Making sense of it all - combining digitized analogue collections with e-legal deposit and harvested web sites

Pär Nilsson

National Library of Sweden

History and collections

- Legal deposit since 1661
- First Swedish newspaper 1645 (Ordinari Post Tijdender)
- Printed collection of newspapers: about 122 million pages (parallel collection at Lund University Library
- Until 1979 legal deposit copies of newspapers to the NLS and 3 university libraries,
- Since 1979 only two copies preserved + one copy used for microfilming

National Library of Sweden

Microfilming for preservation and access

- Microfilming of all Swedish newspapers since 1979 and all major newspapers since the 1950s
- 20 % of the collection available on microfilm by 1979; today 60 %
- Complete collections of all microfilm since 1979 at the NLS and four university libraries
- Smaller collections of microfilm at 70 public libraries, usually local newspapers
- Microfilming of historical newspapers 1983-2008; about 10 million pages mainly mid size or small regional and local newspapers
- Poor technical quality due to insufficient quality control, but important for availability

National Library of Sweden

Digitisation and access – Tiden project

- NLS participated in the Tiden project (1998-2001, w. Denmark, Finland and Norway)
- Result: very manual and small scale digitisation of some of the oldest Swedish newspapers.
- Some of the material typed in due to poor OCR from low quality microfilm and difficult Gothic fonts
- Published using Convera RetrievalWare search engine, but raw interface
- Quite widely used, despite small size collection and access problems
- Not available in the present interface, but necessary to restore in the next system

National Library of Sweden

Digitisation and access – TELplus (1)

- In the Tiden project focus on scanning.from microfilm
- Microfilm collections from 1950 to 1980 (all major Swedish newspapers) very difficult to use for acceptable OCR result
- NLS decided not to use this film, but scanning the printed newspapers still very expensive and slow
- Microfilm collections from 1980- also too uneven in quality
- But some of the technically better microfilms used in TELplus project (2007-2009; WP1)

National Library of Sweden

Digitisation and access – TELplus (2)

- Result: usable images, bad OCR
- For copyright reasons only material up until about 1920s
- Titles chosen was a mix of more well known older newspapers and small local papers
- Very popular despite unorthodox selection of titles and small volume (about 200 000 pages), especially among family historians
- Smaller local newspapers should be included in larger projects in the future

National Library of Sweden

The Digidaily projects (1)

- NLS not to develop large scale digitisation facilities of its own after the TELplus project
- Some in-house projects including ambitious digitisation of all 5600 Swedish Government Official Reports 1922-1999
- Large format and large scale projects with millions of pages to be outsourced
- The Swedish National Archives had many years of experience from digitising church records, maps, etc. at the facility in Fränsta in the middle of Sweden, 430 kilometres north of Stockholm
- EU funding granted for the initial three year Digidaily project (April 2010 to March 2013) and the present one year project (April 2013 to March 2014)
- Development project to develop efficient methods and processes for digitisation of paper originals and routines for future cooperation both between the institutions

National Library of Sweden

The Digidaily projects (2)

- Funded by: the EU Structural Fund for Central Norrland, the National Archives, the National Library, Mid Sweden University, the County Board of Västernorrland, and the newspaper publisher Schibsted Sweden
- During the first project methods and processes were tried out to lower the cost per page while maintaining an acceptable quality
- Estimate for the bulk of the newspaper collection including OCR: from around € 0.25 (glued or stapled but not bound) to € 0.40 (bound volumes taken apart)
- Three important factors: fast and sheet feeding duplex scanners, segmentation/OCR without any manual work, and a second copy of all Swedish newspapers between 1850 and 1978

National Library of Sweden

The Digidaily projects (3)

- Second Digidaily project focused on current newspapers w different editions, supplements, news bills, colour printing, varying paper qualities, etc.
- Investigate the use of high quality digital cameras instead of over head scanners to further lower the price per page
- Development projects with a lot of production:
  Digidaily 1 - 2.5 million pages from Aftonbladet 1830-2010 and Svenska dagbladet 1884-2010
  Digidaily 2 - 2.5 million pages from Dagens industri 1983-2010, Dagens nyheter 1864-2010 and Expressen 1944-2010
- The Digidaily projects use JPEG2000 and METS/ALTO with article segmentation so not possible to the current interface
- New system needs to handle a variety of formats

National Library of Sweden

Digitisation of current newspapers

- The best way to collect, preserve and present today's printed news in digital form? Probably the original PDF-files used in production or processed versions of them.
- NLS in negotiations with Swedish newspaper publishers about agreements, but PDF-files used for printing are no e-legal deposit material.
- The plan is to change the production from microfilming to digitisation of not only the major newspapers, but also all the smaller local newspapers.
- 2-3 million pages per year of current newspapers will keep the present production line alive, with the possibility of adding historical newspapers when there is funding.

National Library of Sweden

Newspaper web sites – harvesting

- NLS is harvesting Swedish all web sites 2-3 times per year since 1997 and Swedish newspaper web sites on a daily basis since 2002
- Newspaper harvesting greatly expanded in 2004 to include about 140 newspaper titles, both large national newspapers and local papers
- The Swedish web archive is only available at the National Library on two computers without connection to the Internet
- Regulated by "Ordinance (2002:287) concerning the processing of personal data in Kungl. bibliotekets digital cultural heritage projects"
- NLS is permitted to collect and store the Swedish "national digital cultural heritage", including all material which can be classified as Swedish on the grounds of "address, addressee, language, originator or sender"

National Library of Sweden

Newspaper web sites – e-legal deposit

- Web harvesting gives the context and the look and feel of the web site
- But web harvesting is only a snap shot and makes it very difficult to capture all articles on a newspaper web site, including updates of articles
- NLS is still harvesting, but will use customized RSS-feeds to capture all new articles published
- Some development work for the newspapers, but no need to archive articles for delivery to the library
- The best possible, if not the complete, representation of Swedish newspaper web sites: the web page as it is harvested once a day together with the individual articles collected by the library through e-legal deposit

National Library of Sweden

Access – copyright legislation and the Personal Data Act

- Two large obstacles for successful use of newspapers in digital form in Sweden: the copyright legislation and the Personal Data Act
- The copyright legislation for newspapers is the same as for books and other publications, i.e. the work is protected for 70 years after the death of the author
- NLS maintains a very strict policy: the library considers only material before 1863 to be completely free of copyright (article written by a 10 year old in 1863, who may have died at the age of 90 in 1943
- An extreme interpretation, but necessary in upcoming negotiations with copyright organizations later this year, concerning collective licenses (change in the Swedish copyright legislation)

National Library of Sweden

Access – the Personal Data Act

- The Personal Data Act in Sweden concerns only living persons, so a possible limit is 100 years
- The Swedish web archive has permission to store and make available even current material containing possibly sensitive information about living persons
- Hopefully this more liberal view on personal data will be the one used also for digitised material

National Library of Sweden

Access – user interface

- No access yet to the digitised material in the Digidaily projects, due to focus on workflow and cost
- Decision earlier this year to develop a solution in house, using the library's long experience from developing its own search interface for the Swedish national catalogue Libris
- The goal is to build a solution that can be adapted to different types of material, not only digitised historical newspapers
- Much of what we will digitise and collect in digital form will have the strong temporal and geographical aspects typical for newspapers; especially true for the online material harvested and delivered through e-legal deposit, where not only the date of publication is important but the precise time, when it comes to developing news stories

National Library of Sweden

Access – one interface for everything?

- In a few years time NLS will have:
  - millions of digitised pages from both current and historical newspapers
  - harvested and e-legal deposit delivered web material
  - a vast collection of sound and moving images in digital formats (since the Swedish audiovisual archive is now a part of the NLS) containing a lot of news material
  - other types of digitised print material (books, journals, ephemera, etc)
- A combined interface for all these types of material will be an excellent tool for all kinds of research concerning Swedish history, society and culture on both a national and local level

Thank you!

www.kb.se

magasin.kb.se

digidaily.kb.se

Questions and suggestions to

par.nilsson@kb.se

National Library of Sweden