Making sense of it all: combining digitized analogue collections with e-legal deposit and harvested web sites - Pär Nilsson, The National Library of Sweden

The National Library of Sweden (NLS) has collected newspapers since 1661, but has of course also in its collection what was published between 1645 and 1661 of one of the oldest newspapers in the world, Ordinari Post Tijdender, which is still continued in the form of a database at Swedish Companies Registration Office. The printed collection of newspapers contains about 122 million pages. A parallel collection is kept at Lund University Library in the south of Sweden. Until 1979 legal deposit copies of newspapers were sent to the NLS and three university libraries, but through a change in the legal deposit legislation in that year only two copies are preserved and one copy is used for microfilming since then.

Microfilming for preservation and access

Microfilming of all current Swedish newspapers has been done since 1979 and of all major papers since the 1950s. Newspaper microfilming in Sweden was originally a private initiative, done in cooperation with the newspapers themselves, but library copies were used for the microfilming of the non-current newspapers. About 20 % of the newspaper collection at the NLS was available on microfilm by 1979 and the collection has been kept at an off-site facility 50 kilometres from Stockholm since 1960. The new legal deposit legislation in 1979 was combined with a decision in parliament to let the NLS microfilm all current Swedish newspapers, including all editions, supplements and news bills. Almost 60 % of the collection is now available on microfilm. Complete collections of all microfilm produced since 1979 and of the older microfilm are kept at the NLS and four university libraries. There are also smaller collections at about 70 public libraries, usually of the local newspapers.

Microfilming of historical newspapers at the NLS was started in 1983 and discontinued in 2008. Most of the larger newspapers had already been microfilmed retroactively before 1983. This made it possible to work with smaller but nevertheless important newspapers. About 10 million pages were microfilmed, mainly from mid size or small regional and local newspapers, sometimes in cooperation with local interest groups or libraries. Even though the technical quality of the microfilm is often poor due to insufficient quality control, this microfilming programme and newspaper microfilming in general has undoubtedly made available a lot of material that is appreciated by local historians and genealogists as well as other types of researchers and the general public. Without the help of microfilm the library could not have made newspapers as available as they have been, without any worries about the preservation of the collections. Some of the lessons learned from the retrospective microfilming project are important to bear in mind in the selection of material for digitisation.

Digitisation and access

Even though microfilming has been an invaluable tool in the preservation and dissemination of newspapers, today it of course also has many limitations. It has become increasingly more difficult to microfilm modern newspapers in black and white, when so much of the layout depends on colour. As a result the NLS decided a couple of years ago to concentrate on the text and make sure that the density of the microfilm is right for this part of the content. Images, graphics and text in colour may or may not be rendered in an acceptable manner. For certain parts of the newspapers, especially supplements with a lot of print in colour, the library accepts density values well outside the limits in the standard for microfilming,

otherwise the text in this type of material would be more or less inaccessible and microfilming would be pointless.

In addition to the technical difficulties, the NLS is also aware of the changes in the film industry and the availability of both microfilm and the equipment necessary for production and use of the film. The microfilming is done for the library by an external company and through these business contacts it is clear that the market is diminishing and that the price of microfilming can only go up.

Newspaper digitisation in Sweden has had a long and slow start. The NLS participated in the nordic Tiden project (1998-2001) together with Denmark, Finland and Norway, but the direct outcome of this participation was a very manual and small scale digitisation of some of the oldest Swedish newspapers. Some of the material proved to be impossible to get acceptable OCR results from and was actually typed in by an external company. The digitised collection was published using the advanced Convera RetrievalWare search engine, but the interface was not adjusted to what was needed and difficult to use. Despite the small size of the collection and the access problems it was quite widely used and is still missed after the final server crash a number of years ago. The image files (TIFF) and the original text (in MS Word!) files are preserved but have not yet been migrated to another system. We will either have to process these files and make sure that the new system which will be developed during the next year is flexible enough to cope with different formats (although perhaps nor MS Word) or we will have to rescan and reprocess the Tiden material.

In the Tiden project the focus was on using microfilm as the basis for scanning. The microfilm collections from 1950 to 1980 were unsurprisingly very difficult to use, if a good quality OCR result was the goal. All the major Swedish newspapers had been microfilmed during this period and the decision was taken not to use this film. This of course hampered the digitisation, as scanning the printed newspapers themselves was still prohibitively expensive and an agonizingly slow process.

Sadly enough the more modern Swedish microfilm also proved to be too uneven in quality to produce good results. Nevertheless an attempt was made to use some of the technically better microfilms in the EU funded TELplus project (2007-2009), where the NLS participated in Work Package 1 (“Making searchable digitised images via OCR”). As could be expected the resulting text contained a lot of errors, both because of the technical quality of the film and on account of the widespread usage of Gothic fonts in Sweden before the twentieth century. For copyright reasons only material up until about 1920 was used and the titles chosen was an unexpected mix of more well known older newspapers and small local papers. Even though the selection of titles was heavily influenced by microfilm quality and the volume was quite small (about 200 000 pages), the web site was very well received, especially among family historians. The main lesson from this project was that smaller local newspapers should be included in larger projects in the future, mixed with more well known national newspapers.

The Digidaily projects

As a result of the participation in the TELplus project the NLS decided not to develop any large scale digitisation facilities of its own. There are some in-house projects including the very ambitious digitisation of all 5600 Swedish Government Official Reports 1922-1999, which is done with a scanner robot. Large format and large scale projects involving millions of pages would instead be outsourced. The Swedish National Archives already had many

years of experience from digitising church records, maps and other types of archival material at their facility in Fränsta in the middle of Sweden, 430 kilometres north of Stockholm. It was only natural for the two government agencies to plan and perform a joint project and EU funding has been granted for the initial three year project (April 2010 to March 2013) and the present one year project (April 2013 to March 2014).

The two Digidaily projects are development project where the NLS and the Swedish National Archives develop efficient methods and processes for digitisation of newspapers in Sweden. The projects also aim at developing routines for future cooperation both between the library and the National Archives. Financiers are 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 and we now estimate that for the bulk of the newspaper collection the cost for a digitised page including OCR will be from around € 0.25 (glued or stapled but not bound) to € 0.40 (bound volumes taken apart). Two important factors in lowering the price are fast and sheet feeding duplex scanners, and a segmentation and OCR process almost without any manual work, but maybe the most important element is that the library has a second copy of all Swedish newspapers between 1850 and 1978. This copy used to a part of the collections at Uppsala University Library, but it was generously donated to the NLS to be used for retrospective microfilming and nowadays digitisation. This copy can be taken apart without any demands for it to be restored or rebound. Working with loose sheets makes it possible to use sheet feeding scanners such as SUPAG Mediascan or BancTec IntelliScan.

The second Digidaily project is much more focused on current newspapers and will involve the handling of a more complex mix of different editions of newspapers, supplements, news bills, colour printing, different paper qualities, etc. The use of high quality digital cameras instead of over head scanners will also be investigated in order to further lower the price per page.

Even though the two Digidaily projects have been clearly labelled as development project, the result will not only be methods for efficient digitisation from paper originals and a low cost per page. During the first project about 2.5 million pages from two major Swedish newspapers (Aftonbladet 1830-2010 and Svenska dagbladet 1884-2010) were digitised and delivered from the National Archives to the NLS. In the second project three more newspapers will be covered (Dagens industri 1983-2010, Dagens nyheter 1864-2010 and Expressen 1944-2010) with roughly the same number of pages. In these two projects we will have a good start for future newspaper digitisation with 5 of our 122 million pages in digital form.

Since the Digidaily projects use JPEG2000 and METS/ALTO with article segmentation as output formats we have not been able to integrate the produced material in the search solution and interface developed for the TELplus material. Once again there is a problem with incompatible formats as with the Tiden project. The new system that the library will develop will of course be based on what is now almost the standard (METS/ALTO), but will probably have to be able to handle a variety of formats.

Digitisation of current newspapers

Since current newspapers exist in a very usable digital form prior to printing, namely PDF-files used in computer-to-plate technology, the best way to preserve and present today's printed news in digital form is probably the original PDF-files used in production or processed versions thereof. The NLS is involved in negotiations with Swedish newspaper publishers about agreements for deliveries of these files on a daily basis, but so far no agreements have been made. Furthermore it has been made clear in the reports preceding the new e-legal deposit legislation that PDF-files produced by a newspaper publisher for printing can not be considered to be e-legal deposit material and therefore not a part of the legislation. This means that even if we enter into an agreement with a publisher, there is no guarantee that the files will be delivered indefinitely.

The plan is therefore to change the production for current newspapers from microfilming to digitisation. This would of course involve not only the major newspapers, which are often in focus in historical digitisation projects, but also all the smaller local newspapers, where much of the daily life in different areas of Sweden is still described. This move from analogue to digital also means the National Archives will be working with legal deposit material for the first time and this involves a lot of communication both with the NLS and with the newspaper printers delivering the legal deposit copies.

The digitisation of current newspapers will probably not differ very much from how historical newspapers are handled. Since all but a few of the newspapers in Sweden are in tabloid format, we will be able to use fast duplex scanning. Since the library does not have any permanent funding or even continued project funding for the digitisation of historical newspapers, a flexible production line for current newspapers is possibly the only solution for a continued digitisation of newspapers in Sweden. This will maintain the existing production line at the national Archives, which will handle about 2-3 million pages per year. If and when agreements on the delivery of PDF-files from publishers become a reality, it will be quite easy to replace some of the current titles, which would be delivered digitally, with historical newspapers awaiting digitisation. If further funding is found, the production line could be scaled up considerably.

Newspaper web sites - harvesting and e-legal deposit

The NLS has been harvesting Swedish web sites since 1997, less than one year after Brewster Kahle and the Internet Archive started the work on Archiving the Internet., which is the name of his article in the March 1997 issue of Scientific American. The goal of the Swedish web harvesting was of course less ambitious than Kahle's and was limited to the top domain .se and .com/.org/.net/.nu web sites considered to be interesting in documenting Swedish web publishing. The Internet Archive also targeted the Gopher hierarchy, the Netnews bulletin board system, and FTP servers, whereas the Swedish project only dealt with WWW sites. The number of captures per year was also smaller in Sweden, usually 2 or 3 in the early years. In contrast the Intenet Archive project has captured the NLS web site 552 times since 1998, i.e. on average 34 times per year but unevenly distributed with peaks during 2004-2008.

The results of the Internet Archive harvesting are available on the Internet through the Wayback Machine since October 2001. The Swedish web archive is only available at the

National Library on two computers without connection to the Internet, in accordance with the "Ordinance (2002:287) concerning the processing of personal data in Kungl. bibliotekets digital cultural heritage projects". This ordinance clearly permits the library to collect and store the Swedish "national digital cultural heritage" as it is published on the Internet, including all material which can be classified as Swedish on the grounds of "address, addressee, language, originator or sender". Information about individuals may be collected and stored in the database "in order to benefit the need for research and information", even if it is sensitive information as defined in the Personal Data Act, i.e. concerns ethnicity, political views, religion, etc. Copying of the archived pages is not allowed, but printing is possible. There is no search facility available, but search by URL of the pages and links presented in the results, with one link for each time the page was archived. The archived web pages have been stored on tape and are fetched to disk on request, which takes about two minutes. The Swedish web archive today has about 1.7 billion objects.

The library had been doing web harvesting for five years when the above mentioned ordinance came into force in 2002. In the same year daily harvesting of Swedish newspaper web sites on a daily basis was started, at first with only a few titles but greatly expanded in 2004 to include about 140 of the newspaper titles. Among the titles there are both large national newspapers and local papers with a lot of material of interest for future local historians and genealogists. Harvesting is done once a day, with parameters set to limit the depth of the harvesting in the web site hierarchy and the number of objects per site. The results vary and there are numerous examples of missing images and style sheets, but despite these problems the NLS now has a ten year archive with daily snap shots of a majority of the Swedish newspaper web sites.

But of course web harvesting has its limitations. While getting and archiving the context of all individual articles is very important when you want to know what newspaper web publishing looked like, it is very difficult to capture all articles a newspaper web site publishes during one day, without also storing a lot of redundant material harvested earlier on. As a part of the recent e-legal deposit legislation in Sweden the NLS is trying to use customized RSS-feeds for the newspaper web sites as a method to capture all new articles published, but also updates of articles. This puts some development work on the newspapers, but also relieves them from archiving articles for delivery to the library.

The broad definition of what is included in the e-legal deposit law is: *Electronic materials that relate to Swedish conditions and are made available to the public in Sweden by transmission through networks*. Electronic material is described as *a defined unit of an electronic recording of text, sound or image that has a predetermined content intended to be presented at each use*; and *the electronic media should be complete and permanent character.* This means that web forums, wikis and other "living" web material is not to be included. To be considered to relate to Swedish conditions the material should *mainly be addressed to a Swedish audience, mainly expressed in Swedish, produced by a Swedish author or performed by a Swedish artist*.

The material should also be unique to the web, which in a strict interpretation of the law would mean that only web published articles that are not also published in the printed newspaper should be delivered to the NLS. The intention here is to avoid redundancy, but the newspapers have made it clear from the beginning that they have no way of filtering out only the articles unique to the web and discarding the ones also published in print. This is in many ways an advantage since the articles on the web often appear in two or more versions and

perhaps only one of them is also available in print. Also web publishing is a different media form from the printed newspaper and should be treated as such, although the content may overlap to a certain degree.

During a two-year period (2013-2014) only a limited number of web publishers have to deliver (or make available) their material, among them the ten largest newspapers, a number of radio and TV channels, the ten largest magazines and journals and a number of government agencies. The library is in the process of developing systems and routines to handle the very large amount of material that will start coming in 2015, when three main (but not easily identifiable) groups will be covered by the law:

- Those who automatically enjoy constitutional protection according to the Freedom of Speech Act (traditional mass media like television, radio, newspapers, books and magazines, and different kinds of organizations and associations)

- Those who professionally produce and/or distribute electronic materials

- State and municipal authorities

To get the best possible, if not the complete, representation of what is published by Swedish newspapers (and other news sources) on the web the ideal solution would be to combine the web page as it is harvested once a day with the individual articles collected by the library through e-legal deposit. In this way both the general layout and context of the newspaper web site is preserved, at least the moment it was harvested, and the development of the articles during the day can be followed until the next day's harvesting.

Access, copy right legislation and the Personal Data Act

Two large obstacles for successful use of newspapers in digital form in Sweden are 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. A not uncommon misconception is that the current owners of the newspaper are the rights holders of the entire content of the publication including back issues, but if the economic rights has not been handed over by the author (quite common in current newspaper publishing) this is not the case. The NLS maintains a very strict policy on how to interpret the legislation and has drawn some criticism for this. Since an article in a newspaper may have been written by a 10 year old identified author in 1863 and this author may have died at the age of 90 in 1943, the library considers only material before 1863 to be completely free of copyright. This is a rather extreme interpretation, but also one that seems quite necessary when we probably will go into negotiations with copyright organizations later this year, after a change in the Swedish copyright legislation, which among other things makes it possible for libraries and others to make agreements with organizations representing copyright owners collectively.

At present it is impossible to know how much such agreements will cost. It will possibly not be very expensive to make agreements for say 1863 to 1913 and newer material will be more expensive, but only negotiations will show the actual cost.

The Personal Data Act in Sweden concerns only living persons and since newspapers is a mixture of information about the living and the dead it is quite possible that the library will

follow the practices of some genealogical organizations, which is to have a limit of 100 years. This would only give us the possibility to publish material up until 1913, but there has so far not been very much discussion about where to draw the line. However, as mentioned above the Swedish web archive has permission to store even current material containing possibly sensitive information about living persons and to make it available for research. Hopefully this more liberal view on personal data will be the one used also for digitised material.

At an evaluation seminar in February 2013 for the first of the Digidaily projects, The NLS and the National Archives received a lot of very positive comments from the two invited international experts Simon Tanner and Edwin Klijn. Both the technical aspects concerning quality and the economical aspects were applauded, but there was also criticism, especially concerning access to the digitised material. In the Digidaily project work on a user interface had not been included, since it was a development project focused on workflow and cost.

The library has investigated different possibilities when it comes to publishing the digitised material. The result of these investigations was a decision earlier this year to develop a solution in house and to use the library's long experience from developing its own search interface for the Swedish national catalogue Libris. The decision also makes it clear that the goal is to build a solution that can be adapted to different types of material. In many ways we will start the development of the system and user interface from scratch. Having newspapers as the first type of material might actually be a very good idea, when so much of what we will digitise and collect in digital form will have the strong temporal and geographical aspects typical for newspapers. This is 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.

In a not to distant a future the library 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, and also other types of digitised print material (books, journals, ephemera, etc). It is not difficult to imagine what an excellent tool a combined interface for all these types of material could be for all kinds of research concerning Swedish history, society and culture on both a national and local level.