


DANISH NATIONAL
LIBRARY AUTHORITY

Digitising Journals

Conference
on future strategies for
European libraries
13.-14. March 2000 Copenhagen
Proceedings

Copenhagen, June 2000

Digitising Journals
Conference on future strategies
for European libraries
Proceedings

2000

Edited by: Vibeke Cranfield

Published by:
Danish National Library Authority/
Denmark's
Electronic Research Library
Nyhavn 31 E
DK-1051 Copenhagen K

Phone: +45 33 73 33 73
Fax: +45 33 73 33 72
E-mail: def@bs.dk
Homepage: www.deflink.dk

Layout: Kühnel Design AS
Printed by: De Facto AS

Number printed: 500

ISBN: 87-87012-65-0
ISBN (electronic): 87-87012-66-9

The publication is also available at:
www.deflink.dk

Contents

5	Preface
7	User-landscapes: Needs and visions of users
11	Digitising journals and the eLib programme
19	DigiZeit (Digitisation of journals)
25	DIEPER - providing web access to retro-digitised periodicals at multiple sites
31	Digitising journals: Highlights from the JSTOR's experience
43	Standards for images and full text
47	Digitisation - technical issues: production at the Göttingen Digitisation Center (GDZ)
53	Metadata and identifiers for e-journals
69	Three stories from the future
77	Selecting journals for digitisation: piecing together the puzzle to create a European model
85	A European model: Organisation
91	Conclusions and recommendations

Preface

The need for discussing digitisation problems at European level became obvious at LIBER's annual meeting in 1999. The idea of arranging a conference emerged and the Danish National Library Authority offered to organise this event.

Digitising Journals: Conference on future strategies for European libraries took place in The Black Diamond in Copenhagen on 13.-14. March 2000. Its organisers were LIBER and the Danish National Library Authority / Denmark's Electronic Research Library in co-operation with the European project DIEPER, the North American JSTOR project and NORDINFO, the Nordic Council for Scientific Information.

The aim of the digitisation conference was to prepare the ground for the development of national policies and organisational solutions at national level and to try to identify practical goals for international co-operation. We went in search of answers to questions like: Why should we digitise, what factors must we concentrate on and what does it involve for libraries? We were introduced to a European virtual library of digitised materials and were invited to "take a leaf out of JSTOR's book"!

Concepts such as European Gateway, European Centre for Journal Digitisation, European Agent, Critical Mass of Digitised Material, Standardisation etc. were all deemed to be important bricks in the building of a European model. The Proceedings of the conference illustrate the

great variety of issues and problems involved in an international digitisation process and also provide a valuable insight into possible solutions.

The mood of the conference was buoyant and visionary and by general consensus an important step forward had been taken.

The programme committee of the conference consisted of:

Trix Bakker, The Netherlands
Sigrun Klara Hannesdottir
NORDINFO
Esko Häkli, Finland
Alex Klugkist, The Netherlands
Elmar Mittler, Germany
Simon Tanner, UK
Jens Thorhauge, Denmark

I should like to take this opportunity to thank the committee for a most fruitful co-operative effort. As a direct result of the conference, a working group has been appointed, whose members are now engaged in further examining - and tackling - the problems posed.

Jens Thorhauge

User-landscape: Needs and visions of users

*Jean-Pierre Bourguignon
Institut des Hautes Études Scientifiques, Paris, France*

First of all, I need to position myself since the task the organisers assigned to me, namely introducing the point of view of users, is different from that of other speakers. This is very challenging since, as a mathematician, I stand at one end of the spectrum of users. Therefore, there is a risk that my presentation be biased by the special relation that mathematicians entertain with bibliographic documents, as explained later.

Two other pieces of information may be relevant here. First, as President of the European Mathematical Society I took part in the establishment of the free server EMIS (European Mathematical Information Service) which includes a Digital library containing already 40 journals. Some institutions have agreed to include back issues of their journals in this virtual library at no charge. Second, as director of the Institut des Hautes Études Scientifiques, I am presently in charge of editing the Complete Works of Renée Thom, one of the Institute's professors who made himself known as mathematician, theoretical biologist and philosopher. This will be presented as CD-ROM whose source will be a TeX-file including also the books he published. We developed the software needed to have the conversion from TeX to HTML available on line, hence making this text a fully searchable document, including the formulas.

As a researcher I use journals very frequently, in several different ways explained later. As a teacher, I also consider that making proper use of libraries is part of the training of a professional scientist. Libraries are wonderful places for intellectual stimulation and work. The new age we are entering in must not break the partnership that we developed with librarians. Even if libraries are going to have an important virtual component, we are looking forward to working with librarians who will remain human beings with renewed, and probably extended, professional skills.

Many points I will raise are likely to have their echoes in other disciplines, although differences in the sociologies of the different fields have to be seriously taken into consideration.

If the basic technical requirements are fundamentally the same across disciplines, it may be necessary to develop extra modules to meet special needs.

My presentation is organised around the questions asked by the organisers to other speakers presenting their projects, but after having turned them around.

1. Which documents ?

The documents we are looking for in journals are most of the time texts

**User-landscape:
Needs and visions
of users**

with formulas and diagrams. Occasionally, but this happens more and more often, there can be pictures, drawings or simulations.

In journals, mathematicians are interested in information of at least three different natures:

- they may be interested in personal views of an author on a topic; in this case the use of a document is not very different from that of colleagues in humanities, and for such a document the user may be satisfied by being able to have a look at the document in a scanned form;
- they may be looking for a precise statement, like the statement of a theorem; for that purpose any reliable source taking up the question is fine, and the original is not indispensable, although a number of mathematical theorems are only stated in one place; note here that the date at which the result was produced has no correlation with the interest of the reader; indeed, there is often no obsolescence in mathematical documents since properly documented mathematical knowledge does not age, a fact that makes mathematicians' needs close to that of historians, and may distinguish them from specialists of natural sciences, where accepted theories have a much shorter life span; note that for this use, the main difficulty is to devise help mechanisms since, for the moment, no search engine is able to locate a statement (in its abstract meaning);

- they may need a special point in a proof; this often requires to have access to a very specific document that should be fully searchable and easily accessible, no matter what its age and state are; here the need may go as far as loading the information or printing it since one may have to think about it for a while until one has grasped in depth the relevant information.

2. Which form for the on-screen presentation ?

On-screen presentation can take two fundamentally different forms:

- images (unavoidable when pictures, or even simulations are to be shown, but, for the moment at least, they are much slower to load)
- structured text (far better for formulas).

It is to be noted that the main advantage of structured text is its full searchability, and that many recently produced texts submitted by scientists are spontaneously in this form because of the widespread use of text processors such as TeX (in various disguises though). The MathML standard, an XML application designed through the collaboration of mathematicians, publishers and software houses, is very promising for the use of mathematical texts on the World Wide Web. For past documents, whose treatment is likely to occupy a good part of the digitising process, to go from a scanned text to a fully structured one still requires a lot of

work if one insists on full accuracy. JSTOR has found a good compromise (which is also consistent with its philosophy of not letting the full data be made available to users). The document is put at the disposal of users as images but in order to make a search possible it is coupled with a structured text which, as such, is kept out of reach of users, thus preventing uncontrolled duplication of the data.

Standards have to be established in order to facilitate the translation, reduce its costs and ensure compatibility. It is there that non-co-ordinated attempts must absolutely be avoided... but this is exactly the purpose of this conference.

3. Which access ?

This item can be understood in several ways. First, it can mean in what physical form is the documentation put at the disposal of users but this question has been dealt with in the previous section. "Access" can also be taken as defining what segment of the documentation is to be covered and which economical organisation is adopted to regulate the access, and it is in this sense that we take it here.

We claim that mathematics is a discipline where systematic and universal storage can actually be achieved. This will not happen unless a co-ordinated multinational effort is put in place. It is so because the mathematical literature is very well-identified and concentrated almost fully in journals that can be taken

from a well-controlled list. Although very substantial, the volume of data has become within reach by modern means.

The most efficient access would certainly be hyperlinks from a bibliographic database to articles. This would of course be ideal since mathematics is served by comprehensive bibliographic databases, Zentralblatt MATH and Mathematical Reviews, that cover almost exhaustively the mathematical production (since 1931 for the first one and 1941 for the second one), and a less systematic one, das Jahrbuch über die Fortschritte der Mathematik covering the period 1868-1943. It seems that JSTOR has already the plan hyperlinking its data to the Mathematical Reviews.

Now, turning to the economy of the whole process, if the cost of the access to the full document follows economical rules presently used for journals (namely that one has to pay to have access to them, even if prices vary very much from one publisher to another), references to digitised documents must remain free of charge without any restrictions.

4. Conclusions

Let me first briefly address points that I did not touch upon during the presentation.

I deliberately did not mention legal issues (copyright, differences in legal systems,...). It does not mean they

**User-landscape:
Needs and visions
of users**

are not important but I felt rather incompetent as far as they were concerned. Achieving a system protecting the rights of authors is extremely important for users. (Note this has been the basis for a long-lasting battle, in particular by the French government in the framework of the World Trade Organisation).

I also remained deliberately naive. Concerning the issues under discussions, publishers, software companies, but also learned societies, adopt commercial strategies. The questions discussed are also partly linguistic, leading to other types of strategic power struggles, the more so that an economically thriving component directly connected to knowledge is developing very quickly in our society.

There are great expectations from the mathematical community - and others. Being able to have full access to a number of journals may potentially have a major impact on the way research will develop. One aspect of this is, if access is affordable, making it possible for remote centres to have wide access to data, an unexpected turning point, that could pave the way to solving at once the huge problem of access to documentation of developing countries for example.

Here is a summary of my main points:

- Standards must be designed very carefully

- in particular one must absolutely ensure access to ancient data (and anticipate the inevitable changes in standards, etc.)
- and also one must take into account the point of view of a great variety of direct users from an early stage.
- In the long run, one should also introduce a principle of public access to sufficiently old data, as elements of the human heritage.

Jean-Pierre Bourguignon
jpb@math.polytechnique.fr

Digitising journals and the eLib programme

Astrid Wissenburg

Information Services and Systems, King's College London, United Kingdom

This paper describes recent activities in British Higher Education in the area of digitisation of journals, by focusing on the UK Electronic Libraries programme (eLib) and related initiatives. It is not intended to give a comprehensive review of such projects in the United Kingdom, or indeed of the eLib programme itself, but aims to give a sketch of the outcomes of relevant projects. The paper will also draw some conclusions as to the benefit of these projects to libraries in UK Higher Education

The eLib programme: background and projects

The United Kingdom has four Higher Education funding bodies (Scotland, England, Northern Ireland and Wales) responsible for the distribution of public money for teaching and research to Higher Education Institutions. Although most of this money is allocated according to set formulas directly to the institutions, a small part is top-sliced to fund national activities. The Joint Information Systems Committee¹ is one such centralised initiative, supported by all four funding bodies. JISC, with an annual funding in the region thirty-five million pounds², provides a high quality national network infrastructure (called JANET) for the UK higher education and research councils' communities, as

well as stimulating and enabling the cost-effective exploitation of information systems. The latter part of this remit has increasingly been fulfilled through providing electronic content and services. Approximately two thirds of the funding is spent on the network, the remainder on electronic content and services³. Initially, its focus has been on supporting research, but through a recent initiative (between ten and fifteen million pounds) a range of projects improving support of teaching will be funded⁴. Also, in recent months the remit of JISC has been widened to include the Further Education sector, which covers vocational education.

In the 1980's major changes took place in the Higher Education sector in the United Kingdom, when the number of universities was doubled following a re-classification of the level of tertiary educational institutions. Subsequently a major review took place of the needs of libraries, best known as the Follet review, after its chair. The Committee's report from 1993⁵ has had a considerable impact on the sector and resulted in several initiatives with significant funding attached to them, totalling more than 100 million pounds⁶. In response to the report JISC initiated the eLib programme⁷. This was not intended as a research programme, but was to provide a

Digitising journals and the eLib programme

body of tangible electronic resources and services and to effect a cultural shift to the acceptance and use of those in place of more traditional information storage and access methods.

Phases I and II (1995-1999) of the eLib programme consisted of about 60 projects in eleven different programme areas⁸. The total funding for these phases was about fifteen million pounds, giving the average project less than £60,000 a year. Most projects were small and only ran for one to two years. The majority of projects finished several years ago, and finding information about their activities and findings is becoming increasingly difficult as web sites disappear and staff moves on. Although there was no programme area dealing specifically with the digitisation of journals, several projects did undertake such activities. The two projects funded in the programme area of Digitisation were concerned primarily with digitisation of journals. In the area of Electronic Journals, the majority of projects focused on creating new electronic journals, but a few did include digitisation of existing journals, as did several projects in the areas of Electronic Document Delivery and On-Demand Publishing.

eLib phase 3 (1998-2001) is intended to build on the first two phases and integrate the outcomes. The five hybrid library projects are exploring the integration of traditional print based with electronic sources creating a hybrid

environment⁹. The four large-scale resource discovery projects are investigating the use of the Z39.50 protocol for search facilities across library catalogues. And, reflecting the increasing concerns in the library sector, one project (CEDARS) focuses on preservation. In addition funding was given to some of the earlier projects to establish themselves as services. In general, the phase 3 projects are larger, both in terms of funding (five million pounds for phase 3) and duration (two to three years). The projects are still running, with the first ones due to finish this summer, making it difficult to draw conclusions on the outcomes at this stage.

Digitising journals - Examples from the eLib programme

In all three phases of the eLib programme there has never been a specific programme area dealing with all issues surrounding the digitisation of journals, although, as mentioned above, several projects did include relevant work. The following examples describe four projects, which included the actual digitisation of journals, and three examples of projects covering related questions.

The Internet Library Of Early Journals¹⁰ digitised a substantial run (approximately 20 years) of six core journals from the eighteenth and nineteenth century, for example Philosophical Transaction of the Royal Society. The reasons given for the digitisation is to improve the availability of the materials to scholars

and to preserve the originals. Although perhaps more a historical resource than a regular scholarly journal, the project is of particular interest for its extensive costing of all phases of the digitisation process, including continued access and availability. The journals are still available freely online.

CLIC Consortium Electronic Journal Project¹¹ is an example of a close collaboration with a publisher, in this case the Royal Society of Chemistry, a professional association. In addition to digitising Chemical Communications, the Royal Society now has almost 22 online journals available, 3 of which are published in electronic format only from 1996 onwards. The digitisation of Chemical Communications has included experimentation with interactive displays of molecule elements to provide added value. The journals are currently available from both the Society and commercial providers such as Blackwells and Swets-Net.

The Journals And Transactions Of The Institute Of British Geographers¹² is another example of collaboration with a professional association. In this case the digitisation was undertaken by HEDS¹³, the Higher Education Digitisation Service, a JISC-supported service providing digitisation facilities and expertise. In this case journal runs were digitised from 1935 until 1997, more than 20,000 pages. The journals are now available in the UK via the BIDS ingenta service.

The BUILDER¹⁴ project is one of the phase III hybrid library projects and

is based at the University of Birmingham. As part of the project the complete run of the journal Midland History (starting in 1971) was digitised, approximately 4,000 pages, again with the HEDS. The Editorial Board based at the University of Birmingham manages access to the electronic version.

The SuperJournal¹⁵ project is an example of a project where the content creation - the digitisation - was the responsibility of the publishers. The project itself developed and implemented an access and delivery application, emphasising the required functionality of such a service. The project provided free access to over 50 journals for its 13 test sites. The licence for these ran out at the end of 1999.

LAMBDA¹⁶ is an electronic document delivery project. A small number of university libraries in Manchester and London digitise and deliver local library holdings, such as journal articles, in either electronic or paper format, but now included. Under current licence conditions, the electronic files need to be deleted immediately and if requested a second time, digitising the item again will create a new version. The project is one of a few examples of a successful transition to a self-financing status, and now includes 10 supply libraries.

Finally, the HERON¹⁷ project is an example of a digitisation on demand service, which will include journal articles. The project aims to develop

Digitising journals and the eLib programme

a national database and resource bank of electronic texts for teaching purposes. The publishers, the Higher Education institutions requiring the materials or HERON itself, will digitise materials. The test phase started in May 1999.

It is difficult to draw any general conclusions based on these examples and the eLib programme in general, but a few observations can be made.

The main incentive for digitisation of journals appears to have been to increase access for users. A second reason was to assist in the archiving and preservation of the materials, this despite the many question marks surrounding digital archiving and long term preservation. The main obstacle seems to have been ownership and copyright, and in most projects that did undertake digitisation, one of the project partners owned the copyright in the journals. The emphasis appeared to have been on providing access to recent years, not digitising long back-runs. As a consequence, close collaboration with the publisher (either commercial or a professional association) was a common element. The actual scope of most digitisation projects was relatively small scale. All of these characteristics can be traced back to the issue of ownership and copyright of the journals.

It is more difficult to comment on the technical standards used: most of the eLib projects undertook their digitisation at least two years ago and standards in this area change quickly. There is indeed quite a difference

between the earlier and later projects. Most projects digitised from paper. The use of microfilm appears to have never been a serious option for any of the projects. The Internet Library Of Early Journals is an exception: it used existing microfilm versions as a basis for scanning of two journals, in order to protect the paper originals. No examples could be found of microfilming the paper originals for archiving purposes as well as a basis for scanning. Rather the digitised version seemed to be perceived as an archival copy. Part of this might have been due to financial restrictions: the only example of use of microfilm showed the scanning of it to be more expensive than paper and the creation of the microfilm version itself would have involved extra work and costs.

There appears to have been a general consensus for resolution standards at 600 dpi for scanning and archiving and at 300 dpi for delivery. The file formats used (with some exceptions) were PDF for archiving and PDF with Hidden text (based on uncorrected OCR¹⁸ versions of the images) as delivery options. Full text versions based on corrected OCR were considered too expensive.

The service offered by most projects in terms of presentation and access of the journals, included all of the following: a table of content, search facilities for bibliographic metadata such as author and title, free text searches based on uncorrected OCR and images of the actual articles (mostly in PDF). Sometimes abstracts

of articles were offered as well. In most cases access to the metadata, such as table of contents and abstracts, was free. For full articles subscription was necessary, especially where titles were produced by commercial publishers and in many cases this subscription was linked to subscriptions to the print publications.

As with the technical standards, it is difficult to draw any general conclusion on costs of digitisation considering that over the lifetime of the eLib programme these have changed considerably so any of the following figures should be treated with extreme caution. The process of scanning was itself considered cheap, but proof-reading to create corrected full text versions of a sufficiently high standard for on-screen display was found to be extremely expensive. Therefore, the majority of projects adopted a compromise between the two approaches: uncorrected OCR to create automated indexes allowing users to search the text. The cost for just scanning appeared to be about 20 pence, but coming down. Scanning plus indexing, based on uncorrected OCR, increased the price to 75 pence. Scanning plus proof-reading upped the price to £3.

Most projects only reported on the costs of the production process itself, omitting the costs of managing these projects and of continued access to digitised journals. The Internet Library Of Early Journals did provide details on this and quotes £4.20 for scanning (including manage-

ment and keyboarded indexes, but not proof-reading), 3 pence per page per year for continued access and 2 pence per page per year for archiving¹⁹.

eLib and related initiatives

The eLib programme has not been the only recent programme within the Higher Education sector in the United Kingdom of relevance to the conference theme of digitising journals, in particular considering licensing issues.

The Pilot-Site Licence Initiative (PSLI, 1996 - 1998), directed by the Funding Councils, aimed to make academic journals cheaper and more accessible for academics and students. Four publishers (Academic Press, Blackwells, Blackwell Science and The Institute of Physics Publishing) offered their printed journals at discounted prices to universities and colleges throughout the UK, and this included the availability of a range of electronic material. The project was not aimed at establishing a service, but wanted to explore the concept of national licences. The evaluations of this initiative²⁰ make it clear that electronic journals were, especially by the publisher, seen as an additional facility rather than a (potential) core resource, although these perceptions did change somewhat over the lifetime of the project. There were savings on the institutional journal budgets for the duration of the project and in most cases that recovered resource remained within the library budgets. However, few subscriptions were cancelled, mostly due to uncertainty about continuation of the PSLI.

The PSLI did have a large impact on the next initiative: NESLI, The National Site Licensing Initiative²¹ (1998-2002), which is directed by JISC.

The Managing Agent for NESLI is a consortium of Manchester Computing at the University of Manchester (which is host to several national electronic services) and Swets and Zeitlinger. There are considerable differences with PSLI: a particular focus on electronic journals as a separate product, the ultimate aim to become self-financing and no government funding for the publishers involved. NESLI co-ordinates the delivery of electronic material and undertakes negotiations with publishers: licences with individual publishers are offered to the Higher Education institutes to subscribe to. It is somewhat unsatisfactory that so far most NESLI deals do not treat electronic journals as a separate product and still combine print and the electronic subscriptions, and that prices are still rising. This appears to be mostly due to the publishers rather than NESLI.

The Future

The JISC has recently published several documents outlining their vision of the future: the Distributed National Electronic Resource²²: "a managed environment for accessing quality assured information resources on the Internet which are available from many sources". Potentially, this managed environment could offer easy access to a variety of electronic resources, including scholarly journals, hopefully at a reasonable price to the HE community throughout the UK. But at the

moment, the view from an institutional perspective is rosy. There are new and imaginative services, but print and electronic subscriptions are still too often combined. Archiving and continued access remains a concern. Very few libraries have dared to start weeding out print journals, and no space savings have been made so far. At the same time there is a proliferation of interfaces and access arrangements, creating a considerable overhead on the local delivery of electronic services. In conclusion, so far institutional costs continue to rise in the absence of a realistic strategy for digitisation of journals and other scholarly resources which has the support of all partners, including the libraries and publishers.

Astrid Wissenburg

astrid.wissenburg@kcl.ac.uk

References

- ¹ Joint Information Systems Committee, URL: www.jisc.ac.uk/
- ² Estimate based on the 97/98 funding allocation. JISC - Constitution and Funding Arrangements, URL: www.jisc.ac.uk/jisc/const.html
- ³ Chris Rusbridge, Towards the hybrid library, D-Lib Magazine, July/August 1998, URL: www.dlib.org/dlib/july98/rusbridge/07rusbridge.html
- ⁴ JISC Circular 5/99, Developing the DNER for Learning and Teaching, 5 November 1999. URL: www.jisc.ac.uk/pub99/c05_99.html

- ⁵ Joint Funding Council's Libraries Review. Report (The Follett Report). Bristol: HEFCE, 1993. URL: www.ukoln.ac.uk/services/papers/follett/report/
- ⁶ Chris Rusbridge, Towards the hybrid library, D-Lib Magazine, July/August 1998, URL: www.dlib.org/dlib/july98/rusbridge/07rusbridge.html
- ⁷ Electronic Libraries (eLib) programme. URL: www.ukoln.ac.uk/elib/
- ⁸ The programme areas are: Access to Network Resources, Digitisation, Electronic Document Delivery, Electronic Journals, Electronic Short Loan Projects, Images, On Demand Publishing, Pre-prints, Quality Assurance, Supporting Studies and Training and Awareness.
- ⁹ Stephen Pinfield, Jonathan Eaton, Catherine Edwards, Rosemary Russell, Peter Wynne, Astrid Wissenburg, Realizing the Hybrid Library, D-Lib Magazine, October 1998, URL: www.mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/october98/10pinfield.html
- ¹⁰ Internet Library of Early Journals. URL: www.bodley.ox.ac.uk/ilej/
- ¹¹ CLIC Consortium Electronic Journal Project. URL: www.ch.ic.ac.uk/clic/
- ¹² The Journals And Transactions Of The Institute Of British Geographers. URL: www.ingentajournals.bids.ac.uk/Pub-info/rgs.html
- ¹³ Higher Education Digitisation Service. URL: www.heds.herts.ac.uk/
- ¹⁴ Birmingham University Integrated Development and Electronic Resource. URL: www.builder.bham.ac.uk/
- ¹⁵ SuperJournal. URL: www.superjournal.ac.uk/sj/
- ¹⁶ LAMBDA. URL: www.lamdaweb.mcc.ac.uk/
- ¹⁷ HERON. URL: www.stir.ac.uk/infoserv/heron/
- ¹⁸ OCR: Optical Character Recognition
- ¹⁹ INTERNET LIBRARY OF EARLY JOURNALS (January 1996 - August 1998) A project in the eLib programme. FINAL REPORT March 1999. URL: www.bodley.ox.ac.uk/ilej/papers/fr1999/
- ²⁰ Report on Phase I of the Evaluation of the UK Pilot Site Licence Initiative, April 1997, HEFCE Ref M 3/97. URL: www.niss.ac.uk/education/hefce/pub97/m3_97.html
Evaluation of the UK Pilot Site Licence Initiative - Phase II, May 1998, HEFCE Ref 98/22. URL: www.niss.ac.uk/education/hefce/pub98/98_22.html
- ²¹ The National Site Licensing Initiative. URL: www.nesli.ac.uk/
- ²² Documents about the DNER. URL: www.jisc.ac.uk/pub/#dner

DigiZeit (Digitisation of journals)

- a joint effort of special subject collection libraries in Germany

*Norbert Lossau and Stefan Cramme
Göttingen State and University Library, Germany*

*The article is a comprehensive summary of the presentation at the conference *Digitising Journals: Conference on future strategies for European libraries**

1 Background information: Retrospective digitisation in German libraries

A coordinated effort to digitise library holdings and other research material began in Germany in 1997, when the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) launched its program for "Retrospective Digitisation of Library Holdings". As a first step, general guidelines for technical issues and selection criteria were laid down¹. Based on these guidelines, several individual projects were started, currently numbering 43, which in total receive about 3 to 4 million DM per year from the DFG².

The main goals of this national program are:

- online access to relevant research collections in libraries
- simultaneous access to frequently used literature
- digital availability of collections which pose difficulties in conventional access (e.g. for conservation reasons)
- intensified use of lesser-known library collections.

So far, the items included in the retrodigitisation program consist mainly of primary sources (papyri, letters, illustrations, ...) and reference works (encyclopaedias, dictionaries). DigiZeit is the first German project to explore the systematic retrodigitisation of scholarly journals.

2 DigiZeit - Digitisation of journals

DigiZeit is structured along the lines of the German system of distributed special subject collections (Sondersammelgebietenprogramm), which has begun to include not only printed material but also digital publications³. The idea of furnishing retrodigitised journals was stimulated by the American JSTOR program, with several modifications demanded by differences in library and publishing structures.

2.1 Overall goals

The main goal of DigiZeit is to enable a better access to core journals from German publishers in the light of the recent trend towards electronic publishing. It aims to improve the information infrastructure, in the first place for researchers in Germany. But another objective of the project is to increase the global visibility of German scholarly journals by presenting them online and integrating them in international services.

2.2 Project schedule

DigiZeit started in February 1999, when a project officer was appointed at the Göttingen State and University Library to conduct a feasibility study for the full project. In co-ordination with all project partners, this feasibility study will be finalised in summer 2000, when it will be the base for a grant application to the DFG. The first production phase of DigiZeit could start at the beginning of 2001.

2.3 Project partners

Currently, nine libraries are project partners for DigiZeit, each responsible for one or more subject fields:

- Staatsbibliothek zu Berlin - Preussischer Kulturbesitz
- Universitäts- und Landesbibliothek Bonn
- Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden
- Stadt- und Universitätsbibliothek Frankfurt am Main
- Universitätsbibliothek der TU Bergakademie Freiberg
- Niedersächsische Staats- und Universitätsbibliothek Göttingen
- Bibliothek des Instituts für Weltwirtschaft - Deutsche Zentralbibliothek für Wirtschaftswissenschaften Kiel
- Universitäts- und Stadtbibliothek Köln
- Bayerische Staatsbibliothek München

Other libraries participating in the program for special subject collections can be included in a second phase of the project.

2.4 Selection of journals

The selection of journals for DigiZeit has to conform to the main goal of providing access to publications which are used intensively. So it was important to identify journals highly relevant to research and with a proven record of high frequency of use. A selective survey of scholars and scientists was carried out in 1999 to get balanced recommendations, which in turn were backed up and corroborated by analysing the coverage in citation indices, review journals and recommending bibliographies. The number of copies held by German research libraries was another factor indicating if a journal is of core importance to its field.

By these means, 57 journals have been selected, comprising about 3 million pages in all. 23 of them started to appear before 1900, the majority of the remainder in the first half of the 20th century. They belong to 14 subject fields:

- English and American studies
- German studies
- Romance studies
- modern philology
- library science
- contemporary art (after 1945)
- history
- sociology
- population studies
- economics
- business studies
- law
- general science
- geology

2.5 Rights holders (publishers and authors)

Taking another cue from JSTOR, DigiZeit wants to cooperate closely not only with libraries and researchers but also with the publishers of the journals. In a first meeting of the project partners with representatives of the publishers, there was a distinctly positive response. It was resolved to build a first prototype server in the next few months, which will include selected volumes to demonstrate the feasibility of the project.

Currently, the majority of the selected journals exists only in print form, although this might change in the future, and for this eventuality a seamless integration of retrodigitised and current digital journal issues is anticipated. Until then, it will be best to imitate JSTOR in defining a "moving wall" for each journal, safeguarding publishers' sales of current issues. Licensing negotiations are prepared at the moment and might be settled soon, even considering some complications caused by German copyright law: the publishers have acquired the right for digital use only for newer articles, published in the last few years, whereas for the majority of the articles the authors can still claim copyright in regard to any digital use. A possible partner in finding a solution protecting authors' rights without unduly inhibiting the project is the German collecting society Verwertungsgesellschaft (VG) Wort, which has been included in the negotiations in the meantime.

2.6 Organizational model

The organizational structures for DigiZeit have not been finally decided yet. It will likely consist of a not-for-profit-organization with a small head office, coordinating several distributed partners and using existing library capacities, but also commercial service providers, if this has any advantages. Part of the production process and/or hosting the document server might be sourced out in this way.

2.7 Pricing and availability

In concurrence with its aim of improving access to important scholarly publications, the main customers of DigiZeit will be universities and research libraries in the form of site licences. Although the start of the project is expected to be subsidized by a grant of the DFG, all costs incurred in the regular service have to be regained from the customers. So there will likely be an annual fee for recovering maintenance costs and allowing the extension of the service. There might be differentiations in pricing just as JSTOR has them.

2.8 Production

(For the technical issues of digitisation see also p. 47 ff)

DigiZeit will presumably use a combination of outsourcing and in-house work. The special subject collection libraries will retain the responsibility for the journals in their subject area, at least regarding the provision of complete copies and the preparation for scanning. A central administration

will coordinate the production process, specify the technical guidelines and parameters, set the time schedules and be the main contact point for external vendors.

2.9 Access

DigiZeit will use an image-based approach for scanning and delivering the journals' content. All articles will be scanned and indexed (for some of the titles, this will furnish the first complete index ever), including reviews, short notes and abstracts, so as to maintain the integrity of the journals. Existing indexes will be included to improve access via subject search.

A full text search is highly desirable, but might initially be restricted to part of the material, depending on the degree in which the OCRing can be automated and executed at acceptable costs (with the special problem that several of the journals have long runs in fraktur type, which do not allow for automatic OCR).

A search, whether in metadata or full text, will lead to the delivery of the page image on the screen, possibly including highlighting of the search terms in the images (see paper on technical issues). There will be no user access to the full text, but of course the possibility of print-outs. The format in which these will be delivered depends on the technical system used, which has not been decided yet. In any case it will be essential to base the system on a database solution, e.g. an extension of the new Document Management

System AGORA already in use at the Göttingen Digitisation Centre.

3 Outlook

The two big goals for DigiZeit in the remainder of 2000 will be reaching an agreement with publishers and authors' representatives and securing funding for the first phase of the full project. Until this will start, possibly in early 2001, the technical and organizational standards for the production process have to be defined as detailed as possible. The demonstration server, which hopefully will be implemented in the autumn, will be of great value in this.

Ultimately, it is to be expected that DigiZeit will be integrated in other national and international projects like the ones presented at this conference.

Stefan Cramme
cramme@mail.sub.uni-goettingen.de

Norbert Lossau
lossau@mail.sub.uni-goettingen.de

References

- ¹ Retrospektive Digitalisierung von Bibliotheksbeständen: Berichte der von der Deutschen Forschungsgemeinschaft einberufenen Facharbeitsgruppen "Inhalt" und "Technik". Berlin: DBI, 1998. URL: www.sub.uni-goettingen.de/ebene_2/vdf/einstieg.html
www.sub.uni-goettingen.de/ebene_2/vdf/empfehl.html

²See list at URL:

www.sub.uni-goettingen.de/gdz/de/projects/vdf.de.html

³Weiterentwicklung der überregionalen Literaturversorgung: "Memorandum", Zeitschrift für Bibliothekswesen und Bibliographie 45 (1998), 135-164; URL:

www.dfg.de/foerder/biblio/memo.html

DIEPER - providing web access to retro-digitised periodicals at multiple sites

*Werner Schwartz
Göttingen State and University Library, Germany*

Digitisation of printed material is becoming more popular with libraries every day. There are a number of reasons for this. One of them and perhaps the most important in the long run is to ease access to printed works that are heavily used. The digitised edition of a book enables users to read it irrespective of the library's location, while protecting the original book from damage caused by excessive use.

Though there is agreement on the usefulness of digitisation no consensus has yet emerged on the selection of works to be digitised, on technical standards, on how to make digital editions accessible and on the conditions of access.

The DIEPER project¹ aims at finding common ground for the digitisation of periodicals only. Why focus on periodicals? Periodicals a user is searching for are often unavailable at his local library. Searching for it, ILL and producing hard copy involve time and relatively high cost for both the user and the libraries involved. On the other hand reading or browsing a journal article on screen is perfectly possible with today's technology. In many cases it will be more acceptable to the user than reading the same on a microfilm reader. But perhaps the most attractive point to the user is his ability to access the document in virtually no time.

Beyond the user aspect there are other reasons for focusing on periodicals: To digitise whole runs of a journal involves high investment and it is absolutely necessary to avoid to digitise once more a journal that has been digitised elsewhere already. Libraries, who want to offer access to a virtual library of periodicals, will try to ensure easy and coherent retrieval in spite of constant changes in information technology.

DIEPER's principle aims may be summarised as follows. It will create a central access point for digitised periodicals. This shall enable anyone connected to the internet to search multiple sites at the same time and, ideally give him access to journals and articles without repeated searching.

a

A database, called the DIEPER Register shall record all periodicals that have been or will be digitised in Europe and beyond. This will allow checking whether a periodical is available already and to link up to the digital archive, where it is held. There browsing and reading will be possible depending on the accessibility of the chosen archive. At least more information on the archive's host and on the periodical may be found. The DIEPER Register is conceived to be the main instrument for bringing a digitisation project in line with parallel initiatives.

**DIEPER - providing web
access to retro-digitised
periodicals at multiple sites**

b

A search engine shall enable full-text searching across multiple archives. This is more than just retrieving periodicals by their title, as can be done with the help of the Register. The DIEPER Search Engine will allow searching tables of contents, authors and titles of articles, and keywords in articles. It will allow global and selective search. Restricting a search should be possible by type of document (articles and/or reviews and/or notes) and by a chosen subject area. All this is dependent on the electronic format used. Minimal standards have to be followed to enable this kind of retrieval.

c

DIEPER is to advocate the use of existing standards while avoiding proprietary solutions, which will not integrate with the global digital library. These standards and acceptable technology will continuously be documented on the DIEPER web site and a minimal standard for indexing digital archives will be defined². To give some examples of emerging standards that gained acceptance in major digitisation initiatives: Image scanning (e.g. minimum 600 dpi resolution; image saved in TIFF at lossless compression³; conversion to GIF, PNG or JPEG for web viewing); filing metadata (DC) and information on the electronic document in XML/RDF; saving the text file created by OCR in XML, tagged according to TEI; identifiers (DOI or URN) shall be unique for every electronic document enabling retrieval in the web without fixed URLs.

d

Last but not least DIEPER will try to support libraries by giving advice on copyright and licensing issues.

Where is the project standing at present?

DIEPER is not restricting itself to devising an infrastructure for the virtual library of digitised periodicals. Testing applications under real circumstances must demonstrate that the choices made are meeting our expectations. The participating libraries have selected six periodicals covering a time span from the 18th century to the present to test the chosen approach. The selected periodicals are all dealing with mathematics to a greater or lesser extent. Focusing on this subject area where a relevant mass of digital documents is available already and where researchers and students alike are familiar with using IT, shall enable relevant results when testing user response. While taking advantage of experience gained in other projects technical standards were reviewed and tested for image capturing, indexing, structuring, and document management.

A task that has been dealt with already is the DIEPER Register of Periodicals. A bibliographic format, compatible with USMARC and UNIMARC, has been defined to include features specific to the electronic edition. This covers information related to preservation and retention. For anyone involved in selecting periodicals for digitisation it is of

particular importance to give information on the general technical features of the original image file created when scanning and on how it is archived.

No derivative can be better in quality than the master file. To record such information DIEPER chose to follow the model given by USMARC #007 for computer files and by the related UNIMARC #135. This allows entering codes for the most essential kind of information (type of electronic resource; special material designation; colour; dimensions; sound; image bit depth; number of file formats; quality assurance target(s); antecedent or source; level of compression; reformatting quality)⁴. On this basis it is hoped that future decision-makers will be able to judge, whether the digitised version of a periodical recorded in the Register may meet their needs or not. To this DIEPER added the option to record retention related information along the lines of a proposal made by UK preservation administrators in 1997. There are four standardised ways of giving information about the intention to retain or not to retain the original and the digital surrogate⁵.

A web form on the DIEPER site allows inputting data of a periodical. This may cover the full range of a bibliographic record for a periodical or may be reduced to minimal information only. In any case the resulting record will be checked by DIEPER and be compared to full bibliographic data available from other sources. The web form asks for at least one Dewey

class number to be assigned to each periodical. If the user does not assign any, this will be done by DIEPER by using DDC class numbers usually of the second level of 99 classes.

Users of the Register will be able to select subject areas of their choice to limit their search. The same option shall be available when using the DIEPER Search Engine. The Dewey classification has been chosen for its wide distribution and use in the international library community. For an example of a record as displayed by the DIEPER Register see p. 30.

The Register was set up as an operational database in 1999. Records are loaded recording all digitised periodicals known to DIEPER. Any organisation having digitised a periodical or which is planning to do so, is encouraged to register it with DIEPER⁶. Only in reaching near completeness in recording digitised periodicals will we be able to avoid duplicating of effort. I regret to say that even at this early stage a few instances of duplication or overlapping have been observed.

What is the project going to do next?

Installing the DIEPER Search Engine will mark the next milestone. It can become operational only when indexing of digital archives according to the minimal standard mentioned above is possible. The partners will therefore enhance existing files and ensure output that will be available for searching.

**DIEPER - providing web
access to retro-digitised
periodicals at multiple sites**

At the same time co-operation with major digitisation programmes such as JSTOR, Gallica and Delta will be sought or intensified. This will be aiming at enabling the Search Engine to retrieve documents from those other sites at the same level as the one at which it is searching the partners' own sites. On another level co-operation with other programmes shall result in making available bibliographic records of periodicals that will be loaded into the Register. To mark the end of technical development within the project phase a demonstrator will be built to test the possibility of sustained use of the DIEPER infrastructure.

Since the partners began digitising they have been aware of the necessity to survey the rights issue. With the help of a legal expert the intellectual property rights (IPR) related to digitised periodicals will be examined. Model licences for access to digitised material will be drafted. In its attempt to ensure that agreements with rights holders will create user friendly conditions of access DIEPER will work closely with related projects⁷.

Preparations for DIEPER to function economically beyond the project phase (2001 ff.) will include the creation of a consortium formed by libraries and other organisations. This consortium shall explore the possibility of continuing to keep up the infrastructure and the service developed during the project phase and to find a model for financial sustainability.

It is hoped that DIEPER will result in a European network for easy and economic use of digitised journals. Enlarging its membership and exploiting the DIEPER infrastructure should lead to a self-supporting agency in the service of users and libraries.

For more information please visit the DIEPER web site or send your e-mail to

Werner Schwartz
dieper@mail.sub.uni-goettingen.de

References:

¹ DIEPER (Digitised European PERiodicals) is a project supported by the European Commission. It started in November 1998 and is to run for 26 months. The project partners are from nine countries: Niedersächsische Staats- und Universitätsbibliothek Göttingen (co-ordinator), Bibliothèque de l'Université Paris V, Centrale Bibliotheek/Katholieke Universitet Leuven, Helsinki University Library, Det Kongelige Bibliotek, København, Springer Verlag Heidelberg, Tartu University Library, Università di Siena - Facoltà di Ingegneria, Universitätsbibliothek Graz, University of Patras Central Library.

² *www.sub.uni-goettingen.de/gdz/dieper*

³ This is the format in which the image file will be preserved for archiving and ultimate migration.

⁴For online documentation of
USMARC codes go to

*lcweb.loc.gov/marc/bibliographic/ecbd007s.html
#mrcb007c.*

⁵A periodical may
be retained indefinitely in original and
possibly in surrogate,
be retained indefinitely in original
format only,
be retained indefinitely in surrogate
format only, or not be retained
beyond immediate use.

⁶Go to

*www.sub.uni-goettingen.de/gdz/dieper/
dpsearch.html*

⁷Most important of these is the
TECUP project:

www.sub.uni-goettingen.de/gdz/tecup

**DIEPER - providing web
access to retro-digitised
periodicals at multiple sites**

*An example of a record as displayed by the DIEPER
Register:*

Periodical: Mathematische Zeitschrift

Published: Berlin ; Heidelberg : Springer, 1918-

Numbering: 1.1918 - 50.1944,1; 51.1947/49, Aug. -

Note: Index 1/25.1918/26 in: 25.1926

ISSN: 1432-1823

Subject: 510

Reproduction: [Electronic ed]

Published: Göttingen: Niedersächsische Staats- und
Universitätsbibliothek, 1999

Series: Mathematica

Note: File of mixed content on optical disc; one
colour/bitonal; 4 3/4 inch or 12 centimetres; image
bit depth is 001; one file format; reproduced from
original; lossless compression; reformatted for
access

Retrievability is medium; Accessibility: in the internet;
Availability: free of charge; Will be retained indefinitely
in original and possibly in surrogate

Shelfmark of original: SUB Göttingen <8 Z NAT 527>

Digitised: 1.1918 - 52.1950; Digitisation planned: 53.1951-

Shelfmark of reproduction: SUB Göttingen <CA 99/127:1-54>

Full text: [Online location](#)

Digitising journals: Highlights from JSTOR's experience

*Eileen Fenton, Kevin Guthrie, Amy Kirchhoff
JSTOR, USA*

Introduction

In the nearly five years since it was established as an independent not-for-profit organization, JSTOR has built a database system comprised of the complete backfiles of approximately 120 journals, many reaching back well into the 19th century. The system offers access to nearly 800,000 items (400,000 of which are full length articles) and approximately 5 million pages from journals in 15 academic disciplines. Reaction to this database from around the world has been enormously positive; more than 700 academic institutions have signed up to support this collaborative organization dedicated to providing a trusted electronic archive of core scholarly journal literature.

The purpose of this paper is to outline some of the lessons learned through JSTOR's experience, with a specific emphasis on the production and technological tools and processes that have been employed in creating the resource. As librarians and technologists from other countries contemplate digitisation projects of various kinds, we hope to offer JSTOR's experience as a baseline from which others can build.

Brief Background and History

JSTOR began its life as a project of The Andrew W. Mellon Foundation. The initial purpose was to investigate

whether digital technologies could be employed to reduce the costs of storing and maintaining long runs of printed journal literature housed on library shelves. At the same time, it was hoped that through digitisation the convenience of access to these same materials could be enhanced.

The Foundation launched a pilot project that included five journals each in history and economics. These ten titles, as they were digitised, were made available to a group of approximately 20 test site libraries. These journals were converted as high quality page images to insure that the database offered a faithful and trusted replication of the original published material. In addition, Optical Character Recognition (OCR) software was used to create text files that could be used to facilitate full text searching of the material. Although the database was not complete, and did not have anything approaching a critical mass of content to attract scholars, it was evident that the idea held great promise. In the summer of 1995, JSTOR was established as an independent not-for-profit organization with its own Board of Trustees. (For a list of present and past JSTOR Board members¹).

The Mellon Foundation provided funds to launch the enterprise, but from the outset JSTOR was charged with developing a self-sustaining economic model. The new organization defined

the first phase of its work to be a database that would house a minimum of 100 important journal titles in 10-15 academic disciplines. In the fall of 1997, a fee structure and levels were established for what came to be known as the Phase I database, and it was promised that JSTOR would complete its Phase I work prior to the end of 1999.

A special charter incentive program was established, and by April 1997 JSTOR had received signed commitments from approximately 200 academic and research institutions. It was an excellent start, and growth and the positive reactions from the community have continued through JSTOR's brief history. As of May 2000, over 700 institutions have signed up, representing every kind of academic institution from the smallest of colleges to the largest of research universities in 31 countries around the world.

One of the surprising aspects of our experience has been the extensive use scholars and students have made of the older published journal literature. These materials, which were used relatively infrequently in paper formats because material in them was difficult to locate, are getting extensive use in the digital medium. In 1999, approximately 1.4 million articles were printed and more than 4 million searches were performed. Usage has been growing and continues to more than double annually, and we expect that well over 2.5 million articles will be printed from JSTOR in 2000.

Production

Building a resource of this scale and magnitude requires meticulous attention to detail from the production point of view. As the JSTOR production staff and processes have developed, we have learned that the basics of production are deceptively easy to describe and surprisingly complex to execute. Through hard experience and trial and error we have learned that challenges in production fall into four broad areas: acquiring source material, assessing source material, creating guidelines for digitisation, and monitoring and controlling the quality of the digital product.

Acquiring Source Material

As with so many aspects of the production process, acquiring source material is more complex than it might first appear. Before JSTOR can acquire the full back run of a journal we first must determine what a journal title encompasses. There is an almost endless array of variations that a journal may take over the course of its publication history. We first research this history using a variety of bibliographic sources in order to determine if there are former titles, supplements, absorbed titles, superseded titles, associated monographic series, or special issues that will need to be acquired as part of the full back run. Using a separate database developed by JSTOR staff, we capture in a single location the results of our research and have documented, for instance, the relationship which the Journal of the Royal Anthropological Institute

has with three other serial titles. Only when we understand what a journal title includes can we be certain of what we must acquire.

When identifying potential sources for a journal, we always turn first to the current publisher, but unfortunately it is the rare publisher who is able to supply a full and complete back run. When this source is exhausted we focus our efforts on two other possible sources, vendors and libraries, both of which can be problematic in different ways. Purchasing back issues from vendors can be costly and we have seen the cost of a single volume sometimes exceed \$100. Obviously economic pressures preclude purchasing large quantities at such rates. Many have suggested that JSTOR access the needed back runs through libraries and several libraries have indeed been willing and able to help us by supplying material. However, as libraries generally do not prefer to make extended loans, have their materials disbound, or give large numbers of volumes away, many libraries are limited in their ability to serve as a source.

The challenges do not end with the acquisition of the physical volumes. As inventory is received - from publishers, vendors, or libraries - detailed records must be maintained about what is received, from whom, and what obligation must be met at the end of digitisation. Again JSTOR staff have developed a special database to facilitate inventory tracking tasks. At JSTOR all material is disbound prior to scanning,

however, some institutions that loan materials to us require that the volumes be rebound prior to their return; other institutions do not impose this requirement. Careful records must be maintained to insure that these types of obligations are faithfully met. Detailed records are especially important as material is sometimes received in duplicate. For instance, a publisher may donate three of the four issues of a volume, but a library who is willing to fill this gap by loaning material will likely loan a complete volume and may require rebinding prior to return of the volume. Without careful records it is not possible to be sure that inventories are complete and correctly processed. All of these elements combine to make acquiring source material a significant and time consuming challenge.

Assessing Source Material

When a complete back run has been assembled, assessing the source material can begin. All source material must be examined closely in order to identify any potential scanning problems or special needs present. Pages may contain printing errors or be torn, missing, or marked by library stamps, user annotations, or insects. Issues may contain special illustrations, finely detailed and/or oversize maps, colored text on colored backgrounds, multi-piece graphs presented through overlay sheets, oversize foldout pages, microform inserts, or other elements that present special scanning needs. JSTOR identifies these pages by completing a page by page review of each and

**Digitising journals:
Highlights from
JSTOR's experience**

every issue to be scanned. We have learned through experience the importance of having and following established standards for identifying problematic pages, guiding the scanning vendors' work, and then verifying that the final product has been created to specification. In addition to noting potential scanning problems, JSTOR staff record in journal specific databases a range of metadata which is used by the vendor in order to ensure that elements such as copyright statements, abstracts, pagination variations, and illustrations are handled correctly. Though the capture of this metadata is costly, it plays a key role in the overall quality of the final product.

Creating Conversion Guidelines

While metadata on the physical source material is important, so too are metadata regarding the intellectual content of the journal. Creating conversion guidelines that enable the scanning vendor to correctly capture this metadata is also quite complex. Because it aims to provide a faithful replication of the printed source, JSTOR is very concerned with accurately and fully replicating various aspects of how a reader experiences the original journal issues. We want to provide the online reader with the same ability to rapidly assess the full issue as the reader of the print original. A reader of the original print document might glance at a table of contents - which may or may not be an accurate reflection of the contents of the issue - and then quickly flip through an issue in order to flesh out his or her assessment of what may be in the journal.

In order to replicate this online, JSTOR expends significant time and energy to ensure that the contextual cues present in the original are reproduced in the digital version as well. We work with our vendor to be sure that the intellectual structure of the original issue is faithfully replicated. If, for instance, a math journal contains a special section devoted to problems and solutions, we will indicate this in the online table of contents - regardless of whether this information appears in the printed table of contents - so that the online reader may quickly assess the structure of the journal issue. We believe that by respecting the structure of the original issue we are able to convey to the reader important information about the context of the article that is being viewed. Capturing this intellectual structure is significantly more complex than simply requiring the vendor to key in the author, title, and page range of each article. In order to assure the quality and consistency of the final product the vendor must be guided in how to identify an article, a title, an author and what to do when these elements are presented in unusual ways. JSTOR's production librarians provide such guidance for each title in the database, doing so at the issue level when a journal's structure warrants it.

Monitoring and Controlling Quality

At JSTOR all of the preparation described above precedes any scanning of the source material. Through experience we have learned that extensive preparatory work is essential

to creating a high quality digital resource. Once conversion is underway and digital material is being received from the scanning vendor, then monitoring and controlling quality become key issues. For this JSTOR relies on automated data verification wherever possible, and elements such as image resolution and metadata formatting lend themselves to this type of check. However, a number of other elements must be manually verified.

Using a tool developed by JSTOR's technical staff, production staff review a statistically valid sample of the image and text files produced by our scanning vendor. The sampling tool, which displays a page image and its corresponding text file side by side, allows our production staff to assess whether the sample images meet established standards for image quality and whether the OCR in the text files for each journal page have been corrected to the required accuracy. Additionally, and using a separate tool also created for the production group by JSTOR programming staff, we verify that the conversion guidelines established by JSTOR's librarians for that journal were followed. Any errors found during this review are corrected by either the vendor or by JSTOR staff.

Production Principles

As JSTOR's production processes have developed over time, four principles of production have become clear. First, form follows function. The shape of the production processes for any given project must be driven by the functionality that is sought for

the project. If, for instance, article title or author name searching is important to the success of the project, this will - or should - have a direct impact on the processes that are put in place to ensure this capability at the appropriate quality level.

Second, scale matters - a lot. As JSTOR has grown we have seen first hand that a project digitising 100,000 pages over the course of a year requires very different management, tools, and staffing than an ongoing enterprise producing 200,000 pages a month. If a digitisation project is expected to grow, it is very helpful to begin the project with production processes designed with scalability in mind.

Third, having the right tool for the job is important. Digitisation projects require tools to facilitate both data management and data verification. Depending on the functionality that a digitisation project seeks to encompass, the necessary tools may not be available as off-the-shelf products. If not, significant development time to create them will likely be needed.

Finally, through trial and error JSTOR has learned that in order to present content well it is important to know what is in it. Decisions regarding the preparation, digitisation and display of the data have to be driven by the nature of the content. Developing an in-depth understanding of the details of the content early in the life of a project will save significant time in the long run.

Technology Challenges and Tools

Just as an array of challenges unique to production drove the development of certain processes and tools, JSTOR's technology base grew according to a very similar path. At first glance, with production it seemed that the journals could simply be boxed up and shipped to the vendors and everything would be automatic - JSTOR quickly learned that this was far from so! In a similar manner, it seemed a simple thing to store images on-line, match these images up with their text files, index those text files and provide access to this system to a number of educational institutions. In this paper we will discuss several areas where we had to develop technological tools to address developments we did not anticipate. These areas include: enabling replication of the database to multiple mirror sites, providing authentication mechanisms to allow multiple methods for accessing the database, offering statistical reporting capabilities, and finally, developing new data structures and code to support more extensive interoperability of the database with databases held by other organizations.

Replication

JSTOR aims to provide a trusted and reliable archive of the journal data within its collection. In pursuit of this objective, JSTOR needed to ensure that, even should something catastrophic happen to one of its servers, the collection would remain protected and available to users. To that end, we currently have four JSTOR servers running at three geographical-

ly dispersed locations. These mirror sites hold separate and duplicate copies of all of the original digitised information. The three U.S. servers are interchangeable and work collaboratively to respond to user activity through the main JSTOR address. In addition, we generate three archive copies of all our data at the point at which it is released onto the public servers. Each of these three copies of archive tapes is shipped to a separate location (such as to our New York office) where is stored far from the servers providing active access.

The above seems a relatively simple approach, however once we began running two servers in tandem, it became clear that there were real challenges associated with insuring that the servers would always be identical, in both data and code. One challenge is imposed by JSTOR's use of frames in the user interface. Due to the stateless nature of the web, it is not unusual for users who are retrieving the framesets of a page to retrieve each frame from a different JSTOR server. Therefore, the computer that serves up the navigation bar must agree with the computer that serves up the article page. Our systems needed to be devised to take this into account.

In addition, although the web is stateless, the JSTOR Production unit is not. The Production Unit makes daily updates and corrections to the data on the public servers. Therefore, the technology unit has developed an extensive infrastructure both to propagate those changes out to all

servers every night, and also to archive those changes. Should we ever need to rebuild the database from tape, we can use the archive tapes and the tapes of the changes made to the database since the original archive tapes were made to rebuild the most recent "state" of the archive.

In addition to the daily data updates, the code that provides access to the data also has modifications made to it at least once a month. As with the data modifications, it is necessary that the code changes are pushed out to the JSTOR servers nightly to insure that all of the servers work seamlessly together and work effectively for end users.

Authentication

In the first year of its existence, it was acceptable only to offer IP-based access to institutional participants through the use of the default authentication system that is provided with web servers. IP-based authentication has emerged as the bread and butter method of authentication for on-line resources in the academic environment. Unfortunately, there is very little flexibility in the default authentication methods of web servers. In addition, IP-based authentication is insufficient in many contexts: it does not work for institutions without stable IP addresses; it does not provide for an easy method of remote access to those institutional users who are not accessing resources from fixed IP-addresses on the campus network; and it does not allow for any new types of access models (for example, individual users).

JSTOR quickly outgrew the default authentication options available on web servers. We developed and now maintain a database-driven model where all authentication information is stored in a relational database. This move offers us important flexibility; for example, we can now work with publishers to provide username/password access to individuals to select journals within the collection. In addition, JSTOR has developed tools to aid institutions in providing remote access to the JSTOR database for their users who are not located at an IP authenticated computer. JSTOR is also very involved with the U.S. Digital Library Federation building and testing the use of certificate-based access to the database. We hope to demonstrate this technology later this year.

The development of an authentication database system has given us the flexibility needed to respond to the needs of publishers, libraries and users. But even the database system itself has needed to change to keep up with evolving technologies. We recently migrated the authentication database from mSQL to Oracle. In addition to providing us with increased flexibility for the future, this step also allows us to push out authentication database updates to all four public servers within 10 minutes of any modifications being made.

Statistics

JSTOR has developed a statistics utility that is used by JSTOR publishers,

**Digitising journals:
Highlights from
JSTOR's experience**

JSTOR participating institutions, and JSTOR staff to track usage of the JSTOR collection. Here again, statistics that could be generated through the use of standard tools and analysis of web server access logs was neither sufficient nor appropriate for our purposes. For example, it is not meaningful to count server hits if some of those hits are generated by displaying logos or accessing navigation buttons. We care about counting things like the number of times an article has been viewed, or the number of articles that have been viewed. Consequently, JSTOR has invested substantial resources to develop a software tool that can analyze our access logs for specific types of accesses.

With print it has always been difficult, if not impossible, for librarians to track how information they have purchased is being used. Electronic resources, however, offer an opportunity for librarians to have much more direct feedback about these decisions, and they are keen to get that feedback. JSTOR's statistics utility allows librarians to view usage statistics by journal title, by time of day, by day of month or month to month, and will graph any of the usage numbers requested. The utility is available 24 hours per day and 7 days per week and can be used to view the institution's usage at any point since the institution became a JSTOR member up to the previous day. For comparison purposes, the statistics package also reports an institution's usage against the "average" usage for libraries within

their classification. The JSTOR statistics package has become a very important resource at JSTOR libraries and has been a model for others. (The specifications for the statistics utility were developed by a task force of JSTOR libraries. The recommendations of that task force were later adopted (with some modifications) by the ICOLC as an international model for the kinds of usage statistics content providers should offer.)

A similar statistics utility is provided to the JSTOR participating publishers. As with the library statistics package, data is updated nightly and information can be broken down for analysis in a variety of ways, such as by usage over the course of a specific year or during a particular month. The system allows publishers to compare usage of their journals to the discipline more broadly and to the JSTOR archive overall, leading to a clearer understanding of the relative value of their older articles to scholars. In addition, information is provided about the most accessed issues and articles for each title. These data help publishers better understand the most valuable articles in their journal runs, and provide insight into possible ways to repackage their electronic material for focused audiences.

We also use the statistics package internally to monitor overall activity in the database. These data are used for a variety of purposes, perhaps the most important of which is to monitor and predict future load on our servers,

information that is vital to guide hardware purchase decisions so that we have in place adequate hardware infrastructure to meet user demand. The JSTOR statistics package has become an irreplaceable and invaluable tool not only internally, but also for our constituent publishers and libraries.

Interoperability

In JSTOR's early days, there was no choice but to develop our own standards for retro-digitisation. There were no models or standards to follow. That said, it was obvious to us from the beginning that we had to set rigorous internal standards and follow them. We constantly evaluate these standards and assess their usefulness against our broad goals to be sure they are meeting our (and the broader scholarly community's) needs.

For example, we originally managed to fit our relational data into a flat file system. It is evident that we must now migrate that approach to a true relational database structure. We also took an established metadata specification (the specification originally developed for the TULIP project) and modified it to suit our specific purposes. There again, in the interest of interoperability, we have developed an XML DTD and currently have the ability to take our metadata structured to our internal specification and turn it into XML. From the XML, we can easily port it to a variety of formats.

Over the past year, JSTOR has increasingly been asked to "share" its

metadata with other institutions, and the need for industry-wide standards has become more important. We are working to co-operate fully and integrate our activities with these developments. Whereas it was difficult to anticipate the need to develop a detailed statistical utility or to prepare for the complexities associated with authentication, as an archive we have always expected to migrate data from one format to another. This is an important part of our planning processes and outlook on a day-to-day basis.

Through our initial experiences, perhaps the most important lesson learned by the technology services unit is that producing high quality data and developing reliable internal standards are of utmost importance. This is especially true of metadata. If rigorous controls are in place, and the data has been entered with appropriate care, the format the data is stored in matters less than the fact that it has been stored in a strictly defined structure. That structure will make it possible to generate software to port the data into new formats when that becomes necessary.

Archiving

JSTOR has developed considerable experience in a wide variety of areas related to the production and delivery of a large-scale database comprised of the back issues of academic journals. In addition, as a project aimed at providing a trusted long-term archive of digital information, JSTOR has been addressing many of the issues revolving around electronic archiving.

**Digitising journals:
Highlights from
JSTOR's experience**

The problems associated with electronic archiving are not simple and have been the subject of considerable discussion recently in the academic community. It is JSTOR's position that there are no technological solutions to this problem. To put it another way, there is no software that can be developed that will allow us to put data in a black box and know they will be accessible in the future. Data needs to be used and refreshed in order to insure that it remains accessible as the technologies for interacting with it evolve. Consequently, the most important question to be resolved when thinking about whether a particular digital item or collection is going to be archived is the entity responsible for its care. Is the organization making a promise about future availability dedicated to that objective and is that objective consistent with its mission? The reason many items in paper remain available to us is not really a technical one, rather, it is more a function of the fact that the libraries that hold them are dedicated to preserving them and maintaining them for future use. The same will hold true for digital collections.

For organizations interested in providing long-term access to electronic information, it is our experience that there are five primary areas that need to be addressed. First, one must make some technological choices about how the data is digitised and stored. This may involve scanning resolution or it may deal with storage formats or database structure. In making these choices, archiving

organizations must remain open to change in order to keep pace with the dynamic technological environment. Our steps to migrate JSTOR software and data structures presented above illustrate the point. Second, one must take special steps to insure the preservation of the data in the event of a catastrophe. JSTOR's multiple mirror sites in three distinct geographical locations in two countries demonstrate an example of this approach. Such redundancy insures that bad luck cannot destroy the entire database. Third, archives must negotiate special relationships with content providers to be sure that they have rights that correspond to the long-term nature of the promise to provide access. JSTOR has special provisions in its publisher license agreement, such as the "moving wall"², to address this issue. Fourth, if they are to be successful, digital archives must establish a special trust relationship with libraries and offer to provide the data to selected or all libraries in the event that the repository is no longer able to care for the electronic collection. Fifth, and finally, organizations caring for archival resources in digital format must establish a realistic economic plan for paying the costs of migration and maintenance over the long run. Just as a library must fight for the resources to build a new set of shelves or renovate the stacks, so must digital archives generate the resources needed to insure that their information remains conveniently accessible.

Conclusion

This paper has provided an overview of lessons JSTOR has learned during its relatively brief existence. Areas covered include the development of production procedures and tools for managing the digitisation process, the creation of specialized software to provide convenient access to these data, and five areas of activity that should be addressed by entities engaged in electronic archiving. It is the hope of the authors that sharing this experience will prove helpful to other projects as they contemplate the conversion of previously published literature.

Kevin Guthrie
kg@jstor.org

References

¹ www.jstor.org/about/board.html

² www.jstor.org/journals/movingwall.html

Standards for images and full text

*Hamid Mehrabi and Henrik Laursen
The Royal Library, Copenhagen, Denmark*

This paper only deals with two aspects of the total digitising process. These two aspects are

- 1 standards for the images and
- 2 standards for the mark-up of the full text.

But they are important because the amount of care invested in those two partial processes will be of considerable benefit to the whole digitisation process.

To clarify our concepts: when literature can be read online it is called full text both when it is presented as images and as ASCII text. In this paper we will call an image an image and the term full text is reserved for the ASCII text version.

A delimiter: we are talking about retro-digitisation, i.e. digitising of older paperbound materials and we are talking about the digitising of texts. So what we have to say is not relevant for texts which originated as electronic files. Neither is it applicable for the digitising of pictures, illustrations, photos, sound or living picture films, which are objects that do not mainly consist of texts.

The image capturing process:

We will start with an example. We show you an image of a page from

a 19th century Danish scientific journal (year 1881). It is a bad image! The next image shows the very same page. It is a good image. What you see are the images prepared for presentation on the web. They are almost equally legible. But if you look at the image information box you will see one remarkable difference: the size of the bad image is 60K, and the size of the good image is 40K. That is to say - the bad image is about 50% bigger.

Even if you look at enlarged versions of the images there will be no significant difference, at least to the human eye. But to prove that there is a difference, we will show you the OCR'ed output from the two images.

If you count the OCR-errors you will see, that on the output from the good image the errors are considerably reduced - to about 15% compared to the output from the bad image.

We want to stress that what you see is the result of unsophisticated OCR. That is, no training, only language selection. We have only applied the French respectively the Danish dictionary that comes with the FineReader Software. In FineReader you can add words to the dictionary and even add supplementary dictionaries in simple text format. So if you have proof-read 100 pages of some scientific journal on botany you can reuse the result as a input to the OCR program. And we have not

Standards for images and full text

made a search/replace process for the most common misreadings.

Still the difference is evident. But I think the most important difference is the difference between proof-reading and no proof-reading. You can all imagine the costs saved when you can avoid proof-reading.

With the level of misreadings in the good result I think it is acceptable to offer full text searching in the non proof-read text. Whether or not you give access to the ASCII text itself is, we think, a matter of taste.

Our personal attitude is that if the user can benefit from it, then let him have it - of course with informative reservations about the quality of the text. Another approach would be to give access to the master image file and let the user have the trouble with the OCR. You cannot use web images as a basis for OCR - the resolution is not high enough.

Both images follow the standard that is actually widely accepted. The image resolution is 600 dpi (or 240 lines per centimetre) and the colour resolution is halftone or 1 bit. This is the standard determined for the DIEPER project. Both images are captured with a Minolta PS 7000 book-scanner using the standard software packet that is supplied by Minolta. The difference is that the bad image is captured with the standard or default settings for exposure light intensity. The good image has been treated - different exposure values

have been checked. Both images have undergone further quality enhancing processing with an image processing application. All these processes can be run as a batch overnight.

The most efficient process for a good OCR result is to apply a Gaussian filter with an appropriate parameter. You can see it as a kind of intensifying of faint lines and removing of speckles.

It is not evident to the human eye which image is the optimal one. But the result of many experiments has proved which resulting image is the best source for the OCR. The whole process must be carried through from start to finish in order to decide which exposure settings and which image enhancing processes are the best ones.

Of course the whole process does not have to be done for every single page. A sample of 10 pages or so from each periodical - as long as the layout and printing quality are the same - will do. Intensive efforts on 10 pages will produce considerable benefits for the remaining thousands of pages!

Conclusion: the standard 600 dpi and 1 bit is only a good starting point. 600 dpi is affordable today with the costs of electronic storage - and bookscanners can scan this quality as quickly as you can turn the pages and press the exposure button. For the Minolta PS7000 as for other bookscanners it means that you can easily scan 200-300 pages per hour. For many texts 600 dpi will give a better OCR result and your web

images can be smaller. For many well printed books 400 dpi and even 300 dpi would give the same result and would also be good enough for the presentation on the web.

But it is only a good starting point - not a sufficient description. And what a sufficient description is has not yet - to our knowledge - been described in theory. It would be interesting and fruitful to have some research on images from which a numeric description of image quality could be derived. For lack of better we must be satisfied with a pragmatic description.

These recommendations were probably mostly of benefit for those of you who are digitising in-house. If you intend to outsource the digitising of your materials you should write specifications that are equivalent to the standards mentioned: master files from which you can derive text images that are small (small file size and still readable on a screen) and from which the OCR output is good.

The mark-up of texts

The high standard that was set for images was mainly based on a cost/benefit analysis of the process costs. This is not the case with the next set of arguments. When it comes to the creation of full texts it will look like the balance sheet you know from your summer vacations: no income, twice the expenses. But still, if you can obtain fine OCR results, the creation of full texts can be done semi-automatically.

The argument for standards for full text is based mainly on a usability point of view. Once you have created a full text marked up according to some standard, you have the basis for a widely differentiated use of the text.

You all know the HTML mark-up for web pages. The first lines of a journal article could look like this: the mark-up is mainly for presentation on the screen. Rich on information for the presentation but poor on structural information. A mark-up with more information on the content could look like this. This mark-up follows the XML-standard. Both texts are ASCII texts or flat files. It can be read by almost any piece of software in the world and allow for recycling or cut and paste. Here is its strength in contrast to the popular PDF Adobe Acrobat format.

Simple scripts in Perl or a simple set of search/replace rules in your favourite text editor can alter the fully marked-up text to be read as you like it. Of course you can also use one of the expensive tools developed for the SGML/XML world.

You can benefit from it in your internal processes

The master file in SGML or XML format can easily be converted to HTML. Or it can be converted to the new e-book formats for downloading by the user. If e-text books become as popular as the producers hope and the stockholders dream of it will soon be

Standards for images and full text

a reasonable service for the libraries to offer.

When creating indices for search machines you can retrieve the data from the master files. We are experimenting with that. Instead of hand typing information about the bibliographical records, you could automatically take this information from the full text file if the layout of the text is consistent to allow this sort of extracting information.

The user that finds an e-text book in your library has of course immediate benefit from your efforts to convert the full text to this format. The same is true for the person who just reads the same text on the web. But if you make the master file available, i.e. the full text with the original mark-up, a user possessing an advanced SGML/XML-aware application can benefit enormously from the effort invested by you in his or her further work with the text.

For the DIEPER project we have chosen to apply the widely used international standard TEI. TEI stands for the Text encoding initiative. The TEI is an international project to develop guidelines for the encoding of textual material in electronic form for research purposes. It was developed in the SGML era for the literary and linguistic mark-up of non-fiction texts. It has proven its vitality not only by the many full text archives that employ the standard but also from the fact that a XML version of the TEI.DTD has existed for a long time. Now we are only waiting

for the new browser generation to become XML-aware. So you can present your XML marked-up text directly.

Now to our conclusions:

1. For your own good:

Take some good pictures!

2. For the user's good:

Use a standard (generalised) mark-up language.

Thank you for your attention!

For technical details,
you are welcome to contact us:

Henrik Laursen
hhl@kb.dk

Hamid Mehrabi
hrm@kb.dk

Digitisation

- technical issues: Production at the Göttingen Digitisation Center (GDZ)

Norbert Lossau
Göttingen State and University Library, Germany

*The article is a comprehensive summary of the presentation at the conference *Digitising Journals: Conference on future strategies for European libraries**

Background information

The Göttingen Digitisation Centre (GDZ) was established in 1997 and is acting - besides the Bavarian State Library in Munich - as national supply Centre for German libraries and academic institutions in the field of digitisation. Funded by the Deutsche Forschungsgemeinschaft (DFG) the GDZ is charged with co-ordinating national efforts towards standardisation and is engaged in testing and providing tools and techniques for image capture and text conversion, bibliographic description, document management, and the provision of online access to digital collections.

The focus of the activities of the GDZ has been and is still on the different fields of technology required to build a digital library. It is equipped with a strong technical infrastructure, financed by the Lower Saxony State Ministry of Science and Culture.

Current Digitisation Projects

The work of the GDZ is based on intensive practical experiences in a number of digitisation projects for the Göttingen State and University Library. A number of collections have been di-

gitised with significant activities in the fields of historical travel literature and North Americana¹, as well as in mathematics. The Jahrbuch-project, building up an Electronic Research Archive for Mathematics (ERAM)², is a joint effort of Göttingen and the Department on Mathematics at Berlin University (Prof. Wegener).

The digitisation of the invaluable vellum copy of the Göttingen Gutenberg Bible is the most recent effort in bringing selected works of cultural heritage to a broad public³

In the European DIEPER (Digitised European PERiodicals) project⁴, a joint effort of eight European Libraries, the GDZ is co-ordinating the technical realisation. DIEPER is aimed to test decentralised scanning-production and unified access over local repositories. Another goal of DIEPER is the establishment of a European database for digitised documents, which, like the EROMM (European Register of Microform Masters), will serve as a central reference to avoid duplicate digital conversion; the database will be located at Göttingen.

The Digital Conversion Process at the GDZ

*Image Capture*⁵

The GDZ captures images from text material in 600 dpi, 1 bit as TIFF ITU Group 4 format. Illustrations can be

captured with up to 8 bit grayscale, high quality colour digitisation up to 36 bit colour depth (both 300-400 dpi, uncompressed TIFF) is possible. All these files are stored as digital masters, derivatives for the Web presentation are generated on-the-fly (TIF2GIF) or in batch mode (TIF to GIF, JPEG, ...).

The digital conversion process works in-house as well as in co-operation with external vendors. Scanning from microfilm and keyboarding of full-text is done offshore.

Joint development of

Scan-software for face-up scanners

Starting the digital conversion process from paper in-house a production software solution was required to drive the two face-up scanners at the GDZ (at this time - 1997 - a Zeutschel Omniscan 3000 and a Minolta PS 3000). The company Satz-Rechen-Zentrum in Berlin was placed with the development of the new program. The first version of "SRZ ProScan Book" was available in late 1997. Today the GDZ uses an continuously enhanced version of this program for the new Zeutschel Omniscan 7000 and the Minolta PS 7000. It meets special production requirements for older books and its features cover e.g. editing of the TIFF-header, production control window with tree view over scanned pages, masking and cropping of pages during the scanning and some image enhancement features like deskewing and despeckling.

Metadata

Metadata play a very important role in the whole conversion process.

Starting digitisation activities in 1997 the main problem was the digitisation of early German text books. Fraktur type was not able to convert automatically with OCR programs. The compromise between presenting only images to the user was the detailed description of the documents' internal structure (chapter, subchapter, figure) as it was covered by the Table of Content, Index, List of illustrations in the print original.

The different types of metadata (bibliographic, structural) are captured in different ways. Recording of the bibliographic description happens in the Union Library Network Catalog (PICA/GBV). MS-Excel is used to describe the structure and pagination of a document. A Java-applet, written at the GDZ is used to convert bibliographic data into the RDF/XML format, another script (written in Visual Basic) converts the proprietary Excel format into RDF/XML. At the end of the conversion process all metadata is merged with another script to the final RDF/XML file and can be imported into the Agora DMS.

The European project DIEPER now addresses the need to retrieve full text, too. The discussion shows, that the combination of RDF/XML for all kind of metadata and TEI/XML as broadly used format for full text will be the way to go for DIEPER. The implementation of identifiers (URN / SICI) is essential to retrieve single items for external access and for referencing between metadata and full text files. A DIEPER e-DOC format,

describing these components, will be publicly available during May 2000 on the Web site of DIEPER.

Joint development of a new Document Management System (DMS), Agora

Following the recommendations of the DFG technical task force to provide a Document Management System (DMS) as a key component for the digital library, the GDZ chose a strategy of collaboration with an industrial software partner (Satz -Rechen-Zentrum Berlin). The main requirements of the task force were to use open, standard formats to ensure a high degree of scalability and interoperability of the prospective digital collections: To this end, the GDZ worked with a database driven system for data import and export, and for handling highly structured documents and metadata. A prototype of Agora, the new DMS, was presented in Göttingen in April 1999 and is now in production at the GDZ⁶. There are currently five Agora installations in Germany and there is an increasing interest outside Germany in the system.

The Agora system is a RDB (Relational database) driven EDMS based on an extensible metadata model⁷. The model can be implemented on different RDB platforms (e.g., Oracle, DB2, and Sybase). An administrative tool (AdminTool), running on Win95/Windows NT, controls all functions. In order to allow for maximum interoperability with other metadata sources, the system works with an import/export format that is based on RDF/ XML. The Java servlet of Agora acts as interface between the RDB,

web server and browser. The communication with the RDB is made through JDBC. HTML templates for the user interface are used to flexibly achieve different views; they are associated with collections through the AdminTool. Based on the structured information in the underlying RDB, elaborate search functionality can be offered to the user.

Agora developers recently integrated the Verity Information Server, a powerful full text search engine, used in a number of significant digital library efforts (e.g. Highwire, Stanford University). The administrator can now offer to users the search capabilities of both the RDB and Verity, making possible not only traditional SQL-queries, but also the range of search functionality that is part of Verity (e.g., fuzzy search and ranking). The inclusion of Verity now makes it possible to offer effective searching, in metadata such as bibliographic fields, titles of chapters, articles, and figures. Later, Göttingen will offer full text searching as well, a feature that becomes increasingly important as Göttingen moves from digitisation of older text material (often in Fraktur type) to 20th century works. Verity is able to search a wide range of document formats from MS-Word over PDF to XML files.

Agora's flexible export functions contribute significantly to interoperability. From its inception, Agora was able to export all data in the RDF/XML format. A recent addition, especially promising for users, was the ability to export PDF files with integrated Bookmarks, created automatically from the structural metadata in the database.

Because of the modularity of Agora, the GDZ has been able to add itself external features to Agora as demonstrated by the successful integration of the on-the-fly conversion of images with the Tif2Gif program, developed at the University of Michigan [TIF2GIF, 1997]. In a related (import) effort, the GDZ has recently developed a tool to convert bibliographic data (Pica/GBV) into Agora's RDF/XML format, and the tool is freely available as Java-Applet from the GDZ's web site⁸.

Online Access

The architecture of the Agora system allows for different ways of access to the digital collections. It is designed as middleware for an Online Library Catalogue as well as a stand alone Document Server for a Digital Library.

Online Union Library Catalogue

The Document Server of the GDZ is via http://protocoll connected to the Union Library Catalogue (PICA/GBV)⁹.

You can start a subject-based search in this catalog and your results represent all kinds of physical representations for a document (print, microform, digital). Choosing the record "electronic edition" the URL leads you directly to the digitised document, administrated by Agora. The URL is designed as cgi-script which ensures the longevity of the address, avoiding manual corrections in the catalogue which may be required when you use physical URLs for single files on a Web server.

Document Server of the SUB Göttingen/GDZ

The Java-servlet of Agora offers a "Simple" and "Advanced Search" mode. The Advanced mode enables an elaborated combination search of traditional bibliographic categories (author, title, place of publication, year of publication) with document types (journal, monograph, multi-volume work) and document structures (article, chapter, figures). You can search a single or multiple collections. The administrator has the option to make all categories of the database accessible to the user via the servlet.

Truncation is possible as well as searching with Boolean operators. The search via the Verity Information Server enables ranking of a hitlist. Browsing collections allows for a first orientation in the Digital Library.

Documents can be navigated via electronic Table of Contents, single pages can be directly addressed by a "Go to"-button.

Full text search by highlighting the search term in the electronic facsimile and navigation via an electronic index (generated from the print original) will be included in Agora this year (2000). The full text search will be based on the powerful retrieval mechanisms of the Verity Information Server, today used for searching metadata.

To date about 1000 documents with more than 350.000 images are available via the Document Server. The collections are continuously growing.

Norbert Lossau

lossau@mail.sub.uni-goettingen.de

References:

- ¹ *www.sub.uni-goettingen.de/gdz/en/projects/itinnorda_en.html*
- ² *www.emis.de/projects/JFM/*
- ³ *www.gutenbergdigital.de/*
- ⁴ *www.sub.uni-goettingen.de/gdz/dieper/*
- ⁵ *www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html#conversion*
- ⁶ *www.sub.uni-goettingen.de/gdz/en/gdz_main_en.html#conversion*
- ⁷ *www.agora.de*
- ⁸ *www.sub.uni-goettingen.de/gdz/gdz-tools/pica2xml/pica2xml_index_de.html
(description in German)*
- ⁹ *www.gbv.de/e-index.html*

Metadata and identifiers for e-journals

*Juha Hakala
Helsinki University Library, Finland*

Introduction

The term metadata has been defined in many different ways. In this presentation, the term means structured description of a resource. When this definition is used, library OPACs are metadata, just as Web index created from data extracted from documents themselves or Dublin Core-based metadata provided by authors and publishers.

From this point of view, libraries and other information intermediaries have three different main options for creating metadata systems for accessing electronic journals and their articles. Traditional MARC-based cataloguing is the most obvious choice. Full-text indexing is another, widely used option. Technologies used for this approach have improved fast, but we may still argue that humans do a better job in describing resources. Third choice, emerging in projects like DIEPER, is embedded metadata - which implies structured documents.

Identification of resources - independent of location information - has always been a high priority for libraries. In the digital world, identification is an even more relevant issue, since unless a resource is identified it is hard to preserve it for a long time. An identifier will also provide unique access to the thing it deals with.

The first half of this presentation will discuss the three methods for describ-

ing journals and their contents. The second half will concentrate on identification, and especially on the role ISSN and SICI (Serial Item and Contribution Identifier) will have in identifying e-journals and their articles.

Traditional cataloguing of journals and articles

Cataloguing of journals is a well-controlled business. ISSN International Centre has established clear rules on how to assign identifiers to journals and what kind of metadata should be provided for the ISSN international database alongside the identifier.

According to the current guidelines, also digitised journals should receive an ISSN and be catalogued into the ISSN database. From DIEPER's point of view, this policy is the correct one. The systems which participate in DIEPER will be available via single access point to human and non-human users such as URN resolution services.

Cataloguing of e-journals does pose some interesting challenges. The web tends to be less stable than the printed world: the names of e-journals change even more frequently than is the case for traditional materials, and journals may not only cease publication but disappear entirely (unless the national library has stored the journal in its digital archive). Cataloguing needs to take these and other issues into account.

Metadata and identifiers for e-journals

The ISSN community is not turning its back on the digital world. Quite the contrary, the ISSN network is busy revising its strategic plans so that they fit into the new environment. The biggest change is taking place in targeting the ISSN system: its scope will in the future be continuing resources. These can be divided into serials and integrating resources. A serial is issued in discrete parts; an integrating resource is updated or added to periodically or continuously. It will be interesting to see how integrating resources will be defined in practical cataloguing work done in national ISSN centres; what is certain is that ISSN will cover a larger part of the Web in the future.

From the point of view of this presentation, the main problem with ISSN is its limited granularity. If a truly efficient system is the aim, identification and cataloguing need to be extended to article level. We will return to identification later, and discuss only the cataloguing here.

Let us assume that DIEPER had decided to rely on MARC-based cataloguing of each digitised article. The five journals we are initially dealing with contain thousands of articles. But we must be prepared for future extension of the service into tens of thousands of articles.

MARC-based cataloguing of articles requires a lot of human resources. Individual libraries are usually not capable of cataloguing large quantities of journals. Co-operation is the best means for avoiding this

problem. For instance in Finland 40 libraries share the effort of cataloguing about 1100 Finnish periodicals into the national article index ARTO. Approximately 65.000 articles are catalogued annually. In February 2000 there were 350.000 records in the database.

According to our experience, one cataloguer can process about 6000 articles per year. This figure contains also subject description, without which a bibliographic record would not be truly useful. Thus maintaining ARTO requires about 10 man years annually, and the cumulative investment is about 60 man years by now. Retrospective cataloguing of the articles from digitised journals would require the same amount of human resources per article. Alas, these cataloguers are already booked for cataloguing current publications.

In addition to scarce human resources there are technical problems which hamper MARC-based cataloguing of articles. Not all integrated library systems support cataloguing of component parts (such as articles). An article index with no possibility for moving from journal description to related articles and vice versa is functionally less than optimal, not least because the cataloguers may need to create journal description manually into the article level record.

Full text indexing

For projects such as DIEPER or JSTOR relying fully on manual cataloguing is not possible. Luckily help is near: a

common approach is to convert scanned images into plain text and carry out full text indexing.

There are people who seriously claim that automated indexing tools will replace humans entirely in the description of electronic (text) resources. It is true that there has been spectacular or at least well advertised progress lately. What follows is a brief discussion of market trends.

Oracle has enriched its relational database product with a module called Intermedia¹. It will enable system managers to store and make available all kinds of documents, including multimedia. Some experts claim that with Oracle and Intermedia no separate text search engine is needed any more. It is likely that in the wake of Oracle other RDBMS providers will add similar functions to their applications.

Once relational database systems become able to incorporating documents themselves, the way has been paved for extending integrated library systems into what has traditionally been called digital library applications. The first ILS to have this kind of functionality is Voyager; its new ENcompass module² "provides consolidated access to archival and digital collections along with the traditional library catalog".

We may then predict that both relational databases and library systems built on top of them will be able to combine bibliographic data and documents. All data will be available via a

single user interface to end users.

We may then ask how good full text searching will be, compared with man-made cataloguing. Since we deal with OCR-converted material, there may - or will - be conversion errors. Therefore it is desirable to have an advanced indexing tool with fuzzy searching. One example of such a text search engine is Excalibur RetrievalWare³. It "supports over 200 document types stored on file servers, in GroupWare systems, relational databases, document management systems, intranets, and the Internet".

It is possible to store digitised text into an Oracle table and then create an index from the data with Excalibur. According to our experiments, the "original" word may still be searchable even if there are 2-3 OCR errors in it (this of course applies only to the Finnish language, which is notorious for long words).

Fuzzy searching will not solve all problems. Alas, simplicity of English language has fooled some developers to think so. In English, no word has more than a handful of inflectional forms. For instance, the verb walk has four forms: walk, walks, walking and walked. That is why traditional indexing programs designed for English have ignored morphology.

Unfortunately most other European languages have more complex morphology. One verb root in Finnish may have 18,000 inflected forms and one noun some 2,000 forms. A morphological analyser is clearly needed for

Metadata and identifiers for e-journals

efficient access. At the moment of writing this there are a number of products available on the market, but they are not yet used very widely in search systems, although adding this component is not too tough.

An indexing program should store both the word itself and all possible base forms in the index. The base forms should be stored in a separate field to enable both exact and morphological matches. For instance, when the indexing program encounters the word thought, it should store thought in the exact-match field and the list (think, thought) in the base form field.

The retrieval program (which does not need a morphology component) should work as follows:

1

Get a search term (base form), e.g. think or thought.

2

Find the records where one of the possible base forms matches the search term. For example, think matches the records think, thinks and thought, while thought matches thought and thoughts.

One of the companies developing linguistic tools is Lingsoft⁴. The information presented above is taken from an article "Indexing and morphology"⁵ available on their Web server. Lingsoft tools have been used for instance in an - for the time being - experimental version of the Nordic Web Index, developed in Center for

Scientific Computing (Finland). A new Finnish NWI database with linguistically enriched indexing module will be made available in summer 2000.

What else can we expect? A lot, it seems. Knowledge management tools such as Autonomy⁶; (see also⁷ for a description of the application) claim to be able to extract meaning from the mass of data we pump into them.

It is impossible to say yet if Autonomy or any other knowledge management application will really become "the Oracle of unstructured data" as the founder of the Autonomy company claims his product will be. Knowledge management companies are thriving at the intersection of two Net-driven trends: the push toward personalising services and the explosion of information in text form. But during the last decade we have seen quite a few Internet companies and trends come and go.

In short: linguistic and knowledge management tools will get better at indexing texts, including both materials born digital or digitised from printed form. On the other hand, collections that we need to manage with these tools are growing very fast. It may well be that these two trends will eliminate one another. At least up to now searching relevant data from the Internet has become steadily more difficult. Improving precision of searching is a vitally important issue.

We have now reached a conclusion that MARC-based cataloguing is too

resource-consuming for large article collections. On the other hand, relying solely on full text indexing is not satisfactory either, although modern inventions are making this option more competitive.

Embedded metadata

Indexing an entire text of digitised articles will provide tremendous recall, but very bad precision, especially if the articles are in plain text (ASCII files) or do not have an internal structure. Therefore in a project like DIEPER which aims at creating very large collections it is important to experiment with embedded metadata.

Instead of creating external resource description such as a MARC record it is possible to embed metadata - structured description of the resource - into the document. This method is efficient only if the resource is structured; then applications parsing it will be able to locate each metadata element separately and deal with it in an appropriate manner.

Syntax

Let us first discuss the subject of syntax. An ASCII file has a very simple structure; it consists of ASCII characters including line feed at the end of each file. If I write an article on metadata which includes several examples of different kinds of identifiers and publish it in ASCII, an application indexing the text has no way of knowing which identifier - if any - has been assigned to the article itself.

Most electronic resources do have more or less sophisticated internal structures. As an example, every image file begins with a header, which contains data the viewer application needs to present the data. The very first data element in every image format is "magic string", which identifies the image format used. As there are more than 100 image formats with different header structures, format identification provided via magic string is vitally important.

Internal structure does not guarantee that useful metadata can be embedded into the resource. For instance the image formats most commonly used in the Internet - GIF and JFIF/JPEG - are very simple and support poorly embedded metadata. TIFF is better in this respect; the price to pay was that TIFF viewers were difficult to develop. And even TIFF is a poor format when compared with text formats such as HTML.

At this stage it is important to make a distinction between proprietary and open formats. The internal format used by Word is structured, but it is owned by Microsoft. The company can change the format any time, and is not obliged to listen to users while making changes. Neither is Microsoft obliged to publish all details of the Word format. It has been claimed that Adobe has not revealed all aspects of PDF, in order to give their own products a competitive edge. This claim, even if untrue, tells something about the suspicion the customers have towards proprietary formats.

Metadata and identifiers for e-journals

As far as I know, there are no published guidelines for embedding for instance Dublin Core-based metadata into Word or PDF documents. This kind of specification may and indeed has been developed by individual projects, but there is no way tools developers could be aware of all such specifications. What works in internal document management system development projects is not applicable to European initiatives such as DIEPER, so we must look for more open standards.

Hypertext Mark-up Language (HTML) is an open specification controlled by World Wide Web Consortium. Anyone can retrieve the full specification of HTML 4.01 from the W3C Web server⁸. In principle anyone can participate in the development of HTML; for instance the Dublin Core community was able to enrich NETA tag with features needed for efficient mark-up of Dublin Core data.

Openness of HTML - both as regards policy and technical details - enables various interest groups to make their own syntax specifications on top of HTML 4. For instance, the Dublin Core community has published in December 1999 an Internet standard which specifies how to encode Dublin Core Metadata in HTML⁹. The people who develop applications which harvest documents from the Internet and extract metadata from these resources may check from this Internet standard how Dublin Core data are - or should be - structured in HTML files.

Dublin Core community has concentrated on HTML, since most Web documents have been and continue to be published in this format. But we can expect that XML¹⁰ will become steadily more popular. It has been said that XML is SGML for HTML. Technically XML is an application profile of SGML. On the other hand, HTML has been published as a set of XML 1.0 Document Type Definitions¹¹.

XML provides us with a good platform for embedding metadata. But if we rely on XML only metadata is machine-readable, but not yet machine understandable. The purpose of W3C Resource Description Framework (RDF¹²) intends to accomplish this. When XML data are declared to be of RDF format, applications will be able to "understand" metadata without any prior arrangements.

XML support is rapidly becoming a common feature in text editors and Web browsers. Whether RDF will be equally successful is by no means clear. However, within DIEPER we have decided to utilise XML/RDF. It is an open and extensible data format, which maps well to what the project is doing. There is also a specification developed by Dublin Core community which specifies how to embed Dublin Core metadata in XML/RDF documents. This draft recommendation, is called Guidance on Expressing the Dublin Core within the Resource Description Framework (RDF)¹³.

Semantics

Digitisation projects such as DIEPER

can not choose metadata format freely. There are several limiting factors:

There must be a specification for how to encode data in the appropriate syntax. For instance, if a project uses HTML, it should not use MARC21 since there is no agreement on how to encode MARC21 records in HTML. But the project could choose SGML, since there is a Document Type Definition for MARC bibliographic format.

It should be possible to convert from the chosen format to MARC and vice versa. Otherwise metadata available in the article database can not be utilised in traditional cataloguing.

The chosen metadata format (and syntax) should be familiar to Web indexing applications such as Ultra-seek. These indexes may provide a secondary access route to the materials.

The format should be internationally known. Choosing a national format such as MAB2 would be an unfair burden for partners who do not know the application.

The format and underlying rules for description should be very flexible. Ideally there should be no obligatory data elements at all; on the other hand it must be possible to add project-internal data elements as well.

There should be off-the-shelf tools for creating metadata in a chosen format and syntax.

For the reasons outlined above the best format choice for embedding metadata into digitised articles is Dublin Core Metadata Element Set. Its home page¹⁴ contains the following scope statement:

"Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies, and commercial organisations."

This quote from the Dublin Core homepage captures in a few words the process long-time Dublin Core activists have seen with their own eyes. The format was designed for layman usage; however, we see more and more information intermediaries such as librarians using the format. The authors themselves using the format are still exceptions from the rule.

The Dublin Core Metadata Element Set was initiated by a rather informal group of librarians, networking people and content specialists who attended in March 1995 what is now called the first Dublin Core Metadata workshop. As of this writing the last such workshop is DC-7, which took place in Frankfurt, October 1999. (For information about these and other Dublin Core workshops¹⁵).

Since 1995 Dublin Core has rapidly gained popularity. In response to the growing implementer community, the

Metadata and identifiers for e-journals

Dublin Core key developers - still the same people who came up with the idea of establishing the format back in 1995 - have taken the steps of formalising the Dublin Core maintenance.

OCLC is formally responsible for maintaining the Dublin Core. A DC Directorate has been formed for this purpose. The fact that OCLC is investing heavily in Dublin Core - without, by the way, being paid for it - is at least for me a proof that there is no conflict between developing and using the Dublin Core and MARC, on the contrary - these formats support one another.

The 15 Dublin Core elements have been divided into three classes (Content, Intellectual property and Instantiation) in the following way:

Content

Title
Subject
Description
Source
Language
Relation
Coverage

Intellectual Property

Creator
Publisher
Contributor
Rights

Instantiation

Date
Type
Format
Identifier

The early versions of the Dublin Core had 13 elements. Coverage and Rights were added not because of superstition, but since the attendees of the third Dublin Core workshop, which concentrated on the use of the format to describe images, thought that these elements are necessary. In the future there will be no more additions: there will be 15 Dublin Core elements for a very long time. One reason for this is that the 15 elements have been "cast in T-shirt"; the list of the 15 elements was printed on the T-shirts given to the attendees of DC-4 and DC-5.

A more serious explanation is that adding any element - even a very useful one - or removing existing elements would most likely undermine the reliability of the Dublin Core initiative as a whole and especially within the implementer community. So, making any changes at element level does not really seem to be a good idea, even if there were good theoretical reasons for doing it.

Dublin Core with 15 elements is a very simple format. But this is not all there is; with element qualifiers and private extensions Dublin Core can be made as complex as needed. A good visual analogy is perhaps a fractal image, which looks simple initially, but a closer look reveals an unending wealth of details.

The existing Dublin Core implementation projects rely heavily on the usage of qualifiers. I am not aware of any project using the 15 basic elements only. It has even been argued that

some elements, such as Date and Identifier, do not make sense (at least to the computers) unless qualified. Be that as it may, qualifiers are an essential part of the Dublin Core, and the need to standardise the core qualifiers to the 15 DC elements is acute.

Given the vital role the qualifiers have for users of Dublin Core, it is somewhat strange that there are no formal agreements on what the core qualifiers for each element are. But this situation will soon change. The next step in Dublin Core standardisation will be specification of the core qualifiers of the 15 Dublin Core elements. Accomplishing this has been much more complicated than the specification of the elements themselves.

At the time of writing this it seems that the Dublin Core Advisory Committee will finalise voting on core qualifiers on 15 March 2000. Working groups developing element qualifiers completed their proposals in December 1999, and voting - which is done at qualifier level - is now well under way. The approved element qualifiers will be published soon after the vote is over.

The proposals are available on the Web; for instance, the proposed Date qualifiers can be viewed at¹⁶. It is not yet known how many qualifiers there will be, but educated guesses often hit the 50-60 range. Once the core element qualifiers have been agreed the expressive power and complexity of the format will grow a lot. But it is important to keep in mind that just as

any element can be omitted, there is no obligation to use qualifiers either.

The metadata format and syntax developed in DIEPER (Enders) will be made as fully compliant with the qualified Dublin Core as possible, in order to maximise interoperability. Just like in many other projects, it is clear that DIEPER will need some internal data elements to achieve the required functionality. These data elements can easily be encoded in XML/RDF in such a way that they pose no problems for applications expecting to and capable of receiving Dublin Core only.

DIEPER will use Dublin Core elements Identifier, Title, Creator, Contributor, Publisher, Language and Subject. Proposals for private elements are ItemNumber, ItemNumberSorting, SerialsNumbering, PlaceOfPublication, FormatSourcePrint and SizeSourcePrint. It is unlikely that other projects digitising articles will come up with exactly the same elements, but this is not even necessary. However, we hope that the projects can agree on using Dublin Core as the lowest common denominator.

Within DIEPER the workflow of the digitising process has been designed in such a way that some metadata can be produced programmatically. Whether this will reduce human effort needed for resource description to manageable level in all DIEPER sites remains to be seen.

Identification of e-journals

In this part of the article we will concentrate on how to identify e-journals and their articles in the most efficient manner.

In any project dealing with electronic resources there is a need to generate a unique identifier to the finest level of granularity. That is, each resource that can be delivered as a separate item, must be identified, and this persistent identifier should serve as a link between the external metadata and the resource itself. One metadata record may thus contain pointers to several resources.

Two questions are discussed separately below:

1

What identifiers to use and when?

2

Should we implement resolution infrastructure in order to provide persistent linking between metadata and resources, and if so, how to accomplish this with identifiers chosen?

Identifier usage

In a European - or possibly global - system like DIEPER it is necessary to avoid usage of internal or local identifiers. These may later conflict with other identifiers used for the Internet documents. Internal solutions will also make it difficult to use URN or DOI resolution services to access the data.

There are at least three traditional identifiers that may in principle be

used for DIEPER materials: ISSN, SICI (Serial Item and Contribution Identifier) and National Bibliography Numbers.

ISSN

The information included herein is partially based on discussions with representatives of the ISSN International Centre. The author wishes to express his gratitude to Ms. Françoise Pellé, Mr. Slawek Rozenfeld and Mr. Pierre Godefroy for the information they have provided.

According to the rules of the ISSN centre, ISSN numbers can (and indeed should) be applied to old periodicals when digitised. If the original printed document has an ISSN, this identifier is also valid for the digital version. ISSN guidelines contain the following chapter:

"A reproduction is a copy of an item and intended to function as a substitute for that item. The reproduction may be in a different medium from the original but it is not a different edition in itself. The ISSN assigned to the original is valid for the reproduction, a new ISSN is not assigned to the reproduction."

From a global point of view it is vitally important to acquire ISSN numbers to the periodicals digitised in projects such as DIEPER and JSTOR. When journals are subsequently catalogued to the ISSN database, there will be a single access point for getting information and possibly also access to the journals digitised in diverse projects. The ISSN international centre

is eager to establish this kind of co-operation with DIEPER and similar projects elsewhere.

In the Internet, ISSN numbers are used in diverse ways. Some publishers embed it only for the homepage of the journal. From the information retrieval point of view, this is a recommended practice. Some vendors also put ISSN numbers into issue home pages and even all articles. This is counter-productive, since then ISSN search will yield very unpredictable results.

To sum up: all digitised periodicals should get an ISSN and be catalogued into the ISSN database. ISSN numbers should not be used for identifying issues or articles.

SICI

The SICI standard (Serial Item and Contribution Identifier Standard, ANSI/NISO Z39.56-1996 Version 2) provides an extensible mechanism for unique identification of either an issue of a serial title or a contribution (e.g. article) contained within a serial, regardless of the distribution medium (paper, electronic, microform, etc.). SICI is based on ISSN and has been designed in such a way that it can be generated automatically from a bibliographic record or from structured text documents.

The SICI home page, with full set of documentation and a link to the SICI generator, is available¹⁷.

The SICI is a combination of defined segments, all of which are required.

These segments are:

1

Item Segment, the data elements needed to describe the serial item (ISSN, Chronology, Enumeration)

2

Contribution Segment, the data elements needed to identify contributions within an item (Location, Title Code, and other numbering schemes in a specific instance of the SICI).

3

Control Segment, the data elements needed to record those administrative elements that determine the validity, version, and format of the code representation. This is the most important segment of the SICI. Interpretation and processing are determined by the Control Segment.

The SICI examples shown below utilise examples using *Mathematische Zeitschrift* taken from (Enders). I am assuming that ISSN has been registered for the digital version of the journal, and it is 0002-8231 (which actually belongs to the JASIS).

SICI for issue 1 in volume 30 would be:

0002-8231(1929)30:1<>1.0.CO;2-X

The Code Structure Identifier (CSI) specifies the code type as a Serial Item Identifier, Serial Contribution Identifier, or other. For issues CSI is 1 (as in above example); for articles it is 2.

Please note that in this fictitious example the check sum (X in the end)

has not been properly calculated. Code "CO" is medium/format code (MFI) which in this case indicates that the document is available on-line. MFI is always "CO" for digitised articles; printed articles would have MFI value "TX". The protocol version, which precedes the check sum, is always 2, until new version of SICI is published and projects start using it.

The Derivative Part Identifier (DPI; in the above example zero before "CO") provides a method for the designation of an identifiable component part of the serial item. DPI can define table of contents (DPI=1), index (DPI=2) or abstract (DPI=3). SICI for the table of contents of the issue above will be:

0002-8231(1929)30:1<>1.1.CO:2-Y

Please note that in this fictitious example the check sum (Y in the end) has not been properly calculated.

For the article "Zum Beweise des Minkowskischen Satzes über Linearformen" published in volume 30, issue 1 of the *Mathematische Zeitschrift* the SICI would be:

0002-8231(1929)30:1<ZBDMSU>2.0.CO:2-Z

The Title Code (in the above example, "ZBDMSU") is constructed from the title using basically any and all title words without attempting to distinguish titles from subtitles. All characters are converted to upper case. All characters not belonging to ASCII-7 must be converted. Therefore "ü" be-

comes "u". If DIEPER-assigned SICIs are used as URNs, then coding practices specified in RFC 2288 should be followed. In practice this means that a few characters such as ">" must be converted to hexadecimal form.

If we want to identify a digitised image of the first page of this article, this can be done by adding the original page number in front of the title code. For instance, if the page was 101, SICI would be:

0002-8231(1929)30:1<101:ZBDMSU>2.0.CO:2-A

SICI seems to meet quite well the identification needs the journal digitisation projects have. There are some requirements in the DIEPER E-DOC format proposal which do not match SICI; for instance, there is an explicit way to specify in SICI that the identified thing is List of Illustrations, but implicitly Title code can be used for coding this information.

SICIs can easily be extended to URNs; this requires only standard prefix (most likely urn:sici:) and character conversion according to the rules defined in the RFC2288.

DOI/URN implementation

In the Internet we need not only identification of resources, we also need persistent resolution mechanism; something that makes possible e.g. durable linking between references and referred documents. The present Internet uses URL for this purpose, with the results we are all familiar with.

There are two systems that may currently be used for creating persistent linking, namely DOI (Digital Object Identifier) and URN (Uniform Resource Name).

Internally DIEPER, JSTOR or similar initiatives would not need a resolution system such as DOI or URN. But we need to provide persistent access to DIEPER resources from anywhere within the Internet. With the DOI/URN umbrella above us, authors could provide persistent links based on these systems to DIEPER journals. This linking mechanism would be immune to URL changes, which are bound to happen over decades and centuries.

Let us compare DOI and URN first.

The Digital Object Identifier¹⁸ is an identification system for intellectual property in the digital environment. Developed by the International DOI Foundation on behalf of the publishing industry, its goals are to provide a framework for managing intellectual content, link customers with publishers, facilitate electronic commerce, and facilitate automated copyright management.

Technically DOI is based on the Handle system¹⁹. There are already DOI-based experimental systems; one of them is described by (Risher).

Although we know that DOI can for instance link references and referred documents, it may not be appropriate for digitised journals which may

lack commercial potential. DOIs are not free; there will most likely be an annual payment for each DOI. Neither is assignment of DOIs cost-free: it is necessary to register publisher ID, which will form part of the DOIs assigned by this organisation.

The DOI business model in its current form is not viable for projects like DIEPER who deal with materials with limited commercial value. Technically the system seems to be solid enough, but Handle system is unfortunately not a standard, but just a technology. However, there are attempts to standardise DOI which, if successful, will improve the situation in this respect.

Uniform Resource Names²⁰ framework is being developed by the Internet Engineering Task Force's URN Working Group. Standardisation is not yet quite complete, but the aim is to finish the work in the near future. Eight Internet standards which define most of the URN framework have been approved, and only three documents remain Internet standard drafts. These documents are generally regarded as mature.

Despite some criticisms towards the URN system from W3C, IETF remains committed to it. This is good news for national libraries and the ISSN International Centre, who are among the early implementers of the URN system.

URN consists of three parts. Every URN will start with string "urn:". This

Metadata and identifiers for e-journals

will make it possible to find URNs even from documents that do not have structure. Second part, Namespace identifier (NID) specifies the system used as identifier. For instance, NID for ISSN will most likely be "ISSN". The third part, Namespace Specific String (NSS) is where the identifier is put.

Once NID has been agreed on to an existing identifier, it is technically trivial (and cost-free) to generate URNs from the existing identifiers. For instance, to create an URN from ISSN you need only to add prefix urn:issn: into every ISSN. Of course not even this is necessary; if a URN resolution request arrives, the local application can easily drop the URN prefix and deal with the namespace specific string only.

Contrary to DOI, the URN string does not contain any hint as regards where to find a resolution server. As an aside, it is not clear how long the publisher ID in DOI will give a valid hint on location of the resolution server, since publishers are usually more short-lived than their publications. In the case of ISSN, it is easy to find a resolution server, since the ISSN database maintained by the ISSN International Centre can resolve all ISSN-based URN. A user may not get the home page of the journal if it has disappeared or moved to other location and URL in the bibliographic record is out of date, but you will always get a description of the resource.

At the moment there are only experimental URN-based resolution services, of which one has been built by the

ISSN International Centre. The ISSN implementation includes plug-ins for Netscape Navigator and Microsoft Internet Explorer. These plug-ins enable the users to type ISSN-based URNs into the Location window of the browser. The users get back a description of the journal, which includes one or more URLs.

There have been some preliminary discussions about how to support resolution of SICI-based URNs via the ISSN database. This would enable efficient co-operation of the ISSN database and DIEPER archives, if URNs are utilised in the latter. Of course this mechanism would be scaleable to all similar journal archives.

Technically using the ISSN database as a gateway would not be too difficult. Each journal description should have two pointers, one in 856 \$u to the homepage (if in existence), another in a tag not yet defined to the URN resolution service which will be able to handle SICIs based on this URN. The latter address will be the same for all journals sharing the same archive.

Without the existence of the ISSN database SICI would be a problematic identifier to resolve within the URN system. As a dumb code it does not provide any information on where to find a resolution service. ISBN on the other hand, does contain such a hint. The problem is that time renders these hints useless. When publishers and even countries disappear, ISBN will lead users astray. Therefore there are serious plans for revising ISBN too in such a way that it becomes a dumb code.

SICI namespace has not yet been registered within the URN system, but for experimental purposes DIEPER and similar projects could easily expand SICIs into URNs by adding urn:sici: -prefix into the current SICIs and storing the resulting URNs in their databases.

At the same time the projects should negotiate with ISSN International Centre about practical tests, which may in the future lead to closer practical co-operation. Without the assistance of the ISSN International Centre there is no feasible way to resolve SICIs, so support from ISSN IC is a prerequisite for efficient URN utilisation.

Any journal digitisation project might of course establish a "poor man's" URN resolution service immediately by indexing URNs and making them searchable internally, but this will not add much value to using SICI for searching purposes. Searching URN instead of SICI via e.g. Alta Vista means just some extra typing with the same end result. The real value of using URNs comes from the establishment of resolution services.

Implementing URN resolution service locally is, according to the ISSN International Centre, relatively easy if the search system already supports HTTP protocol. Thus it can be hoped that this kind of experiments become common in the near future. Creating the mechanisms needed for supporting URN resolution on a global scale may unfortunately be more complicated and time-consuming. Although the technologies proposed by the

URN Working Group have been tested and experimental software exists already, the distance from experiment to production may be long.

Juha Hakala

juha.hakala@helsinki.fi

References

Enders, Markus: E-DOC format for DIEPER. 2nd edition, 17.02.2000. Will be published on DIEPER Web site *www.SUB.Uni-Goettingen.de/gdz/dieper/*

Risher, Carol: Reference Linking with DOIs: a case study. D-Lib Magazine, February 2000. Electronic resource, available at: *www.dlib.org/dlib/february00/02risher.html*

¹ *www.oracle.com/intermedia/*

² *www.endinfosys.com/new/encompass.html*

³ *www.excalib.com/products/rw/index.shtml*

⁴ *www.lingsoft.fi/*

⁵ *www.lingsoft.fi/doc/indexing/morph.html*

⁶ *www.autonomy.com/*

⁷ *www.wired.com/wired/archive/8.02/autonomy.html*

⁸ *www.w3.org/MarkUp/*

⁹ *ftp.funet.fi/pub/doc/rfc/rfc2731.txt*

¹⁰ *www.w3.org/XML/*

¹¹ *www.w3.org/TR/xhtml1/*

¹² *www.w3.org/RDF/*

Metadata and identifiers for e-journals

¹³ www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/

¹⁴ purl.org/dc/

¹⁵ purl.org/dc/about/workshop.html

¹⁶ www.mailbase.ac.uk/lists/dc-usage/files/datefinal.html

¹⁷ sunsite.berkeley.edu/SICI/

¹⁸ www.doi.org/

¹⁹ www.handle.net/

²⁰ www.ietf.org/html.charters/urn-charter.html

Three stories from the future

*Kirsten Strunck and Jens Thorhauge
Danish National Library Authority, Copenhagen, Denmark*

Summary

Scenario 1

The costs and other difficulties in connection with digitising printed journals full scale were considered too extensive, so digitising of printed material is only done for preservation reasons.

The use of printed materials from the libraries is stable and for instance monographs are still often published in printed form.

The electronic search facilities have improved, records, tables of contents, abstracts are digitised for a growing amount of material.

The European co-operation on inter-lending has improved and each country runs a "printing house" that delivers articles on demand in whatever possible form. Not only foreign but also many national libraries rely on this national journal-printing house for their non-core journal use.

Scenario 2

The professional use of information is nearly entirely based on electronic information. Hence to avoid that the huge mass of scientific and research information especially stored in printed journals should not be utilised digitisation programmes were - at different speed - started in all European countries.

From the beginning it was obvious that standards were needed and

consequently - after a great number of conferences - the European Digitisation Centre was established with a co-ordinating function, and the task to digitise the European documents which the committee and its working parties select.

The old type of library with huge collections in printed form have been dramatically reduced and replaced by a small number of more deposit-like libraries.

Scenario 3

Researchers use a mix of printed and electronic information. In the first place the minor European countries decided to digitise the bibliographic entries but not the full text, because the costs would lead to a total change in library priorities. The English journals were however digitised and the growing use of them - and the decrease in use of national journals - was one element in the growing anglofication.

Hence various national policies were established with the common objective to select the most important and most used printed documents for digitisation. The selection can be based on citations or a specified number of requests. Obviously the researchers think that more should be digitised, but on the other hand they would not appreciate a further decrease in the service of the libraries.

The three scenarios are designed to provoke a discussion. It is important

Three stories from the future

to stress that they are not an attempt to predict the future, but rather to visualise various possible consequences of decisions we are going to make in the near future.

The scenarios should have a high probability according to the premises on which they rest. They might be slightly exaggerated, but are in principle down-to-earth images of different possibilities.

The idea is not to choose one of the scenarios - and naturally none of them will become true in the primitive form in which they have been drafted. The idea is rather to combine various elements and try to choose the best way to reach the goal we are heading for.

Scenario 1

Idea: It is the records (bibliographies and catalogues) of the information production of the past which are digitised, not the full text. Library service is of high standard and the bibliographic effort is considerable.

A European student, sitting at his workstation, is searching for information about the origin and development of Darwin's theories. The interesting question for the student is what he can gain access to directly in digital form, and what he has to order from the library in printed form. The initial search shows that there is a huge number of articles in digital form, and likewise a huge number in old printed journals. He only finds "The Origin of Species" from Darwin's own hand in

digital form. So he decides he'd better ask a librarian to help select the articles, orders a handful of the most obvious titles in paper form from the university library, prints another handful of obvious new articles and sets off for the library after making an appointment with his favourite librarian.

The library available to the student is a hybrid library. All types of information media in printed form are kept in libraries which also give access to electronic resources to those users that do not have the necessary facilities themselves to use the network. The digitisation technology which was formed towards the end of the 20. century has ensured that preservation threatened documents have survived in digital form.

Out of concern for the preservation of the cultural heritage and also the worry whether information in non-digital form will be used optimally, it has previously been considered to digitise the existing printed documents. High on the agenda has been the digitisation of periodical literature, which for centuries have reported on the progress of research. The task would be quite enormous and very, very demanding on resources.

Periodical literature takes up a lot of room in the libraries' collections. On the other hand the volumes of these works does not increase very much at all, as electronic publishing of scientific articles has won the day early on in the 21. century. The preservation

of periodical literature is not particularly threatened either, as it was quite common already towards the end of the 20. century to copy articles for the users instead of lending. The result of the discussion on digitisation of journals has therefore been that one only digitises the decidedly preservation threatened documents, preferring instead to spend the money on improved bibliographic systems and professional advice.

Retrieval and localisation of information media in printed form as well as electronic documents are guaranteed through a bibliographic system in electronic form.

Libraries, bibliographers and lately commercial firms have for centuries been constructing a bibliographic system consisting of records and the works which the documents contain. Based on the knowledge of for example author, title and the subject of the works you may here find the documents that will fulfil a need for a certain kind of information or experience.

Catalogues of individual libraries' collections have been produced for more than 1000 years. Bibliographies, too, go back a long way. Since the middle of the 20. century great energy has been exercised globally to make it a national obligation, that each country records the national knowledge production in national bibliographies. UNESCO and IFLA have published national bibliographic recommendations which were

adopted at international conferences. Together the individual countries' national bibliographies represent a global recording of knowledge production. The documents were kept in the libraries - and preservation for future generations is guaranteed through legal deposit.

Catalogues and bibliographies recorded previously primarily at document level which gave rise to a commercial industry of Abstracting and Indexing services that for the main part indexed the articles in the scientific journals, often with an extended contents analysis in the shape of subject headings and abstracts. Many national bibliographies now also include recording at article level, just as many libraries have bought access to the contents of the article bases for those journals which they have on their own shelves.

Even before it became quite common to publish documents in electronic form, information technology was being used for the production of bibliographies and catalogues. The advantages of the electronic bibliographies - both as regards production and application - are so great that retro-conversion of printed catalogues and bibliographies into electronic form has been firmly supported - also financially.

National bibliographies are financed by government means, catalogues via the libraries' budgets. Abstracting and Indexing services are commercial. Retro-conversions are often financed by project means.

Three stories from the future

The development of the bibliographic system and the desire for a common utilisation of it has brought about standardisation at national as well as global level. The standardisation applies to both the bibliographic data and systems and to the communication between systems.

In order to exploit the bibliographic resources to the full, a gateway to the bibliographic system has been established at European level. The national interlibrary loan co-operation between libraries, which already existed in some countries in the 20. century, has developed into a European system of lending paths, with common routed orderings via Internet.

So for most European countries the situation is that the recordings - and not the full text is digitised. But at a number of conferences in the beginning of the 21st century it was agreed that each country should run a "printing centre" with access to national journals and deliver copies in whatever possible form - on demand. It was also recommended that the search facilities were improved by digitising the table of contents, abstracts etc. from core journals. In most countries it is still an ongoing process. For the core journals in English the situation is different, many of them are digitised in full scale, due to the fact that they have readers all over the world.

Meanwhile the student meets the librarian and returns to his home with Darwin's chefs d'oeuvres, some art-

icles and a handful of the absolutely most important monographs on Darwin - and a print of the selected bibliography that was produced with the help of the librarian.

Scenario 2

Idea: It is the information production (the journals) of the past that is digitised as professional information use is nearly entirely based on electronic information. Accordingly the huge library collections have a deposit character and the budgets of research libraries have been changed dramatically.

A British sociologist is wondering how around the year 2000 one used the Dream society to categorise the next historical step in society's development. He starts a search from his workstation and quickly finds a rather limited number of articles. Obviously the term was only used for a short period, and never became paradigm-setting. A quick glance at them reveals to him that they refer to an even more limited number of books, one of which, The Dream Society, was published in 1999 in New York and Copenhagen. His headache increases - the answer to his question turns out to be more difficult to find than he imagined. He knows that the chance of finding it digitised is very slim indeed and that the chance of finding it in printed form in England is fifty-fifty.

At the same time a meeting is taking place in the European digitisation committee's working party for digit-

isation of journals. The European digitisation committee was appointed at the beginning of the millennium after a number of European conferences on the subject for the purpose of outlining the guidelines for digitisation of European information media in printed form. The committee has a number of working parties which deal with questions about digitisation of certain material and publication types.

By the start of the new millennium it became obvious to the library and information world that information which was not accessible in digital form within a very short time would be dead information. Studies and observations of the users' information searching behaviour pointed to the fact that their experiences with information searching on the Internet would be normative. In order to secure the use of the extensive scientific periodical literature for the generations to come, an almost total digitisation of this type of document has been given the highest priority.

Political awareness of the creation of the network society and the interest in preserving the cultural heritage lead to the establishment of the European Digitisation Centre, financed by EU means.

The Centre's primary task is to digitise the European documents which the Digitisation Committee and its working parties select for digitisation. When the Centre was established, digitisation was a comparatively expensive process. At the same time

there was no European standard as such for digitisation. This fact weighed very heavily when choosing a central European solution for the digitisation task. By establishing a European Digitisation Centre one would gain administrative advantages and avoid the problems that might occur in connection with using digitised documents which had been digitised according to different standards.

The digitisation of material is a co-operative effort between the libraries of the individual European countries. When the Digitisation Committee and its working parties have selected the materials to be digitised and worked out a prioritised order for starting the process, the publishing country's legal deposit library lends the material to the Digitisation Centre.

Retrieval of the digitised documents is guaranteed by all documents containing metadata in accordance with the Dublin Core metadata standard. Some European national bibliographies have also introduced the practice of adding the digital document's URN to the existing national bibliographic record.

The Digitisation Centre also fulfils an archive and maintenance function. When digitising documents they are also being microfilmed - to be on the safe side. Due to problems surrounding the durability of digital data and the technological development - concerning hardware and software as well as the development of certain standards -

Three stories from the future

migration of the digital data will be necessary from time to time. This continuous maintenance of the digital data is simplified by the central solution to digitisation.

Back to the meeting in The European digitisation committee's working party for digitisation of journals. Following extensive examinations of the use of digitised documents, the Digitisation Committee is re-examining its selection policy. So far a number of the absolute core journals within a broad spectrum of subjects has been included in the digitisation process. As it turned out, however, there are some "dead" digital articles in stock, which means the selection criteria must be reconsidered.

It is the revision of the selection policy which is on the agenda for the working party for digitisation of journals. Once more it must be debated whether:

- one should digitise the core journals within every subject
- the research results of some subjects become obsolete so quickly that a digitisation would be wasted, and the digitisation should therefore be concentrated primarily on the core journals of the humanities
- one should not focus on journals at all, but on those articles in the journals that are of lasting value. The question then arises how does one decide what qualifies as being of lasting value:
- selection on the basis of bibliometric studies. Scientific journals and their articles are primarily targeted at the

research world, therefore citation pattern must be the decisive factor in determining which articles are the most important

- selection based on the use of the articles. Only those articles which the user in fact tries to get access to are alive. Therefore digitisation must be done "on demand".

The meeting starts in a heavy atmosphere of recycled arguments.

Meanwhile the social scientist has found the old catalogue of The British Library. For a brief moment he was happy - the Dream Society should have been in the library, but turned out to have disappeared. He will have to wait to get it from Europe - and pay for it, too.

Scenario 3

Idea: Only memorable highlights of the cultural heritage are digitised - works with signal value for the owner institutions plus journal articles that are often used. Physical libraries continue to be important and offer good service

An astronomer is wondering how Tycho Brahe - a Danish astronomer - could make quite precise observations of the planets even if his theory of their movements was incorrect. He starts searching for information on Tycho Brahe, and is particularly interested in his instruments.

Journals are electronic - and have been for decades - but older journals you cannot count on unless they are in English.

Digitisation is done on demand, and particularly sought after articles are selected - slowly a digital journal base has been built up.

There is broad general and professional interest in research into the cultural heritage. Fields that had been considered dead for a long time are revitalised. The more obscure theories in all fields are re-examined. The easy bibliographic access to the cultural heritage and research of past generations, the subject entries, the linking and path systems on the net have resulted in a more intensive use.

The interest was also relatively high in the beginning of the 21st century. At that time it became increasingly common to publish in electronic form and to communicate in all manner of ways using the information technology. It was considered whether it was necessary to digitise the contents of information media in printed form of the past in order to preserve the cultural heritage. For centuries research had been reported on in periodical literature and one was anxious that this resource would not be exploited to the full in an electronically influenced society.

The sheer volume of information media was frightening, however, and seemed to demand far more resources than were available for digitisation. If one did not digitise documents in printed form some of them would perish. But that was nothing new. One had never previously been able to - or indeed expected to be able to preserve the entire cultural heritage. But - unlike

before - the possibility was there of obtaining financial means for digitisation and preservation of the most valued treasures.

Hence quite different strategies developed. Now in particular English and American literature has been digitised, including the old journals some of which have been published in an unbroken line since the 18th century.

On the other hand minor languages, for instance in Europe, such as the Scandinavian languages, Dutch, Portuguese, Czech, Hungarian etc. - that often have a strong learned tradition - were forced to make calculations as to the cost of a quality digitisation. Consequently it was decided in most cases that the price was too high as it would swallow the best part of the library budgets and lead to dramatic cuts in the traditional library service - including closing down many minor libraries. So the counter strategy decided upon was to continue developing the bibliographic tools for the printed and electronic material.

After a decade bibliometric studies revealed that the use of older journals in English increased while the use of printed journals in non-English languages decreased. This pattern made publishing of research results in the minor languages nearly disappear. In consequence several political programmes were established. The most important point was that the most used articles

Three stories from the future

should be digitised. In general articles were delivered from the large research libraries in a primitive digitised on-demand version. Three requests for an article would lead to governmental support for a quality digitisation and a free access on the net.

Likewise many of the small nations with their own language have government support programmes for publishing research in the national language to avoid a devaluation of the mother tongue.

At the beginning of the 21st century there was no common policy for digitisation in Europe. Therefore each library digitises to an extent which reflects its ability to provide resources for the process. The works that are being digitised are either those threatened by the ravages of time or works which the owner institutions consider to be of special value. Rumour even has it that the digitisation resources are being used on prestige projects - choosing those works for digitisation which the owners will gain most credit for.

Just as in cultural historical museums one only finds mummies of those Egyptians who could afford quality embalming, one finds on the net only the digitised works which enterprising institutions have been able to obtain the financial means for.

The astronomer has found a lot of information about Tycho Brahe and his contribution to astronomy, including a number of articles on his instruments and his observations. And as a European library has deemed his

work on the astronomical instruments worthy of a digitisation project, he has also been able to study on his screen pictures of these examples of the technology of the past.

Jens Thorhauge

jth@bs.dk

Selecting journals for digitisation: piecing together the puzzle to create a European model

*Hazel Woodward
Cranfield University Library, United Kingdom*

As we have heard from a number of previous speakers, selecting and digitising journals is very much a global activity. We know that there are many digitisation initiatives being undertaken in individual institutions and organisations world-wide but it is often difficult to find out information about these initiatives. There are also a number of consortia-based journal digitisation projects. Some of the consortia involved in digitisation have a subject focus - digitising for both preservation and access in a particular specialised subject area. Others may be regional or cross-sectoral consortia - public, special and academic libraries working together with art galleries and museums, for example, to spread the cost of preservation and widen access to important materials. And last, but not least, we are all aware that at national and international levels there are also many organisations and institutions undertaking digitisation of a variety of information resources - including journals. The question we need to address in this forum is, how can we get all these initiatives to feed into a cohesive European model for journal digitisation?

In this presentation, I intend to address selection issues from a local, consortia, national and international perspective. As far as local digitisation initiatives are concerned the critical factor that affects selection decisions is the mission of the organisation:

"Each institution developing digital collections will have a distinct organisational mission and context which will define and influence the scope of its activities, its criteria, strategies and procedures for (selection) acquisition and collection development" (Arts and Humanities Data Service)¹.

We all have different missions. My mission at Cranfield University² - a relatively small, specialised, post-graduate institution - is quite different from that of a large academic research institution with a remit to build a large archival research collection. My library is very much to do with access; we do not hold large collections of materials but we do spend a large amount of the budget on electronic information resources and on document delivery. Thus, when libraries are deciding what to digitise at institutional level, they must examine their own mission in order to define the selection criteria to develop both digital and printed collections. By so doing, they will be applying consistent selection criteria to the acquisition of optimum resources for the "hybrid" library.

The types of journals that are being digitised at local level tend to fall into two categories. Firstly, journals which are created in-house and for which copyright is owned. The importance of copyright has been discussed in previous presentations and it is an issue that is highly significant in terms of digitisation decisions, not least because obtaining electronic

**Selecting journals for digitisation:
piecing together the puzzle
to create a European model**

copyright permissions is an exceedingly tedious task. I have direct experience in this area as I was director of one of the Electronic Libraries Programme (e-Lib)³ projects - Project ACORN (Access to Course Readings over Networks) - from 1996 to 1998. The aim of this project was to create/acquire digitised journals articles for an electronic short loan collection and one of the main areas of activity was obtaining the electronic copyright permissions from individual publishers.

The second category of journals which may be considered for digitisation at a local level, are those contained within special collections which may be deemed of national or international research importance. Currently, large collections of printed research materials are buried deep in the archives of libraries world-wide and digitisation is an excellent way of preserving these items, improving dissemination of what is available and providing access to the material. The fact that many automated library management systems are developing "digital media archive" modules, almost certainly means that the number of local digitisation projects will increase dramatically. At Cranfield we are the first UK University library to install Sirsi's Hyperion⁴ software which facilitates the management, storage and retrieval of multi-media and digital collections and we are actively developing our strategy for building collections to support the learning, teaching and research within the University.

As our academic colleagues develop their e-learning materials and distance learning courses, there is going to be an even greater need for us to work closely and collaboratively with them to create, and provide access to, collections of digitised materials - some of which will be journals. In parallel with local scanning and digitisation it is likely that libraries will also purchase pre-digitised materials on demand from sources such as Project HERON⁵ in the UK. HERON is attempting to set up a national database of digitised journal articles and book chapters and to make these available to the academic community. The advantage of such a database is that it provides information on what has already been digitised, thus saving duplication of effort, as well as providing information on rights clearance and an easy method of purchasing materials.

Let us now move on to examine consortia projects involved in journal digitisation. There are a number of interesting projects worldwide. Astrid Wissenburg spoke earlier about the various eLib projects such as the Internet Library of Early Journals⁶ and some of the Hybrid Library projects such as MALIBU⁷, BUILDER⁸ etc. which include elements of journal digitisation. There is another strand of funding available within the UK called the Research Support Libraries Programme (RSLP)⁹ and, as one example from this programme demonstrates, the Glasgow Digital Library Project also includes journal

digitisation. And just to show that my examples are not all UK-based, the US Arches Project¹⁰ is providing an on-line repository for collaborative digitisation projects.

As far as selection within consortia initiatives is concerned, projects normally focus upon a particular subject area or group of collections. Unlike at local level, where individual institutions make selection decisions according to their own organisational mission, consortia-based projects are looking at a wider picture. And because most of the projects require external funding, selection criteria have to be carefully formulated, presented and supported in the project proposal, and be subjected to peer review. Those projects which are deemed of sufficient interest to the community will be funded. An important benefit of receiving external funding is that the funding body will also co-ordinate the dissemination of information about the project. So, for example, if you visit the Research Support Libraries Programme, the Joint Information Systems Committee (JISC)¹¹ or the eLib Web pages, they give full details of their programmes and further information on all funded projects. However, there is no central service which provides information on all UK journal digitisation projects.

Finally we move on to the national scene and there has already been much discussion within the workshop on this topic. Prior to writing this presentation, I spent some considerable time browsing the Web

and looking at various national library Web pages. I was particularly interested in seeing what information was available on these sites about digitisation, digital preservation, and national policies on these topics. The following provide a few examples:

The Australians are very active in the area of digitisation and they have a number of committees examining, and producing policies and guidelines on, issues pertaining to the digitisation of their national heritage - including the selection of online materials.¹² Selection criteria include:

- be about Australia
- be written by an Australian of recognised authority
- be on a subject of social, political, cultural, religious, scientific or economic significance and relevance
- have long term research value (e.g. peer reviewed journals).

In Canada, the National Library of Canada has an Electronic Collections Co-ordinating Committee¹³ which has produced a "Networked Electronic Publication Policy and Guidelines". Similarly, in the UK the JISC is co-ordinating IT strategy for UK Higher Education. The JISC Committee on Electronic Information is developing the Distributed National Electronic Resource (DNER) - a national strategy for developing quality assured information resources.

"The DNER is a managed environment for accessing quality assured information resources on the Internet which

**Selecting journals for digitisation:
piecing together the puzzle
to create a European model**

are available from many sources. These resources include scholarly journals, monographs, textbooks abstracts, manuscripts, maps, music scores, still images, geospatial images and other kinds of vector and numeric data, as well as moving pictures and sound".

Within the DNER, the National Electronic Site Licence Initiative - NESLI¹⁴ is currently working on providing wider access to electronic journals. As a member of the NESLI Steering Group, I thought it might be helpful to summarise for you the context and the aims and objects of NESLI:

- NESLI forms part of the wider DNER initiative
- NESLI aims to:
- increase and improve access to e-journals in higher education institutions (HEIs)
- negotiate value-for-money deals for e-journals for HEIs

The NESLI initiative has been quite an uphill task. It is overseen by a Steering Committee who have appointed a Managing Agent whose job it is to deal with all aspects of negotiating with publishers, subscriptions, access interface, and authentication. The Managing Agent also promotes the use of the NESLI Model Licence, another interesting element with the potential to save both libraries and publishers a great deal of time and effort. An overriding aim of NESLI is to separate print and electronic subscriptions. Although some of the publishers' offers currently available still tie electronic

subscriptions to print, the Managing Agent is in negotiation with many hundreds of publishers at the moment and many of them are saying that they will shortly be allowing the separation of print from electronic. This is an extremely important development.

Another very interesting, and relatively new development is one that involves a commercial supplier - Chadwyck-Healey - who have announced that they intend to digitise a minimum of one hundred back runs of journals each year for the next three years, to provide full text back-up for their Periodicals Content Index (PCI). I would suggest to you that this is going to impact upon the current economic model, now that we have commercial companies moving into large scale digitisation projects.

So how do we go about developing our European model? It seems to me that one of the most useful things we could do would be to start developing a European Gateway to digital information which could provide a single point of access which would bring together information on, for example:

- policies, strategies and guidelines
- organisations and Web sites
- data documentation and standards
- database of digitised journals
- news, events, discussion lists.

Such a gateway would enable all European libraries to find information quickly and easily and, in many cases, save us re-inventing the wheel in terms of policies and guidelines or

re-digitising materials. As a model, I would highly recommend the National Library of Australia's Preserving Access to Digital Information (PADI) Website which is well developed along these lines.

Adding content to the European Gateway could be done at a variety of different levels. We know that many journals are being digitised within different organisations and projects but many of the projects are short term and will not be funded indefinitely. Moreover, many of the exit strategies of these projects are unclear as to what is going to happen to the digitised journals, once the project funding is finished. The European gateway could provide a repository for such titles. But perhaps more importantly, there are many thousands of journals published by learned societies and commercial publishers and the digital archive of these titles is growing rapidly. Most major scholarly journal publishers now have five or six years back issues in digital format. Clearly many European libraries would like to have access to these back issues and would be willing to pay for this access - as can be seen from the success of JSTOR. Although, as yet, we have only limited information about usage it is becoming reasonably clear that improving access is going to increase the usage of this material. I would very much agree with David Bradbury's comments about the fact that libraries have a bad track record when it comes to dealing with journals. Most libraries only catalogue journals at title level - ignoring the

valuable article level information which in many cases has consumed significantly more than half their materials budget. Access to the actual content of journals is provided by guiding users to databases that rarely give any indication of whether a title is held in the users library. After searching the database the users must then return to the library OPAC in order to discover whether useful references are available locally. Such a situation can be vastly improved by providing electronic access; but can we organise this on a European scale?

In order to achieve a "European Centre for Journal Digitisation" we would need co-ordination, collaboration and strategic planning at a European level. Indeed, there is already collaboration in this area. Project Nedlib - the Network European Deposit Library - involves a number of national libraries across Europe, and it may be possible to develop that project to encompass such a Centre. Clearly, we would have to ensure that there is representation across Europe from all the national libraries and other national bodies involved in electronic resource provision. I would further suggest that we might look at the NESLI model in terms of content gathering for the European Centre. I feel sure that there would be a role for a Managing Agent to undertake the negotiation with publishers, liaise with all the various organisations involved in journal digitisation (such as the Higher Education Digitisation Service [HEDS] in the UK) and to manage licence agreements and subscriptions.

**Selecting journals for digitisation:
piecing together the puzzle
to create a European model**

Of course, such an initiative would require at least some start-up funding but it could possibly be self-funding after a start-up period. As with NESLI, the Managing Agent would be guided by a Steering Committee. Many different issues relating to selection would require careful consideration. For example, should journals in the humanities and social sciences be digitised first? After all, there is a body of research in the print environment that shows usage of this literature to be heavier over a longer period of time than that in the science and technology disciplines. However, this does not mean that scientists do not require digitised backruns; chemists, for example, will tell you that they also frequently require access to older chemistry material. Perhaps the way forward is to think in terms of subject clusters. I ask you to cast your minds back to the keynote speaker - Professor Jean-Pierre Bourguignon, a mathematician. There is clearly a lot of interest from academics and researchers in obtaining access to core journals within their own subject area. Subject portals are already developing rapidly across a wide range of disciplines. It may be that our European Managing Agent could start up discussions with some of the major learned societies. I am sure that they would welcome the opportunity to widen access to information for their members; indeed, there is evidence that their members are exerting pressure on them to think about how they might expand access to older materials in their particular areas of interest.

There will be many challenges and many opportunities if we do move towards the setting up of a European Centre for Journal Digitisation. However, I believe that, as many of the workshop speakers have demonstrated, some of the important building blocks have been created. We just need to start stacking the blocks together.

Hazel Woodward

h.woodward@cranfield.ac.uk

References

¹Arts and Humanities Data Service:

www.ahds.ac.uk/

²Cranfield University Library:

www.cranfield.ac.uk/cils/

³Electronic Libraries Programme:

www.ukoln.ac.uk/services/elib/

⁴Sirsi Ltd:

www.sirsi.com/Prodserv/Dma/hyperion.html

⁵Project HERON:

www.stir.ac.uk/infoserv/heron/

⁶Internet Library of Early Journals:

www.bodley.ox.ac.uk/ilej/

⁷MALIBU project:

www.kcl.ac.uk/humanities/cch/malibu/

⁸BUILDER project:

www.builder.bham.ac.uk/

⁹Research Support Libraries Programme:

www.rslp.ac.uk/

¹⁰ Arches Project:

www.rlg.org/strat/proarch.html

¹¹ Joint Information Systems
Committee:

www.jisc.ac.uk/

¹² Australian PADI gateway:

www.nla.gov.au/padi/

¹³ National Library of Canada:

www.nlc-bnc.ca/pubs/

¹⁴ National Electronic Site Licence
Initiative:

www.nesli.ac.uk/

A European model: Organisation

*Esko Häkli,
Helsinki University Library, Finland*

Europe is a multinational, multicultural and multilingual entity. Therefore, special efforts are needed to create a truly European organisation for activities which are based on contributions of individual countries and institutions. Therefore, I will in the following not present a ready-made blueprint of the future organisation, but ask questions and discuss issues which should be taken into account when drafting an organisational model.

Targets for European co-operation

Research in European culture and other European topics requires access to the sources of the European heritage and to the research dealing with it. This access should not be limited or narrowed by national boundaries. For the first time in the history of our civilisation we have been given a possibility of creating a single source of access to the distributed European resources. On the other hand, the emerging common virtual space in the Internet is creating a new type of interdependence of all players who are active in this field.

The rationale of a European model is a unified access to the collected digitised European resources, regardless of where they have been published and where they are kept.

Therefore, the aim of our efforts must be a European Virtual Library

- which is based on distributed national resources
- which is financed nationally
- which will be created in a co-ordinated manner.

In other words, the European Virtual Library will be a decentralised enterprise which, nevertheless, is efficiently co-ordinated. All actors participating in it e.g. have to accept the same technical standards, they have to pay due attention to the common European interests when selecting the titles to be added into the common pool and they have to make their reformatted resources available via a common gateway.

In order to create a significant resource which will have an impact on the use of European publications all over Europe and elsewhere, efforts have to be concentrated on digitisation on a large scale. Only the necessary critical mass will be enough to make the results attractive and useful.

The Context

Discussion about organisation of digitisation on a European scale has to take into account at least the following facts:

- a European Virtual Library has to make use of all the advantages the common digital space in the Internet can offer and to ignore the geographical distances

A European model: Organisation

- reformatting of the existing collections has to take place in individual countries which should share the costs
- access to the resources which have been created in a decentralised manner has to be provided by using a common gateway; the common access point will of course not prevent individual countries from creating their own national infrastructures
- a European programme has to be highly selective, because the variety of European journals is extremely rich and because the digitisation process is expensive.

What can be organised?

I want to maintain that all European models will be based on decentralisation. The only question is, how decentralised they must be and how much can be done together, at a European level. It is not realistic to count on financing of large pan-European projects, nor on creating a large common production facility for a number of countries. But it is important to discuss the functional goals which can be set at the European level. Several options can be drafted, e.g.:

- only a union catalogue of digitised European resources
- a union catalogue plus a well co-ordinated production scheme: development of common policy for application of standards and other relevant principles
- a common gateway to the distributed

- digitised European resources
- a European database of digitised European (full text) resources.

Of course, it is possible to proceed step by step, starting from a less ambitious goal and proceeding towards more ambitious ones, provided that a long-term policy has been formulated and adopted. To make sure that the same standards are applied, efforts have to be focused on co-ordinating the production in the participating institutions. Easy access has to be the ultimate goal and it requires not only an adequate gateway but also access to the necessary bibliographic information.

Co-ordination with other European efforts

We can not develop our goals in isolation. A meaningful discussion requires a blueprint of the development of the European Virtual Library at large. Digitised, reformatted resources are only a part of such a broader concept. Other important parts are e.g. the development of cataloguing networks and the use of electronic resources which are born digitally. A number of projects aiming at developing gateways nationally or internationally have also to be taken into account. In a common virtual space all efforts have to be co-ordinated. Due attention has, therefore, to be paid e.g. to the plans of the European national libraries aiming at creating an open European network of the national bibliographic OPACs, which are being discussed within the CoBRA initiative.

The question of permanent access in the future is even more complicated, partly because there are, so far, no adequate technical means to guarantee a permanent availability. But the permanent access is also a question of organisation. The licensing principles, which European libraries have adopted, include a strict requirement of permanent access to the resources which once have been purchased through a license. But how to arrange it? The original supplier may not be able to provide it in all eternity and we may not even be interested in it. Therefore, the question arises how to organise the permanent access to the trade publishers' electronic publications. If libraries have to assume the responsibility for that task, how would the archiving of reformatted digitised resources be organised in that connection?

Within the CoBRA initiative the European national libraries have recently started looking more closely at these issues. It is obvious that also archiving has to be arranged in a decentralised manner. As a matter of fact, the permanent archiving is closely connected with the legal deposit of electronic publications which sooner or later will be introduced in most European countries. Legal deposit includes automatically a responsibility for securing permanent access of the deposited publications. Two questions have to be considered in this connection: 1) legal deposit can not automatically provide open access to the deposited material, 2) it is not yet clear, whether digitised publications will be subject to legal deposit or not.

It is even more important to look at the arrangements which have been set up for acquisition of electronic publications by using collective licensing. Two main issues are of particular interest, conditions of access today and permanent access in the future. We will face these questions also when making our reformatted resources available. It may not be possible to grant access to digitised resources in all countries free of charge and according to similar principles. In that connection licensing arrangements may be needed. But, who will grant the licenses and who will purchase them? How can this business be coordinated at European level? What kind of relationship could this licensing have with other licenses? Take JSTOR as an example. There are a number of individual European libraries who have purchased a license to JSTOR but there are also a couple of countries who have purchased a country license. Could there be a Europe-wide license? Hardly. What will the model we are discussing here look like?

What has to be organised at European level?

Because it is unrealistic to count on a significant European financing for the efforts we are discussing here, my feeling is that the framework of a European organisation needs to be limited to the minimum. There are already a number of other examples, such as CENL, CERL, CoBRA and EROMM. Some of the projects financed together with the European Commission may also result in more

A European model: Organisation

or less permanent organisation. Among these projects DIEPER is, of course, of special interest.

What are the issues which necessarily have to be dealt with at the European level? At least the following aspects are relevant in this connection:

- **development of a European policy:** the work to create a European Virtual Library has to start from the production of digitised resources. Therefore a European policy, as part of a global policy, will be needed to formulate the goals, to draft the necessary organisation such as a co-ordinating body and to develop an overall scheme for the co-operation; different options for the structure of such a scheme can be discussed
- **standardisation:** due to the present stage of standardisation the only guarantee for permanent access to the digitised publications is the use of open standards and an adequate level of ambition in connection with the quality requirements. Standards are necessary also in choosing the identifiers such as URN and the metadata formats such as Dublin Core. An obvious task for the European co-ordination is, therefore, guidance in the application of standards. It may not necessarily be self-evident that an agreement between national and European views can be achieved easily. Different opinions can be found already about the very basic issues, because they are closely connected with the costs

- **selection** to become a truly European enterprise the programme has to develop a European profile for the selection of the publications to be digitised and not only focus on co-ordinating the results of the national programmes. The European profile must encourage the national schemes to select items of European interest into their digitisation programme. A special problem is created by the items published in minor languages. They may have a great national relevance and therefore receive a top priority in the national scheme. A truly European scheme can not only build on the publications from the great countries

- **access** as mentioned already above, the ultimate goal of the European exercise in digitising journals is an easy access to the European resources. This issue may, however, be very difficult to organise. National legislation and different policies of individual institutions may effectively prevent from choosing a unified approach. Organisation of the access is an issue even if the intention only is to create a common gateway to the distributed resources. Individual countries are already developing their own gateways or portals into their collected electronic resources. They can also be planned as subject interfaces or subject portals. Shall the digitised resources we are discussing here have a gateway of their own? What will be the advantages of such an approach? In addition to such aspects access also

requires close co-operation with other players, such as the International ISDS Centre etc.

Access to the reformatted digitised resources also requires that copyright arrangements are in place. Attractive selection of digitised resources on a large scale can hardly be created on the basis of freely available publications only. Collective arrangements is the only feasible way but it will not necessarily be easy to develop them. Ultimately the copyright issues have to be solved at the national level but it may be necessary to develop common principles in European co-operation

- **archiving:** even if the main goal of the European digitisation programme will be an easy access to digitised European publications, the question of permanent archiving has also to be tackled as has been mentioned already. But how much archiving has to be organised at a European level? Primarily the digitisation will take place at the national level and, therefore, it would be only natural also to place the responsibility for archiving on the national level. On the other hand, standards and dissemination of know-how belong necessarily to the tasks of the European co-ordinating body.

How to proceed?

To make any progress a European policy will be needed. Therefore a co-ordinating or planning body will be needed for drafting the policy and planning the first practical steps.

The policy has to be based on reliable information about the national schemes and it has to pay due attention to the national priorities and the possibilities of individual countries participating.

From a policy point of view it is important that all interested countries can participate in the planning. Of course, we can ask whether a project like DIEPER could be used as the platform for a broader European approach. Because my library is participating in DIEPER I am very much interested in such a possibility but I am afraid that we can not proceed simply by creating facts and only offering other countries an opportunity to join. A European programme can become successful only if all interested countries participate in the planning on an equal footing and if they together decide what they want to do together. A European programme has to be understood as a part of a global programme which makes European resources available all over the world.

Conclusion

The following elements, at least, have to be discussed when drafting a European digitisation programme:

- drafting a policy with common goals
- setting up a co-ordinating body
- developing and co-ordinating the level of the know-how in participating countries: common standards, high level of expertise
- creating a balance between the national and European efforts

**A European model:
Organisation**

- integration of the programme with other relevant programmes in the same virtual European space.

Esko Häkli

esko.hakli@cc.helsinki.fi

Conclusions and recommendations

*Ann Matheson
National Library of Scotland, United Kingdom*

I am sure you will all agree with me that we have had a most stimulating two days here in Copenhagen, with high quality papers and excellent presentations. We have been given a bold vision for a distributed "European virtual library of digitised materials" in a global world. We have already seen during these two days some of the foundations of the "European virtual library" through the work of DIEPER, DigiZeit and Delta: at the international level we have heard about the large-scale work of JSTOR, and I have been very impressed by the willingness of Kevin Guthrie and his colleagues at JSTOR to share their experience and their findings with us in Europe. The foundations are being laid by existing projects within Europe but there is a strong sense in the Conference that we are now at the point at which we need to develop some large-scale programmes in Europe to form a critical mass of digitised materials. We must recognise that we have a long-term mission here and we are only at the start of that mission.

My task is to pull together the main themes and the discussion during the two days of the Conference.

1. Why should we do it?

1.1. For users

The users have "great expectations": we had an inspiring paper yesterday from our user who set out the visions

and dreams of users for virtual libraries in their subject disciplines.

1.2. For the mission of libraries

- to improve access to collections
- to give parallel access to heavily-used material
- to promote the use of lesser-known collections
- to increase the visibility of "national" materials in the global world.

1.3. For the material

- although primarily an access-related tool there are preservation dividends.

1.4. In response to external factors

- funding bodies and paymasters may require it either for prestige or through need.

2. What factors must we concentrate on?

2.1 Care in the selection of content

- national libraries

will be concerned with establishing national policies

will be driven by cultural imperatives or cultural mandates

- university libraries

priorities will be research-orientated

All decisions must be related to the Library's mission: and it should be noted that the traditional

Conclusions and recommendations

mission of libraries is now being subject to change.

2.2 Careful considerations of access requirements

2.2.1. To meet users' need

- must be *open*
- must ensure *compatibility*
- must involve *users* in some of the decisions
- must ensure access to "old data" through archiving
- must try to find a way to ensure that access to knowledge is not determined by economic power.

2.2.2 To determine the library's mission

- to provide *easy* and *continuing* access.

3. What does this involve for libraries?

3.1 Consensus on common technical standards

- image scanning to a minimum of 600 dpi
- metadata created in XML/RDF
- full text created by OCR in XML/TEI
- unique identifier to locate the full text (URN or DOI).

3.2 Cost considerations

- economic factors are significant (actual costs, how to secure the money) but it is important to balance cost considerations with the value of what is created
- there is already existing experience on which to draw in JSTOR and DIEPER, etc.

3.3 Legal issues

- There are copyright and rights issues to be considered
- JSTOR has valuable experience in developing relationships with publishers on which to draw.

3.4 Digital documents must be managed

- essential to use document management systems
- avoid duplication by checking the European Register of Periodicals
- consider "free" access to EROMM to assist with identifying if a journal has been or is going to be digitised; and to encourage notification of decisions to digitise journals to the Register.

3.5 Must maintain permanent access

- sustainability - no single or simple path
- long-term archiving

archiving digital data is different to paper

archiving digital data is linked to other issues, e.g. standards

collaboration is essential.

4. A "shared activity" for libraries

- independent and interdependent

we must share decisions on selection

we must share some decisions with users

we must share information with one another

we must share our expertise

we must share our tools - we should not reinvent tools in Europe where they have already been well made in North America

we must share the task with publishers

we must share responsibility for maintenance and for long-term archiving.

5. Mottoes

Think it through!

Work local

Work national

Work European

Work international

Be prepared - the unexpected always happens!

Ann Matheson

a.matheson@nls.uk