IFLA 2013 SATELLITE MEETING ON NEWSPAPER & GENLOC SECTIONS

15 August 2013

Digitisation of historic newspapers and voluntary digital deposit of newspaper pre-print files in the National Library of Estonia

Krista Kiisa
Digitisation Coordinator
National Library of Estonia

The work with newspapers in the NLE involves 3 main activities:
Digitisation of historic newspapers
Harvesting on-line newspapers from the web
Voluntary deposit of newspaper pre-print files by publishers

Workflow for current newspapers.

NLE's priority in working with newspapers at the moment is acquiring and archiving pre-print files of currently published newspapers. We started the preparations for this already some years ago. The current Legal Deposit Act in Estonia applies to printed materials and to web publications, but it doesn't apply to electronic pre-print files. For long-term preservation purposes, we did microfilm all the paper newspapers received under legal deposit for many years. During the recent years of economic slowdown, when our budget was drastically reduced, we had to take the difficult decision to stop the microfilming of newspapers. Though the Legal Deposit Act allows us to collect and archive web publications, we didn't have enough resources to extensively harvest dynamic data from newspapers' websites. As the content published on the web and the content on paper are usually two totally different things, it was decided that more efforts should be made to clarify the message to publishers how useful it actually is for them to archive their pre-print files of the paper edition in the National Library of Estonia. We didn't have the law supporting us, so in negotiations with publishers we had to stress the good will and the valuable experience it gives to the publishers in terms of the future legal deposit system that we hope will come into force in 2014 with the amended Legal Deposit Act.

We started negotiations with the biggest daily newspaper called „Postimees" (http://postimees.ee), and got surprisingly good feedback. The very first pre-print files started to arrive into our ftp server in 2007. Not so regularly and properly as we had dreamed perhaps, but the first step was made. That was a showcase with a well-known publisher, from where it was much easier to start convincing other publishers. Due to limited IT support for a period of time, we kept a moderate pace in getting new agreements with publishers, but over several years we had nearly half of newspaper publishers sending their pre-print files to us voluntarily. A frightening experience was the crash of one publisher's server that resulted in the loss of all data. As many publishers do not appear to have regular backup facilities, it definitely was a moment for them to consider the possible dangers. It was an advantage for us in the negotiations, as in the agreement we promise to give their original files back any time they need them.
But as we know, the appetite grows when eating. We wanted more! And what's more important: we wanted to have them not some time during the day or week, not after some reminders, but we wanted them early in the morning, right before the working day starts. We started to imagine how we could save more money by not having to buy lots of additional paper copies for compiling the Database of Estonian Articles called *ISE*, (http://ise.elnet.ee/search), where about 200 articles are added to the database each day. Why not do this work from electronic pre-print files, being ready to start at 8 am and with no need for extra paper copies.
We didn't have any other efficient means to use, so we focused on this method.
In autumn 2012, negotiations between National Library of Estonia and **Estonian Newspaper Association (EALL)** started. EALL is a non-profit organisation working in the common interests of newspaper publishers. At the moment EALL unites 40 newspapers published in Estonia, with a total daily circulation of 510,500 copies. EALL was very interested in cooperation. It was evident that our idea would be profitable for all partners.

The aim of the National Library of Estonia is to have the content of newspapers archived in its digital archive DIGAR (http://digar.nlib.ee), with a flexible user access over Internet. We are in the middle of a campaign to invite people to use our digital archive more. And we wanted to arrange the in-house workflow for cataloguing the articles into the articles database from pre-prints operatively the same day the newspapers are published. Last but not least, having the pre-print files, we can avoid the cost of microfilming or further digitisation.

The Estonian Newspaper Association supports the idea that we archive the electronic pre-print files, and will provide publishers with their files back whenever the publishers need them again. Even more, EALL saw here a business model for themselves. The opportunity to use NLE's server as an intermediate station, where media monitoring companies, who have negotiated licences from EALL, can download the pre-print files for their commercial use. As monitoring companies need to finish their work before the working day begins, it was now already the EALL's concern that the National Library would receive the pre-print files very early in the morning so that they would be accessible for the media monitoring companies. We were very happy to hand over to the EALL the unpleasant task of keeping the negotiations with publishers.

All electronic pre-print files were decided to keep immediately open for use on the premises of the NLE. That's something we definitely want to back up in any circumstances. The access restrictions for broader use via Internet are defined by the publishers. And to be honest, at the moment there are only two publishers who don't allow free Internet access to their pre-print files. Most publishers have set 2 weeks, 2 months or 3 months restricted access time. During that time the files are closed for open access, after that period they are open for all users of the NLE digital archive DIGAR.

As a final result, the NLE is now part of a value chain, acting as a mediator between private companies. This has resulted in a better coverage of newspaper and journal titles that we archive, and has also given us an opportunity to improve our services both to commercial users as well as regular readers.

Today we have 41 publishers who are daily (or more precisely, nightly) sending their newspaper preprint files to our archive, altogether around 110 titles. As mentioned before, every publisher has the right to determine access restrictions for public access over the Internet. But on the premises of the National Library the newspapers are accessible electronically in all in-house computers on the same day they are published.
More or less it means we have 99% of currently published central daily or weekly important newspapers in our hands same time they are sent to printing house – it means hours before they are actually sold. This number reflects only the number of either daily, weekly or centrally published newspapers. It doesn't include the so-called small scale newspapers published by schools, churches, institutions, organisations, different societies and companies. The amount of such newspapers is currently 342 titles in Estonia. It is a very time and work-consuming area needing extra staff.

Another work process actively developing is the web harvesting.
At the moment we use the selective method for registering and archiving the newspapers from the web. The archiving activity runs according to the Legal Deposit Act, in force since 2006.

According to the Act the digital legal deposit system is administered by the National Library of Estonia. The general criteria for selecting and archiving web materials are: their publication, identifiability, exhaustiveness, long-term and permanent value, place of publication (specified according to publisher). All these criteria were fixed by the Web Archiving Experts Working Group –a joint advisory body comprised of members form leading memory and research institutions. The objective of the working group is to provide advice on selecting the material for archiving; the broad-based working group also represents the interests of current and future researchers.

The National Library of Estonia has the right to harvest Estonian web content. Access to the archived websites is open unless right holders impose a restriction. So far only a small amount of the most valuable sites has been harvested. A pilot study has been conducted on archiving a large volume of websites and the first complete harvest of the Estonian web domain is planned for 2014. Since it is impossible to foresee what is important for future researchers, it is necessary to harvest the snapshot of the whole Estonian national web. Newspapers, published only in web, are planned to be harvested according to the frequency of publishing, in most cases it means, daily.

Last but not least the activity with the longest history in the NLE is **digitisation of historic newspapers.**
The digitisation of older newspapers is done predominantly from microfilms that were created in the beginning of the 1990's. This means that the quality of scanned images is often quite poor. Unfortunately that's the only choice we've got since due to the very complicated history of Estonian library holdings after World War II, actually no library in Estonia no longer has the complete newspaper collection in it's physical holdings. The nearly complete microfilmed/digitised one is a result of co-

operation where single physical copies were gathered from different libraries only for short-time microfilming purposes in the 1990's and are now accessible from the webpage http://dea.nlib.ee

At the moment it is still an image database, where search is possible only by title and date.
Though there is no fulltext search available yet, it's still quite actively used.
There are 380 titles, altogether 1,3 million pages, mostly in the Estonian language. The database gets around 300 single visits every day. For example, the Estonian radio's morning programme is compiling its old newspapers editorial column using actively our old newspapers' database for that.

Our next plan is to OCR and run the article segmentation on these pages. The activity started thanks to the Europeana Newspapers project. This is a great opportunity to gain know-how from project partners and gives us the possibility to OCR and run an article level processing for half of our digitised newspaper images. The technical work of OCR and article segmentation is done by 2 project partners: Innsbruck University and CCS (Content Conversion Specialists GMBH).

As described in my presentation, there are 3 different databases, 3 different places the users must search if they want to find newspapers. That's something we want to change in the very nearest future. For that reason we plan to acquire portal software as a single point of access to all newspapers – historic digitised, current pre-print files or harvested from the web- from one site. The current plan is to open the new newspapers' portal in 2014. We also plan to enable crowdsourcing there, as we hope that with the help of our users we can improve the current quite poor result of OCR, gained from microfilm scanned images.

Until then we need to be satisfied with the different places for searching the newspapers:
http://dea.nlib.ee
http://digar.nlib.ee/digar/lihtotsing?m=s&l=ajaleht&q=ajalehed

Although we have worked with newspapers for many years – microfilmed them, digitised and described them, harvested them from the web and archived their print files – and have seemingly provided services to our readers, we still find new ways of improving the service and also to enlarge the coverage of titles that readers can access through our portal. This demonstrates that despite being "old news", newspapers continue to be at the forefront of digital library developments and sometimes driving them.