LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich, Printed in Germany

Digitisation for Access to Preserved Documents

by MAJLIS BREMER-LAAMANEN

INTRODUCTION

Today the digitisation of our collections is a goal for libraries all over Europe. The choices we make in digitisation and preservation now will have a significant impact on the future. Do we only emphasise access? How do we enable access and preserve our originals in a qualitative and productive way? What will actually be left of our cultural heritage in the next millennium?

In this paper I am going to look at promoting access to preserved originals mirrored by the experience at the Helsinki University Library, the National Library of Finland:

- Preservation activities as platform for digitisation and OCR
- Processing access to collections
- The future – looking ahead

PRESERVATION ACTIVITIES AS PLATFORM FOR DIGITISATION & OCR

For a long time microforms have facilitated access of remote collections to library users. Large reformatting programs have in Europe and USA been the action of saving brittle paper collections, like books and newspapers, printed on acidic paper. In the 1990s the preservation discussion has gradually turned to digitisation matters. Anne R Kenney wrote in RLG DigiNews in 1997 that microfilm besides being a preservation medium for "slow-fires" of acidic collections "…could become an important legacy measure for coping the "fast fires" of digital obsolescence" (Kenney, 1997). The discussion of using microforms as platforms for digitisation is today a relevant issue. We have all heard about the Brittle Book projects in USA (Chapman, 1999; Conway, 1996). In these projects hybrid approaches were tested: digital images were transformed to microfilm and microfilm was used as a platform for digitisation.

Recently both IFLA Newspapers Section and the Research Libraries Group (RLG) have given guidance on microfilming for digitisation. IFLA Newspapers Section has published *Microfilming for Digitisation and Optical Character Recognition* in 2002 as a supplement to the microfilming guidelines. The RLG Guidelines for Microfilming to Support Digitization, Supplement to RLG Microfilming, was published in January 2003.

There is hence an ongoing interest in microfilm as a safety platform for preservation. When again we deal with collections printed on more sustainable paper conservation and preservation measures are needed to prolong the lives of the originals. These measures can be carried out as a part of the digitisation process. Preservation and digitisation programs can successfully collaborate. Libraries and archives have the major responsibility for digitising and preserving their collections. That is why we have to have a digitisation and a preservation program - a strategy for our users of today and tomorrow. We do also need to know in which condition our collections are to make the final choices within our plans.

THE NATIONAL LIBRARY

The National Library of Finland – Helsinki University Library, was established in 1640, and has had Legal deposit right since 1707. It is carrying out most of its preservation and digitisation services in the Centre for Microfilming and Conservation, which was established in 1990 in Mikkeli, 220 kilometres from Helsinki. As the responsibility of Library's digitisation services are here since 1999, the work for all projects, programs and co-operation activities on preservation and digitisation are organised and processed in the Centre, with over 40 employees.

Digital Program for Libraries

The Library Sector – co-ordinated by the National Library - has in 2002 accomplished a Digitisation Program for the cultural heritage. The target for digitisation starts with the copyright free material from the Middle Ages to the end of the 19th century. Within this field the 19th century is prioritised because of the demand of the material. Access to Historic Newspapers, journals, books and ephemeral print are of great importance. Also older collections were considered nationally and internationally important. Our lost cultural heritage due to a disastrous fire in Turku in 1827, could be digitised from Swedish collections. Dissertations from the same era have a great demand, as well as the special collections in the participating libraries. Like the Nordenskiöld map collection, part of the Memory of the World Register of UNESCO. The National Library will promote copyright solutions for access via digitisation of more recent materials.

Preservation program in the National Library

Microfilming is playing an important role in our preservation program. Newspapers have been filmed on a large scale since 1951. Today, since 1997, all current newspapers including local and some free newspapers have been microfilmed. Books have been filmed from 1966 onwards primarily on microfiche. Retroactive microfilming of

newspapers, books and journals is part of our preservation program. The chosen books consist of unique titles of Finnish literature from the era before 1810 and of more recent literature from the 19th century. The preservation program includes also other preventive actions for current legal deposit material like boxing, protective enclosures, binding and retrospective measures like conservation for the most unique material already damaged in use or storage.

Condition surveys

After having established the goals and priorities in our digitisation and preservation programs we need to make the actual choices. In order to do this we do need a condition survey to establish the condition of our collections. These surveys should be based on standards or internationally accepted methods in order to be comparable internationally. As microfilms are used as a platform for digitisation, a condition survey was done in 1997 based on a random sample of 500 microfilms of 40.000 films. The period included all films from 1951-1996. [1]

It was found that more than half of our older films needed re-filming. Most of the material would have been in a good shape if:

- newspapers were filmed unbound or bound volumes were filmed on a book cradle under a glass according to existing standards
- there had been a separate master for duplicating copies in early times
- the density readings were acceptable

These findings correlate with the IFLA Guidance for microfilming for digitisation and OCR, which adds the importance of a low reduction rate for maximum quality. The condition survey of our paper collections began at our library in 2002 and is continuing. It is based on the international surveys made in US (Stanford and Yale), Sweden and Holland. A random and stratified sample of 3.700 books from our scientific and literary collections from 1810-1944 was chosen out of 150.000 books. The results will have a great impact on digitisation and preservation priorities and co-ordination within the programs and conservation measures can be directed to specific problems.

PROCESSING THE ACCESS TO COLLECTIONS

Helsinki University Library had in 1997 chosen digitisation of all newspapers as one of the main targets in its reformatting program. And as practically all newspapers in Finland

and the other Nordic countries have been microfilmed, it was possible to use microfilm as a platform for digitisation needs. This resulted in the

Nordic project Tiden on Historical Newspapers, 1998-2001, co-ordinated by Finland in which the Nordic national libraries from Sweden and Norway, and the Århus State Library in Denmark participated, supported by the Nordic Council of Scientific Information, the NORDINFO and the participating libraries. The National Library of Iceland, together with Greenland and the Faroe Islands was in close cooperation with the project.

The Objectives of the Nordic Project

The main objective of the project has been to use microfilm as an intermediate for future digitisation. Microfilm enables large-scale production of digital images. In Norway and in Finland we have built production lines for digitisation of newspapers from film. The objective has been to start with a part of the newspapers and when this project is finished, integrate the digitisation of the newspapers to the libraries' ordinary functions. In Denmark and Norway the objective has been to get as much of the collections on the web as possible – with minimal search possibilities. That is search on title, date and place.

In Finland and Sweden we have included full text search, which enables search on each word in the textual content. In addition in Finland an index of articles with hierarchical search terms from 1771-1890 will be available.

The Newspaper Contents

The contents in the Digital Newspaper Library of Tiden have been chosen according to the importance of the newspapers and copyright possibilities in the member countries. Kungliga Biblioteket chose to digitise Post-och Inrikes from 1640-1721 in Sweden. Some predecessors to this actual first newspaper have been digitised from 1620-1630s. This means that information starting from the 30-year war until 1721 in Europe is included and available today for full text searching. In Finland we chose to build a digital platform for the day to day life in the 18th and 19th centuries. In this project every newspaper, containing 44 titles, is digitised for full text searching from 1771-1860, starting with our first newspaper *Tidningar utgifne af et Sällskap I Åbo* in 1771 containing 140.000 pages. This project has been continued by a second project, which will include all newspapers until 1890, with 800.000 pages more. Denmark continues in the Tiden project, almost where Sweden stopped, with *Adresseavisen* from 1751 – 1890. Shorter periods of other titles have also been included, like *Berlinske Tidende* 1863-1865, *Dannevirke* 1862-1864 and *Faedrelandet* 1863-1865. Norway is covering almost

two centuries, the 19th and 20th century with *Den norske rigstiden* 1815-1882 *Adresseavisa* 1802-1900 and *Norlands Avis* 1893-1978.

The number of pages we have on the net at this point is actually of minor importance. More value lies in the development and innovations we have made to build the production line, for digitisation and full text searching, for our large library collections. As I know the Finnish project best, and as it is the most complete I will introduce our production line to you.

PRODUCTION LINE

The individual steps of the process are as follows:

- Microfilming: refilming the newspapers if the quality of the present microfilms is not good enough.
- Digitisation: scanning of the microfilms; the scanner software has been made to identify the varying page images and the films are scanned automatically.
- Optical Character Recognition (OCR): conversion of the images to text files; requires many adjustments and training of the software especially for text in Gothic letters.
- Identification: identification of title, issue, date, pages and attachments of the newspaper takes place in principle automatically but requires some human treatment.
- Database import: importing the prepared material into the database system

The original

The quality of the original is essential for all reformatting activities. Old books, maps and newspapers are often not of the best quality and cleaning, repair and conservation measures might be needed. Older newspapers are small in size continuously growing to the size of A1. Problems are encountered as the paper is turning yellow in the 19th century and many types of fonts and languages are used on the same pages in one newspaper.

Microfilming

When using microfilm as intermediary the quality of the microfilms is the key to the success. Of course, the quality is crucial also for the preservation in general because we have to accept the simple fact that in the long run newspapers and acid books will

survive only on microfilm. Digital long-term preservation is still sought for. We are always using the smallest reduction ratio possible for microfilming, not exceeding 16 or 18 x. The size of the smallest e-letter influences the OCR- of Gothic text as much or more than the reduction ratio.

Digitisation

Most of our digitisation experiences via microfilm are from old historical newspaper collections. We are using black and white, 400 dpi, 1-bit digitisation for the OCR-process. Too sharp images might influence the OCR- of Gothic text in a negative way as the letters become more individual and needs more teaching time. We are thus dealing with material where compromises have to be made. (The newspaper image files would be too large to handle smoothly in greyscale. And as the microfilm has a limited dynamic range some of the problems diminish in black-and white, and some, like the quality of photographs, are getting worse. But with the text and OCR in mind some of the bleed through in the papers is extinguished and the result is even better than direct scanning from the newspaper.) The digitisation process is automatic. Very seldom changed density in the films require adjustments by the operator. In a production perspective scanning from film is much faster than using the originals when oversized items like newspapers are concerned.

Optical Character Recognition (OCR)

Because of the majority of Gothic letters in our newspapers, the OCR conversion requires much effort. We are using Fine Reader by ABBYY for the OCR, which also provides support for Finnish and Swedish languages. It is not too sensitive to react too much on the unevenness of the quality of the original pages and is able to process the text in batches after training. In our experience the most important factors influencing the OCR quality of the conversion are the text font, language and reduction rate. Roman style, Swedish language and a low reduction rate gave the best results in our newspapers. At the moment we have had OCR read 250.000 pages. For a newspaper with acceptable print the identification rate exceeds 95 %.

Identification

The last point in this process, the entering of each image and its textual version should aim at automation when dealing with large quantities of material. Usually, every image has to be addressed on behalf of the title, date, number and page. We have built software for the identification of large quantities of images. By indicating the title and ISSN-number of a publication this can automatically be used for all chosen newspaper numbers. Each newspaper number, when chosen, gets in turn automatically correct page numbers.

When the identification step has been completed the software creates an XML file that is copied to the server along with images and text files.

THE ENGINE ROOM

Much effort has been put into developing a proper functional architecture for the digital newspaper library, as a part of an overall digital architecture. This will include all kinds of materials and the production of the digital versions as the digital original, the digital master and the digital image on the web and the textual versions for full text searching. URN identifiers are generated automatically. Solutions have to be applicable for other types of material such as maps, periodicals and books. The database will also be connected by search and retrieval protocols to the National Portal and database of the Scientific Libraries in Finland, enabling fuzzy search, vocabularies and indexes to older digitised collections.

Text search software

From the very beginning it was obvious that a hundred per cent OCR conversion is impossible. That is why two decisions were made. The first one was that the basic tool for the users of the digitised newspapers is the digital facsimile of the original pages. The ASCII version will be used for searching purposes only.

The other decision was to use RetrievalWare from Excalibur, software that could manage a limited percentage of errors. This is also needed for the retrieval of the old fashioned language used in the newspapers. It has among other features a so called fuzzy search function which is able to identify the searched words even if one to three letters would differ from the word sought for – because of misspelling or old-fashioned language. (It processes the words as bit-strings and uses pattern recognition to find matches. Therefore RetrievalWare was chosen. Finnish and Swedish language support has been added to the normal search functions.)

THE FUTURE – LOOKING AHEAD

The Nordic Digital Newspaper Libraries were opened to the general public on 25th of October 2001. By the end of 2002 the total searches of pages rose to 1 million. Systems have been running without problems since the launch. We have received plenty of user feedback on the system. Most of it has been positive along with some error reports and improvement suggestions.

Access is available by:

- online browsing
- title search
- search by date (in which you can actually see the day of the week)
- free text search via fuzzy search function
- article search by indexed words

The Nordic countries are still working together to enhance full text fuzzy search possibilities between the Nordic newspaper databases including the new co-ordinator Iceland. This pre-investigation has been funded by the Andrew W. Mellon Foundation and will be ready in spring 2003. Co-operation is and has been successful. A broader access to our digital strategies and results will in the future be gained in the EU Minerva project (a Ministerial Network for Valorising Digitisation Activities in the EU Member States), a network that promotes sustainable access to our cultural and scientific heritage. Institutions can present their digitisation policies, programs and projects and obtain good practices [2]. You are welcome to participate. Co-operation is the platform for sustainable access.

NOTES

1. The survey was based on the ANSI/AIIM MS45-1990 standard: Recommended Practice for Inspection of Stored Silver-Gelatin Microforms for Evidence of Deterioration and on the Finnish standard, which is in accordance with the former: SFS 5808: Inspection of the Stored Silver-Gelatin Microforms for Evidence of Deterioration 1998.
2. Username: mindemo – passwords: mv1731

REFERENCES

1. Kenney, A.R.: "The Cornell digital to microfilm conversion project: final report to NEH". *RLG Diginews* 1, 1997. http://www.rlg.org/preserv/diginews/diginews2.html#com
2. Chapman, S., P. Conway & A.R. Kenney: *Digital Imaging and Preservation Microfilm: The Future of a Hybrid Approach for the Preservation of Brittle Books*. Washington, DC : Council on Library and Information Resources, 1999 (http://www.clir.org/pubs/archives/hybridintro.html), and Conway, P. *Conversion of*

Microfilm to Digital Imagery: A demonstration Project. Performance Report on the Production Conversion Phase of Project Open Book. New Haven, CT : Yale University Library, 1996.

3. *Microfilming for Digitisation and Optical Character Recognition*, 2002

4. Dale, Robin L. *RLG Guidelines for Microfilming to Support Digitization*. RLG, 2003. http://www.rlg.org/preserv/microsuppl.pdf

WEB SITES REFERRED TO IN THE TEXT

The Andrew W. Mellon Foundation. http://www.mellon.org/

Helsinki University Library, The National Library of Finland, http://www.lib.helsinki.fi/english/infoe/index.htm

IFLA Newspapers Section. http://www.ifla.org/VII/s39/snewsp.htm

Memory of the World Register of UNESCO. http://www.unesco.org/webworld/mdm/register/index.html

Minerva project. http://www.vilmamedia.fi/minervaeurope

NORDINFO. http://www.nordinfo.helsinki.fi/index_eng.htm

Research Libraries Group (RLG). http://www.rlg.org/

Tiden - A Nordic Digital Newspaper Library. http://tiden.kb.se