LIBER QUARTERLY, ISSN 1435-5205
© LIBER 2004. All rights reserved
K.G. Saur, Munich, Printed in Germany

Challenges in Digitisation - Paper given in honour of Esko Häkli, Helsinki, 23 November 2001

by ELMAR MITTLER

INTRODUCTION

The digitisation of information seems to be the realization of mankind's dream. We can - in principle - get seamless information everywhere and at any time. Digital information seems therefore to be a victory in terms of space and time. But, on the contrary, if we look at the reality of digital information on the World Wide Web, it sometimes looks more like a nightmare because of the overwhelming mass of material, lack of quality control, and lack of stability.

For these reasons electronic information presents a real challenge to libraries, the 'professional information institutes' of the world.

Three main aspects of digitisation and the digital library will be discussed in this paper:

- Retro-digitisation
- Research and the digital library
- Publishing as a distributed system

RETRO-DIGITISATION

The digitisation of printed material is one of the great challenges for national and research libraries as a way of improving access and availability. The overwhelming mass of material makes selection a problem. There are two main developments in this field:

- digitisation of cultural heritage (for an example from the SUB Göttingen, see the digitisation of the Göttingen Gutenberg Bible)
- digitisation of research material. The Deutsche Forschungsgemeinschaft e.g. is funding a research-based retro-digitisation programme.

A positive side effect of the digitisation of cultural material is that it allows libraries to combine activities intended for a wider audience with the provision of better access for researchers. There are new opportunities to unite materials that are scattered in different libraries internationally (e.g. the Library of Congress's programme Meeting of Frontiers, which is bringing together Alaskan and Siberian materials from its own collections, the National Library of Russia (St Petersburg), the State Library of Alaska (Anchorage), and the State and University Library of Göttingen.

But it is not only libraries that can be partners in this new method of virtual collection building for research purposes. It is a challenge for all types of heritage institutions - archives, libraries and museums - to bring together different materials relating to an individual writer, artist or subject.

For such projects to be successful, inter-operability between these different activities is essential. There are different levels of inter-operability. *Technical inter-operability* covers hardware, networks, data types, application compatibilities and protocols.

Standards are needed: for retro-digitisation of printed material, the GDZ Digitisation Centre in Göttingen uses the standards: 600 dpi for imaging, and Dublin-Core and XML for metadata description.

The Digital Library Federation's newly developed METS standard will be an interesting development for the exchange of digitised material.

Just as important as the technical standards is *information inter-operability*, which covers areas such as: content range, language, metadata, conventions of naming, and user interfaces.

It is certainly one of the main challenges for national and research libraries to move to common rules in the field of descriptive cataloguing. It is to be hoped that the new steps towards MARC Harmonisation, in which LIBER is taking a leading role in promoting to European activities, will succeed. But libraries must join with the whole research community - including researchers and learned societies as well as with other heritage institutions - to introduce minimum standards for inter-operability of metadata such as Dublin Core.

In Germany, the national library, Die Deutsche Bibliothek, is engaged on the Metalib project with the SUB Göttingen, with the aim of developing nation-wide standards in co-operation with the research community.

A third factor, which is often under-estimated, is *social inter-operability* covering personal and organisational rights and responsibilities, which have to be settled. In addition, it is necessary to develop partnership and mutual trust to achieve successful projects and long-term cooperation.

A new culture of partnership is necessary to meet the challenges of the virtual world. It will be essential for co-operative activities to be conducted on a basis of trust and stability if we are to succeed in creating a stable virtual information environment.

Journal Digitisation

The digitisation of journals is a good example of the necessity of co-operation. The first project of this kind was JSTOR, which was begun with a grant from the Mellon Foundation. It is now working quite successfully as a not-for-profit organisation. It was also a co-operative activity for many libraries, helping to bring together the holdings necessary to complete sets of journals. The DIEPER project, which demonstrates different activities in Europe, was more closely concerned with standards: it is based on the development of a single access point for digitised material from different countries, and this is more closely in line with European needs than the development of a nation-wide not-for-profit organisation like JSTOR.

The Digizeitschriften project, a co-operative activity by leading German libraries, is supported by the Deutsche Forschungsgemeinschaft. Sets of leading German journals will be digitised in the first phase.

DIEPER was a European project involving the State and University Library of Göttingen, in collaboration with the University Library of Finland, the University Library of Paris 5, the University of Pisa and the University of Patras.

JSTOR has good links with publishers, making possible the digitisation of older material. The concept of the “moving wall” - a period of three to five years between the original publication date and delivery time via the JSTOR service - protects the publisher against cancellations of subscriptions for recent materials. But some publishers are preparing retrospective sets of their publications. Elsevier seems to be the first publisher investing a substantial amount of money (about US$ 40 million) in retro-digitisation. In this particular instance, libraries may feel uncomfortable: a publisher with extremely highly priced journals identifies a new field of additional financial income.

Retrospective digitisation and copyright

“A library, that’s a good concept”, a publisher once said, comprehending the work of research libraries. In terms of copyright he saw an additional opportunity to exploit - and Elsevier may see its activities in a similar way. But perhaps the publisher made a mistake. If you look carefully at the copyright situation, you can see that material published before 1925 may be out of copyright. After 1996 the publisher will normally be considered the owner of both the electronic and the print copyright. Electronic publishing has been recognized during this time. So the author may have assigned the electronic copyright to the publisher without making a special contract. But the period between 1926 and 1995 is doubtful for the publisher. The legal situation seems to be clear: the author is the copyright owner of the electronic copyright, because he was not able to transfer the copyright into electronic publishing since it did not yet exist. This would, in effect, mean that every author has to be asked for permission for retro-digitisation - a seemingly totally unrealistic task, although in the case of the Finnish Electra project, which dealt with living authors in a relatively small country, this task was carried out quite successfully. The Digizeitschriften project has instead worked out a contract with the publishers and the German collecting society (Verwertungsgesellschaft Wort), which represents the authors, to solve the problem.

RESEARCH AND THE DIGITAL LIBRARY

The World Wide Web began as a communication medium for the research community at CERN in Geneva. Up to the present there continues to be a lack of clarity between communication and publication on the Internet. Traditionally, a publication such as a printed monograph, multi-volume work or a journal was a well-defined entity. But when is a text on the Internet published in the proper sense? The most significant part of the Internet comprises current communication - information about conferences, preprint versions of papers, discussion lists, chat rooms, etc. And often there are dynamic 'publications', where authors present new versions without showing the differences from previous versions.

There is another challenge for the digital library. Digital media are often dynamic in character. They comprise work in progress, or they are databases and are no longer linear text. In addition, co-operative virtual libraries on special subjects such as 'Meeting the Frontiers' are no longer static activities. They are more and more combinations of text, pictures and video-clips. If they are combined with researchers' comments, chat rooms etc, they are no longer definable entities - they are research in progress. They will become a 'digital research cluster'. These materials - not the official 'publications' themselves, which will also be part of these clusters – but the combination of all these materials will provide most interesting documentation of research and the way in which research is progressing. If 'research clusters' of this kind can be hosted and archived by libraries, the research library will become a research document in itself. It may be that this research-based activity is mainly a task for university libraries. Close technical co-operation and shared activities between national and academic libraries are, however, a real necessity if we are to archive these research materials and make them available for long-term access.

PUBLISHING AS A DISTRIBUTED SYSTEM

Copyright and alternative publishing activities

The 1996 WIPO Copyright Treaty and the 2000 European Directive on the Information Society have made the position of the publisher in the publication chain stronger than ever before. The key for access in the future will be licensing, and no longer to such an extent legal exceptions for libraries or for private use. In this situation activities like SPARC or the Public Library of Science (PloS) are of increasing value. SPARC wants to build a more competitive market especially for science journals, and to develop alternative communication schemes. The Public Library of Science is encouraging authors only to give their copyright to publishers if they agree to allow free access after six months.

There are many new and alternative business models in the digital world, as the results of the European project TECUP demonstrate:

- self publishing (by the author or institution),
- pre-print server (from learned societies or research communities),
- subscription model (electronic),
- pay per use (additional access to - seldom-used - electronic journals or new style ‘journals’ without issues),
- pay by author (publication in the electronic journal is costly).

Access is the real challenge in the digital world. You may be allowed to access or you may not:

- through consortia at national or regional level libraries try to guarantee cross access for the whole group of (bigger and smaller) partners;
- some publishers are offering different pricing models for smaller or bigger libraries;
- access via document delivery seems to be an additional way for seldom-used material.

But all these access models depend on the budgets of the university, the library or the user. There is a danger that we will have information rich and information poor people in the academic field of the future. But isn’t there another way outside the traditional publisher-driven system to assure academic information exchange in the digital world?

The distributed communication and publication system

1. The Open Archive Initiative

There are two different philosophies on how to create access to digital material. You can collect all the material in one big cooperative database, or you can collect it by harvesting it at points of interest. The open archive initiative tries to combine these two philosophies:

- data provider offers material with a standardized (Dublin Core) metadata set,
- service provider offers collected metadata in a special field providing access to the material stored by the data providers.

The Open Archives Initiative (OAI) can provide a structural model for the distributed communication system of the future. The players will be authors, institutions, universities and their libraries as data providers, and learned societies, research institutions and specialized libraries as service providers.

2. The digital university

Academic education is becoming more and more electronic-based. Universities are increasingly offering multimedia-oriented educational material and courses. Academic institutions without a networked environment are no longer compatible with the world of learning. The trend is moving in the direction of the digital university. Ideally, this academic institution must provide students and researchers with a personal, an intra-institutional and an Internet-oriented communication (and publication) system.

It should have three main levels:

1. Private level
2. Intranet level
   - 2.1. Course level: course material
   - 2.2. Institutional level: educational material
3. Internet level
   - 3.1. Communication level - preprints, dissertations, etc.
   - 3.2. Publication level - peer reviewed material; university press

The estimated content of a digital information system in such an e-university will provide as much content as that of a big publisher like Elsevier. Official publishing will be only a marginal part of the data storage and processing in the academic information systems. On the other hand, a development of this kind gives a real chance to change the situation in the publication chain between the academic world and publishing houses

(Roosendaal, Geursts & Van der Vet, 2001). By building a distributed system of open archive initiative content providers, combined with service providers of learned societies, libraries etc., research communication will be given a new foundation. The role of the publisher in the future will be to select high quality material out of this system, aggregating it with added value in peer reviewed journals, monographs etc. It may be that this system will really function as a distributed system, so that the papers are stored on university servers; or it may be that the added value papers are stored in their final version on publishers' servers - in any case access to academic material will be assured and the position of researchers and universities will have improved.

Only libraries - national and academic libraries - and the international community of libraries can build the worldwide infrastructure for

- this new system of distributed communication combining OAI-servers
- quality services for cataloguing and subject classification (e.g. via metadata sharing)
- free access via portals and
- online delivery of digital and digitised material.

In short, the real challenge for libraries, in co-operation with researchers, universities, learned societies and publishers, in the digital world is to build a worldwide infrastructure for the standardized communication, information and publishing system of the future.

But if they are to realize this vision, libraries need more leaders like Esko Häkli, who combine vision, co-operative skills and practical common sense.

Acknowledgements

I would like to thank Ann Matheson for having looked through the English version.

REFERENCES

Roosendaal, Peter A., Th.M. Geursts and Paul van der Vet: "Higher education needs may determine the future of scientific e-publishing". *Nature*, 18 September 2001.http://www.nature.com/nature/debates/e-access/Articles/roosendaal.html

WIPO Copyright Treaty and Agreed statements Concerning the WIPO Copyright Treaty (adopted in Geneva on December 20, 1996). WIPO Publication Number 226, ISBN: 92-805-0706-0, 56 p. http://www.wipo.int/clea/docs/en/wo/wo033en.htm

WEB SITES REFERRED TO IN THE TEXT

DDB - Die Deutsche Bibliothek. http://www.ddb.de/index_e.htm

DIEPER – DIgitised European PERiodicals. http://gdz.sub.uni-goettingen.de/dieper/

Digizeitschriften – das deutsche digitale Zeitschriftenarchiv. http://www.digizeitschriften.de/

DFG - Deutsche Forschungsgemeinschaft (German Research Foundation). http://www.dfg.de/en/index.html

DLF - Digital Library Federation. http://www.diglib.org/

Funded Projects in the DFG Program Retrospective Digitisation of Library Holdings. http://gdz.sub.uni-goettingen.de/en/vdf-e/vdf-liste-e.shtml

JSTOR – The scholarly journal archive. http://www.jstor.org/

Göttingen Gutenberg Bible. http://www.gutenbergdigital.de/gudi/start.htm

LIBER MARC 21 Interest Group. http://www.kb.dk/liber/division/taskforce/marc/index.htm

Meeting of Frontiers. http://frontiers.loc.gov/intldl/mtfhtml/mfsplash.html

OAI - Open Archives Initiative. http://www.openarchives.org/

PloS - Public Library of Science. http://www.publiclibraryofscience.org/

SPARC - the Scholarly Publishing and Academic Resources Coalition. http://www.arl.org/sparc/

SUB – Staats- und Universitätsbibliothek Göttingen. http://www.sub.uni-goettingen.de/index-e.html

TECUP. http://gdz.sub.uni-goettingen.de/tecup