries all over the world. This simply results from the fact that to give access to a printed publication, you need a physical copy. These collections however, are built on the basis of specific profiles to serve specific audiences. They are not comprehensive and systematic in the same sense as deposit collections are. Moreover, preserving our heritage is not the primary aim of these institutions. Yet, there is some guarantee in their sheer numbers. They will not all burn down on the same day. (If they do so, we will probably have lots of other problems to worry about.)

Publishers have a passive role in this model. Certainly, they provide copies for the deposit libraries. We are very grateful for that, but they take no active part in the archiving process as such. In the paper world, there was apparently no need for this.

Both components in the dual model have inspired current initiatives to tackle the archiving issue, as I will mention later on. As to the role of the publishers, I would argue that in the digital world a more active role from their part should be required.

What makes the digital world different?

Let's move now to the digital world. What makes the digital world different? To identify the differences we must look into the basic characteristics of digital objects. Each of these characteristics has direct consequences for the archiving issue.

Digital objects are omnipresent. If there is a single copy anywhere, it can be accessed from everywhere. A single copy is sufficient to serve all readers worldwide. Omnipresence apparently is a mixed blessing. On the one hand, it makes life easier, both for publishers and libraries. On the other hand, it makes them feel uneasy about potential consequences. Digital distribution of scholarly information is more efficient and economical, but it makes publishers anxious about unauthorised use. For libraries there is no need to store their own copy of a digital publication if it is accessible online elsewhere. But at the same time, it makes them worry about future access. Therefore they demand guarantees for so called 'perpetual access'. Omnipresence paradoxically leads to anxiety about future presence.

Digital objects are volatile. They can be changed and adapted easily. That's a strong point in the case of online databases meant to provide up-to-date information. It is a weakness however in the case of scholarly publications, where adaptation comes down to manipulation. Volatility has added new dimensions to the issues of authenticity and integrity. In the digital world, safeguarding integrity is a prominent component of the archivng challenge.

Digital objects are extremely perishable. If there are floppy disks still lingering in your office, you

should hope they don't contain essential information. The floppy may not fit into the machine on your desk. So you better take your loss now. This is the preservation problem in a nutshell. In this respect, the good old clay tablets are by far superior to our modern media.

The extremely short lifespan of digital objects creates the necessity to develop preservation techniques right from the start. Printed information can be accessed directly and easily, provided you have obtained a physical copy, and provided that either your eyes are good enough or you have the right pair of glasses. However, to access digital information you always need an intermediary instrument. You always need a computer, consisting of hardware and software that change rapidly over time. From the viewpoint of permanent accessibility this represents the major difference between the two worlds. In the next section I will go into this a bit deeper.

Digital objects are also very fertile. In Dutch we would say, they breed like rabbits. The volume of digital publications is growing very rapidly. This, however, does not necessarily pose a major problem. Experience at the Koninklijke Bibliotheek indicates that there are considerable economies of scale. It would mean that we should exploit them to counter rapid growth of the volume. Yet, there is a related phenomenon that does give rise to more of a challenge: the diversity of formats and software programmes. The greater this diversity, the more complicated your preservation toolbox will have to be.

I want to illustrate two fundamental facts. Firstly, preservation never can be a passive process; it implies repeated actions all through. Secondly, you have to adapt your toolbox continuously as technology develops.

Components of digital publications

All digital objects consist of a bit stream stored on a disk, a tape or whatever. The storage medium deteriorates through time, but you can easily copy the bit stream to a fresh medium.

In the bit stream is a logical format, e.g. Word Perfect. These formats rapidly become obsolete. Word Perfect by now seems to have been replaced by MS Word.

Finally you need an interpreter or rendering tool in order to access and decipher the formatted bit stream, e.g. your PC. Again, these rendering tools also rapidly become obsolete.

Deterioration and obsolescence are the problem. What can be done about this?

Preservation measures

The key concepts are: refreshing, migration and emulation. Beware, however, because here again there is confusion on terminology. What I mean by refreshing sometimes is called migration. To make things worse, what I mean by migration, sometimes is known as conversion. So I will try to clearly define the terminology.

Refreshing means transferring the bits and bytes to a fresh physical storage medium. This is the least part of the problem. It can be done, but it has to be done repeatedly and it will take resources and organisational discipline.

As to the format and the rendering tool there are several strategies one might follow. The most widespread method is migration to a new format, but one has to accept then as a drawback that some information might get lost on the way. The alternative is emulation: instructing a new rendering tool to behave like an obsolete one. Experiments done jointly by the Koninklijke Bibliotheek and the Rand Corporation proved that this is a viable technique. The method however is labour intensive and therefore costly.

Over the last couple of years a new method has been developed together with IBM: the Universal Virtual Computer. It's based on a combination of migration and emulation. Out of self-protection I will not go into the details. I am afraid I don't understand it myself. For details, consult the Koninklijke Bibliotheek's Web site for the e-depot project:
http://www.kb.nl/kb/resources/frameset_collecties-en.html [viewed 21 June 2004].

Strategies

Whatever strategy you choose to follow, it will always imply repeated actions. You will never have a dull moment. That's a certainty. There is uncertainty, however, as to what your actions exactly will have to be, because we don't know what future technology will be like. Therefore a permanent research and development effort is needed.

In the subtitle of my presentation there is one letter that has a special meaning. It is the last 's' in the word strategies. The plural is not accidental. By strategy I refer to both organisational strategies and preservation techniques. The plural indicates that in both fields several strategies might prove to be viable in the end. Moreover, the plural is meant to express uncertainty. Over the past years we all have witnessed rapid change. There is no reason to doubt that technology will keep on surprising us. The same holds for institutional and organisational responses to technological change. It's hard to prophesy what the dominant model for the dissemination of scholarly information will be like, ten years from now.

Uncertainty, however, does not discharge us from taking action. We all have to plot our course. When plotting a course, beacons are to be sighted. By this I mean developments or insights you relatively feel sure about. My beacons consist for a large part of insights gained from practical experiences at the Koninklijke Bibliotheek. These beacons comprise the requirements that must be fulfilled for a permanent archive to function properly.

Requirements for permanent archives

The first one looks a bit self-evident and silly: permanent archives presuppose permanent commitment. Self-evident or not, it is a fundamental requirement. A permanent archive that is meant to be credible should provide reasonable guarantee for continuity. Long-term preservation must be a major component of its mission and must belong to its core business.

Permanent archiving takes substantial resources, both organisational and technical ones.

Qualified manpower is needed. However intelligent your storage equipment may be, you never can do without human intervention and expertise. You also need dedicated ICT equipment, specifically designed for the job.

Moreover, sustained R&D efforts are required. There will never be a moment that will allow you to say that the job is completed. Technology will keep changing. Whenever new platforms or new formats emerge, you will have to face new challenges and prepare for counter attack. Again and again you will have to devise the means for maintaining accessibility. It's a never-ending story to keep your ever-changing toolbox up-to-date.

The good news is that there are also considerable economies of scale. The fixed costs of a permanent archiving system are relatively substantial. Once your system is working well, you can expand the storage capacity relatively easily, and costs per unit will go down. It primarily means expanding the number of disks, tapes or whatever storage medium is used. The more terabytes you store, the more economically your expertise will be deployed, together with the organisational arrangements and the preservation techniques.

So these are my beacons in a permanently evolving and uncertain environment. The future may prove that I am wrong, but this is the best I can offer now. These are the insights on which the strategy of the Koninklijke Bibliotheek is based.

From the requirements stated one can draw two basic conclusions. First, these requirements tend to narrow down the number of possible candidates for permanent archiving. Candidates should have the resources and the will to engage in a major long-term commitment. This means that long-term preservation should feature prominently among their strategic goals and be part of their mission. Permanent archiving cannot be a sideline activity or by-product.

The second conclusion is related to the economies of scale. It's an economic law that economies of scale inevitably result in a degree of concentration. Exploiting economies of scale therefore calls for co-operative efforts. Research and development efforts should also be shared. Preservation techniques should preferably be open source techniques. It wouldn't make sense for each research or university library to try to establish its own permanent archiving system. In the case of STM publications a handful of permanent archives, wisely spread around the globe, might suffice.

Archiving strategies for digital publications

It appears to me that, as regards the archiving of digital publications, three strategies are currently emerging. Yet, each of these strategies is still, one could say, in its pilot phase. As they are under way I can only explain them tentatively. I also have to make up names for them. I will describe them as *the Safe Place Strategy*, *the LOCKSS Strategy* and *the Institutional Repositories Strategy*. In all

the three of them storage of digital publication is a core issue, but they differ in how they emphasise long-term preservation.

I argue that only the first one, the Safe Place Strategy, makes long-term preservation its primary goal. The second one, the LOCKSS Strategy, aims at combating librarians' anxiousness on future access. Yet, as ingenious as it is designed, it's not clear to me how they envisage tackling the challenges of long-term preservation. The third one finally, the Institutional Repositories Strategy is primarily designed to serve other goals than permanent archiving. I am afraid advocates of this strategy severely underestimate the challenges of long-term preservation. This being said, in all probability you will not be surprised to hear that my library adheres to the first model, the Safe Place Strategy.

The Safe Place Strategy is directly derived from the requirements I stated earlier. From these requirements it follows that permanent archiving should be taken care of by a limited number of institutions, dedicated to this task. Permanent archiving should be prominent in their mission. The model clearly draws its inspiration from the deposit system in the printed paper world. In this view national libraries are natural candidates for permanent archiving. This has been their mission all the way through. Being a national librarian, I might be biased. This thought already crossed your mind of course, so I better bring it up myself. Therefore, I hurry to add that other institutions also may qualify, provided they meet the requirements and provided they are willing to take part in the global arrangements that are needed. Large library co-operatives could be an example of such institutions. STM publishers might wonder whether one single Safe Place is good enough to serve their purposes. For reasons of security and for political reasons it might be wise to select a number of Safe Places.

The next model is quite the opposite of the Safe Place Strategy. Instead of relying on a number of dedicated institutions, they seek safeguard in large numbers. It clearly draws its inspiration from the second component of the dual model in the paper world. I named this strategy after the LOCKSS initiative, a co-operative venture supported by the Mellon Foundation. The least you can say of it, is that their acronym is ingenious: Lots of Copies Keep Stuff Safe. In order to safeguard future access, libraries should request their own copy of digital publications to be stored in their own electronic stack room. The more libraries do so, the better chances are that future availability can be guaranteed. In the heart of the project is an electronic device that checks automatically whether a publication is still available in your stack room and whether it is still intact. The principles of authorised use are explicitly acknowledged. It's an elegant strategy. It has the attractiveness of any decentralised model, guaranteeing the common good as the outcome of free decisions made by autonomous agents. There is however a serious drawback. Long-term preservation implies permanent development and application of a preservation toolbox. As far as I can see, this is missing in the model. LOCKSS primarily responds to the anxiety of librarians about dependency on publishers for future access. It neglects the intricacies of long-term preservation technology.

The Institutional Repositories Strategy is closely related to the Open Archives Initiative. Its primary goals are not in the realm of permanent archiving. To begin with, academic institutions want to display the intellectual output of their faculty. That's only a natural and legitimate ambition. Yet, for the advocates of this strategy there is more at stake. They claim a role in the dissemination process of scholarly information. I do have views on this as well but I won't disclose them now. You then should have invited me for another session in your programme. You already discussed this issue earlier in your conference. The Institutional Repositories Strategy tends to underestimate or to neglect the requirements for permanent archiving. Safeguarding future accessibility is no by-product that automatically derives from establishing repositories. So, as for this purpose I have little confidence in the model. Universities indeed should take their responsibility for future access. May be in doing so, they might rely a bit more on co-operation with Safe Places as hosts and guardians of their intellectual output.

Why publishers should care

What about the role of publishers? Will they be passive, like they were in the paper world? There are good reasons for you to care and to get in-

volved. The most obvious reason is that customers ask you to take care of the problem. The "perpetual access" issue represents a direct commercial interest. There is another obvious reason. You want to safeguard authenticity and integrity of your publications. Permanent archiving, when implemented properly, by its very nature takes care of this.

I think there are yet two other good reasons. Publishers as a group form a major player in the field of standards, formats and software developments. Big players must behave in a responsible way. It might help enormously if preservation needs were to be taken into account right from the start, when implementing new developments. Publishers cannot be missed in the permanent process of developing preservation measures.

Finally, publishers share responsibility for the lifespan of the intellectual output of their authors. Authors don't like the idea of their output being lost.

So, I take the opportunity to extend a standing invitation to you, to actively participate in the preservation process. I also take the liberty to draw up a work programme for concerted actions.

Concerted efforts called for

We need to explore business models that help to recover the costs of archiving. I hate to bring up this financial issue. You probably find it typically Dutch. On the other hand costs and sales must be concepts that cross your own mind every now and then. Permanent archives serve large communities that extend far beyond traditional geographical boundaries. We must find ways to share the burden.

As I argued, a sustained commitment to the development of long-term preservation procedures and techniques is fundamental. In this process we need of course involvement from ICT vendors, but no less from the publishers.

In the printed paper world there is a clear allotment of responsibilities. Similarly we will have to develop global arrangements for the deposit and permanent archiving of digital publications. We will be very pleased to work with you on this programme.

References

LOCKSS (Lots of Copies Keep Stuff Safe). 2004. URL: http://lockss.stanford.edu/ [viewed 21 June 2004]

Royal Library, The Hague. 2002. IBM/ Koninklijke Bibliotheek Long-term Preservation Studies http://www.kb.nl/kb/hrd/dd/dd_onderzoek/dnep_ltp_study-en.html [viewed 21 June 2004]

The State of Digital Preservation: An International Perspective. (CLIR Reports, no. 107) Washington, DC: Council on Library and Information Resources, 2002. Full text (PDF and HTML) available online: http://www.clir.org/pubs/reports/reports.html [viewed 21 June 2004]

van Wijngaarden, Hilde & Erik Oltmans. 2004. Digital Preservation and Permanent Access: The UVC for Images. URL: http://www.kb.nl/kb/hrd/dd/dd_links_en_publicaties/publicaties/uvc-ist.pdf [viewed 21 June 2004]

Editorial history:
paper received 13 May 2003;
final version received 1 May 2004;
accepted 21 June 2004.