World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

toukokuu 23, 2005

Code Number: 019-E
Meeting: 97 Newspapers

Connecting to the past – newspaper digitisation in the Nordic Countries

Majlis Bremer-Laamanen
University of Helsinki, Mikkeli, Finland

Newspaper collections are targets that have a great demand on the Internet from researchers as well as the public at large.

Our digitised historical newspaper collections and the born digital ones are connecting the users to places, questions, nations and human life over centuries. Incidents from the past are suddenly easily accessible. The past is living in the present.

Today I will talk about newspapers in Finland. I will also give an update on the Historical Nordic Newspaper Project – one of the pioneers in digitising historical newspapers.

Newspapers

Newspapers are perhaps astonishingly, still the most important media in Finland. The Norwegian national library tells us that Norway is the country with the largest newspaper reading population. We can say that newspapers have a very high status in all the Nordic countries.

All over the world newspapers have adjusted to new sources of media like radio, movies and television. Today they face “ghosts” like the Internet. How much will the new Internet behaviour interfere with the life of today?

Way of life in Finland

The average use of newspapers in Finland is almost an hour per person per day and it has not diminished. [1]
The daily paper is a way of life. It is delivered to your front door in the morning, to be enjoyed with a cup of coffee or tea. It gives the reader a moment of peace and comfort together with the national and local news before the day starts. Free newspapers are delivered on the subways and trains on your way to work.

Media channels

The look and feel of many newspapers have changed in appearance to a modern outlook, tabloid format, actively reaching out to its customers and the youth at schools, delivering information, science, leisure and advertisements. Surprisingly perhaps, the heavy users of computers are heavy newspaper readers. Over ninety per cent of young Finns read a newspaper each week.[2]

Newspapers are the media channel that daily is best reaching the Finns. Hence, over half of all money spent on advertising is in the newspaper business. It is for example the major channel for information on stock exchange today.[3]

Newspapers are the most important research media for research in Finland. They are the prime source of investigation in more than half of the research projects in Finland and almost in half in Sweden. They are used as source material for research in the fields of media, history, political science, sociology, pedagogy, art, business, natural sciences and technology.[4]

As such the interest in our historical newspapers is high. Newspapers are the most used individual group of the Finnish national collection.

The role of online newspapers

Online newspapers are produced all over the world, also in the Nordic Countries. These papers are usually available via the National Libraries. We have about one hundred newspaper titles on the web among the 900 periodicals in Finland.

Surprisingly the role of the Finnish online paper is to support the paper version. Thus their monetary importance is still quite low. These papers are improving and the overall development in knowledge society will influence their use. Also paper look-alike editions of newspapers are available on the web in Finland since 2002.[5]

Converting large volumes today

[1] www.sanomalehdet.fi/en/tietoa/index.html, page 2
[2] www.sanomalehdet.fi/en/tietoa/index.html page 3
[3] Hufvudstadsbladet, 2005, April ; www.sanomalehdet/en/tietoa/index.fi, page 3
[4] www.sanomalehdet.fi/suomenlehdistö/fi
[5] www.sanomalehdet/en/tietoa/index.shtml, page 4

John S. North´s description of newspapers is relevant and sheds some light on the reasons for the rather late start for newspaper digitisation projects.

> Periodicals bibliography is a much neglected field, for understandable reasons. First, it is massive: periodicals easily outdo monographs in sheer volume of publication. Second, no clear definition of a periodical is generally accepted, and the working definition varies from library to library. Third, any one periodical is likely to change in some of its primary bibliographical elements from issue to issue (title, subtitle, format, editor, publisher, proprietor, frequency, printer, size, etc.). Moreover, periodicals are often considered ephemeral: stale news, cheap popular information, trivial records. In short, throw-aways. They are often poor quality paper, arriving in libraries unbound and in endless irregular succession, so are unwieldy to shelve and catalogue, and are seldom to be found in complete runs, seldom well indexed. They are the nightmares of librarians and bibliographers.
>
> John S. North

The Waterloo directory of Scottish newspapers and periodicals 1800-1900.

The newspaper holdings have been considered a nightmare for digitisation as well. The size of newspapers grew in the late 1900th century to four times the size of a tabloid of today. The poor print, the poor quality paper and the use of Gothic Fraktur and Roman text in the Nordic countries made a challenge for digitisation. Even more of a challenge was the Optical Character Recognition needed to make the text searchable from the image of the paper.

Digital newspaper projects are a new hot topic in Europe and around the world. The British Library is going to digitise and give free text search to 2 million pages. Austria and Estonia are well under way with their newspaper projects. The United States (The Library of Congress) is planning newspaper projects. So is also the National Library of Australia.

The Nordic Historical Newspaper Project – TIDEN

One pioneer in the field is the Nordic TIDEN project, starting in 1998 and launched on the web http://tiden.kb.se in 2001.

The libraries participating were the Royal Library of Stockholm, the National Library of Norway and the State and University Library of Århus. The coordinator for the project was Helsinki University Library, the Centre for Microfilming and Conservation.

The aim of the TIDEN-project was:

- to test criteria for microfilm as a platform for digitisation and full text search
- to build production lines for the digitisation of newspapers
- to integrate the digitisation to the libraries ordinary functions
- to give a continuous widening access to newspapers out of copyright

Today the amount of online pages has risen from 400.000 at launch to 1,6 million. In Finland, Sweden and Norway full text search is available.

Automation

When dealing with large collections like newspapers the production has to be as fully automated as possible. This was one of the aims when TIDEN started. The possibilities to do so are far better today than five years ago. Today we are changing our half automated processes to a faster automated process. We are working together with the Royal Library of Sweden to test and enhance their line and our production line. The Centre for Microfilming, Conservation and Digitisation of the Helsinki University Library – the National

Library of Finland is situated in a smaller town Mikkeli in Eastern Finland. The Norwegian National Library in Mo and Aarhus Staatsbibliotek will also be able to follow up our results.

The process

The first step in the production line is the digitisation of newspapers, from microfilm or from the original. Newspapers have been a main target for the reformatting programmes in Finland, Sweden, Norway and Denmark since 1950. This makes it possible for us to use microfilm as intermediary for digitisation if the film quality is high enough.

The process of digitising newspapers includes

Microfilming	- refilming the newspapers if the quality of the present microfilms is not good enough
Digitisation	- scanning of the microfilms; in black and white or grey scale
OCR	– conversion of the images to text files; requires many adjustments and training of the software.
Identification	- of the title, issue, date, pages and attachments requires some human treatment
Database import	- by a separate software

IFLA Guidance

When using microfilm as intermediary the quality of the original newspaper and the microfilm is the key to success.

Some advice is given in “Guidance on the best practice for microfilming of newspapers in preparation for possible future digitisation”. 2003. English, French, Spanish and Chinese versions are available on the IFLA-net.
The Guidance was based on the information gathered within the TIDEN project.
Information is also available on the TIDEN web-page at http://tiden.se.

Goals today

Our goals in Finland are to require:

- an industrial production environment
- an automated optical character recognition (OCR) of both Fraktur and Roman text even within a page
- highest possible quality
- highlighting of search words on the newspaper page
- cost effectiveness
- speed of the process
- xml-METS-standard (Library of Congress)

The results of the OCR- conversion in Finland and Sweden have shown that there are several factors influencing the quality of the conversion, the most important of them being the text font, language and reduction rate. From the very start of the TIDEN-project it was obvious that a hundred per cent OCR conversion is impossible. Due to the old language and the great mass of text proofreading was not our way to enhanced search. It was thus decided that the ASCII versions of the text would be used for searching purposes only.

The basic tool for the users was the digital facsimile of the original pages. A retrieval ware with fuzzy search possibilities was chosen in Sweden and Finland to identify the search words even if one to three letters would differ from the word sought for. The search tool process the words as bit-strings and uses pattern recognition to find matches.

The speed and automation of the production environment is essential when comparing the process 5-8 years ago to the possibilities today. The automated processes offer the coordinates to each word in the paper. Thus users get better service as the words are highlighted. Other improvements are available. Images are of a better quality as the microfilm scanners and the OCR-software are able to handle greyscale images automatically. Previously, Roman text had to be especially trained for the OCR-software. Now we are looking at a breakthrough where the text is interpreted automatically.

There are interesting vendors offering reasonable production automation environments on the market. Helsinki University Library and the Royal Library of Sweden tested one of them in 2004. The test period gave us very good information about the key issues for the new production line when carrying out a request for tenders.

If a suitable production environment could be established, we would like to use it for other digitised collections as well.

Connecting the past

In digital futures the digital collection management will be an exiting challenge. It is not enough to digitise our collections and put them on the web with enhanced search possibilities. The infrastructure around our collections is changing fast as well as the expectations of our users. We have to improve access in various ways, including automatic translation to our old holdings, including meeting places for users. We have to connect them to other people, to holdings in other heritage sectors and so on. We do have to make the decisions on the roles of libraries. We have to decide which services are free and which are not.

Today the Historical Newspaper Library in Finland has found its place. It is highly regarded among researchers especially in the field of history and languages. It has also found its way to the public at large via information in the media and from person to person. Many a person is doing some genealogy with the help of the newspaper library from home. Finnish immigrants are using the library, finding information on their families leaving Finland, following them to their new country. Last year 150.000 visits were recorded and 1,8 million pages were searched for. Still, we have to extend the use to education in schools.

People are very enthusiastic about the library. It is easy to use in its Google-like approach. The users want more and we will try to meet their needs.

One future aim has been to coordinate search to the Nordic historical digital newspaper library. Perhaps also that could be a nice challenge for the future.