Submitted on: 28/07/2014

Implementation of digital deposit at the National Library of Norway

Svein Arne Solbakk
Digital Library Development, National Library of Norway, Mo i Rana, Norway
E-mail address: svein.solbakk@nb.no

Abstract:

The paper will outline the strategies at the National Library of Norway for capturing digital born publications directly from the publishers, even though the publication itself is on paper or other physical media. Also the actual implementation of these strategies and lessons learned in the process will be presented. This includes both the dialog with the publishers, the technical solutions for digital capture, and the processes for handling digital publications within the library. The paper will focus on a concrete implementation for newspapers. However, the principles are general.

The paper will also briefly cover how the most important parts of the Norwegian Internet are currently captured for preservation.

Finally, the digital deposit and the web harvesting will be placed into the context of the ambitious digitization program at the National Library of Norway.

Keywords: digital deposit, newspapers, digital library, web harvesting

1. A digital national library

The National Library of Norway established an ambitious digitization programme in 2006. The ultimate goal is to become a digital national library. In this context, a digital national library is not just a popular service on the web. It represents a completely new way to fulfill the role of a national library.

To become a digital national library, it is necessary to:

- Digitize the non-digital collection
- Collect the digital born masters of non-digital publications
- Collect the publications that only exist in digital formats
- Negotiate with copyright holders to be able to make the collection available in digital formats

- Make the digital collection available to the users where they are, when they want, in the way they would like

All of these actions require extensive focus and a lot of resources over a long time period.

The digitization of the non-digital collection at the National Library of Norway is expected to take 25 – 30 years. More than 50 people work full time with the digitization of a variety of materials. The current status is that 80 % of the books (375 000 titles) and 20 % of the newspapers (16 million pages) that have been published in Norway are digitized. In addition, large amounts of journals, photos, handwritings, posters, music and other sound recordings, video, film and broadcasts, have been digitized.

A lot of agreements with copyright holders have also been made, that makes it possible to give access to the digital formats. The Bookshelf agreement (2012) gives the National Library rights to make all books published in Norway in the year 2000 or earlier, available to everyone with a Norwegian IP-address. Currently, 175 000 book titles are available for free reading in Norway, and we expect to have 250 000 titles available in 2017.

The National Library has also made agreements with newspapers that make it possible to give access to 35 newspaper titles in digital format in Norwegian libraries. Nine of the titles are available for everyone. Due to these agreements, 50 % of the pages of digitized or digital deposited newspapers are available in Norwegian libraries.

Another example is an agreement made with the major radio broadcaster in Norway, where we currently give open and free access for everyone to more than 36 000 radio programs. This includes news broadcast on radio from the 1930s up to yesterday’s news.

However, the focus of this article is action point two and three in the list above. Therefore, the rest of this article is about the deposit of digital cultural heritage.

2. Digital deposit

The first digital deposit at the National Library of Norway was implemented at the end of 2004. From that point, radio broadcasts from the Norwegian Broadcast Corporation have been deposited as files via the Internet, and the transport of physical tapes stopped. In 2008, digital deposit of the first television broadcast channel was implemented. Currently, digital deposit via the Internet has been implemented for all nationwide radio and television broadcasts in Norway. In addition, digital deposit is implemented for the local broadcasts from the Norwegian Broadcast Corporation.

The broadcasters as well as the National Library have gained from the transition to digital deposit. It is much more efficient, and far cheaper than logging the broadcasts on physical tape media, and transporting the physical media on the road.

For printed materials, however, if digital deposit is implemented, the National Library still claims the paper version. It is therefore not so obvious what the publisher gains from digital deposit. In addition, for the newspapers, the National Library also wants to give access to the digital version of the newspapers in Norwegian libraries, to be able to stop the expensive microfilm production.

Still, the Norwegian legal deposit law limits the mandatory legal deposit to the format of the actual publication. Thus, the National Library cannot claim the digital files that are used to produce a paper publication. To get these digital files, we have to sign an agreement on digital deposit with the publisher.

The National Library of Norway therefore invited Norwegian newspaper publishers to collaborate on digitization as well as digital deposit. In such collaboration, the National Library handles the digitization and OCR, and takes responsibility for long term preservation of the digital master files both from the digitization and from the digital deposit. The newspaper publisher pays 50 % of the digitization costs, and gets a complete copy of the results. When collaboration on digitization is established, we also implement digital deposit of the newspaper in question.

So far, agreements have been signed with publishers covering 35 newspaper titles, and there are negotiations going on continuously with newspaper publishers to add more titles to the list. We have focused on some of the largest newspapers in Norway, but at the same time wanted to have a broad representation of digitally available titles from all over the country. Some of the agreements are limited to digital deposit, while quite a few of the publishers also collaborate on digitization of historical issues of the newspapers.

Even though 35 out of more than 250 active titles do not seem to be a lot, due to the focus on the largest titles they represent a large percentage of the number of published newspaper pages per year in Norway. This is an important fact since the digitization of the paper issue of a newspaper immediately stops when a digital deposit agreement with the publisher is made. This represents a significant saving for the National Library, and resources are reallocated to digitization of historical newspaper issues.

In addition to digital deposit of the newspapers, the National Library started digital legal deposit of eBooks in 2012. Currently, eBooks in the ePub and in the PDF-format are collected and preserved in the digital long term storage.

3. A digital deposit solution for newspapers

The objective has been to establish digital deposit for most of the active newspaper titles in Norway within a five year time frame. This means to be able to monitor and follow up the digital workflow on a daily basis for approximately 250 newspaper titles.

Also, the publishers use different digital production systems, and they have implemented a large variety of digital workflows. To avoid having to implement 250 different solutions for digital deposit, the National Library therefore has defined a standard for digital deposit of newspapers in Norway. The standard specifies

- Naming of files to be deposited
- File format
- Image resolution
- Delivery method

Every page in the newspaper should be a separate PDF/A file, and the images should at least hold the same quality as in the files used to print the paper version. All the files for a complete issue of a newspaper should be packaged in either a tar or a zip file. The packaged file should be placed at the publisher's ftp-server, available for the National Library. All the pages from the paper version of the newspaper must be included in this package.

The naming convention for single pages is:

<newspaper-name_subname_zone_date_volume_number_issue>-
<sequencenumber_pagenumber_appendixname>.pdf

A real life example – the first page of the news appendix of the morning edition of the newspaper Aftenposten on June 7th 2006, for area zone 1:

aftenposten_morgen_1_20060607_147_253_1-1_001_nyheter.pdf

The naming convention for the packaged file of one issue is:

<newspaper-name_subname_zone_date_volume_number_issue>.zip/.tar

All letters should be lower case, and there should be no subcatalogues. The file names contain sufficient information to rebuild the published newspaper in a digital environment.

For quality assurance, there is also defined as a recommended option to produce an md5 checksum of the packaged issue, and put it in a separate file together with a list of all the file names contained in the packaged issue. The naming convention for this file would be:

<newspaper-name_subname_zone_date_volume_number_issue>.md5

If some of the defined elements are not relevant for a given newspaper, "null" is inserted. The most complex newspaper in Norway was used as a case when the naming convention was defined. Most newspapers do not have subnames or zones.

As a rule, every title is collected daily from the newspapers ftp-server. However, we require the publisher to have sufficient space on their ftp server to keep the issues for the last 14 days. In this way we have a small buffer if something goes wrong.

The first step after the file transfer is to validate the packaged newspaper issues. If the optional md5 file is available, the md5 checksum of the received package is compared to the checksum in the md5 file, and the file list is checked against the files in the received package. The next action is to validate the correctness of the PDF format, using JHOVE (2014). If the complete package validates correct, the newspaper issue is deleted from the publishers ftp server. If something seems to be wrong, the issue is put on hold, and the publisher is contacted to get correct files.

Next step is to run every file in the newspaper issue through docWorks (2014) to extract the text from the pages, and to format it in the XML/ALTO format (2014). Most of the pages will have hidden text that is used directly, but often some parts contain graphical elements, and then they are run through an OCR engine to extract the text. Typically this may be the case for advertisements. The XML/ALTO files are used both to index the text from the newspapers for search in the digital library service, and to be able to highlight search hits in the text.

After docWorks, access quality JPEG2000 files (2007) are generated for every page to have a single way to handle digitized and digital deposited newspapers in our digital library service.

Then all files related to the newspaper issue is wrapped in a METS container (2014), and metadata is extracted from the file name and inserted into Mavis (2014), a database used by the National Library for a variety of digital object types. The complete METS container is ingested into our digital long term preservation infrastructure, the access files are transferred to our image server, and the METS and XML/ALTO files are used to make the newspaper issue searchable and available in our digital library service.

If the newspaper has fallen out of copyright, or an agreement allowing us to is made, the text is also made available for external search engines, and functionality in our service is offered to download a PDF version with hidden text. The PDF is generated on the fly from the JPEG2000 files and the corresponding XML/ALTO files. This is, however, seldom an option for digitally deposited newspapers, but more relevant for digitized historical newspapers.

So far, digital deposit is implemented for 15 newspaper titles. Adapting the digital workflow of the publisher to deliver a copy to the National Library confirming to the defined standard, tends to take a lot of calendar time. Also, changes in a publisher’s workflow often introduce unexpected challenges.

The workflow at the National Library has to a large extent been automated. However, some critical elements require human follow up. First, the 250 active titles have a broad range of publication patterns. Some have several issues daily, while others are published only once a week. So far, it is manually monitored whether a digital issue of a given newspaper title is received according to the expected publication pattern. If an expected issue has not appeared, the publisher has been manually contacted to solve the problem. This is possible to carry out for 15 titles, but it would be extremely time consuming to do manually for 250 titles.

Also, deviations in the workflow, that may cause a newspaper issue to stop, so far have not been reported by the system. It has consequently required a lot of effort to follow up the workflow to discover deviations and resolve problems that have occurred.

To be able to scale the digital deposit to all active newspaper titles, a new system has been developed which purpose is to support the digital deposit of newspapers. The system is currently in test, and the plan is to put it into production in September.

The expected publication pattern for each title is put in to the system, and the system monitors that each title is deposited in accordance with the expectations. If a deviation occurs, an email is sent by the system to the publisher requesting the expected issue.

Also, the system polls the different steps in the digital workflow, and if a deviation occurs, an alarm goes off. The system offers a dashboard where the complete picture is drawn. I.e. if something has gone wrong with any of the deposited titles, it will appear on the dashboard. In addition, the status of every single title may be monitored in the system. When resolving an error, the operator may set a new status to inform the system that the deviation has been taken care off.

So far, the systems looks very promising, and we look forward to put it into operation this autumn.

4. The way forward

The National library is currently negotiating with a large consolidated group of newspapers (approximately 80 titles), and we expect to make an agreement with them on digital deposit in 2014. Digital deposit of the 80 new titles will then be implemented over the next two to three years. This will represent a major milestone in our digital deposit of newspapers, and it will allow us to make a considerable reallocation of digitization resources in the direction of historical issues.

Also, a revision of the Norwegian legal deposit law is under work, and it is expected to include the right to claim the digital born version of paper publications. If such a revision is made, it will enable the National Library to scale up the digital deposit at a higher pace. Anyhow, the National Library will continue its focus on getting new agreements to implement digital deposit for as many of the active Norwegian newspaper titles as possible.

The newly developed support system is planned to be expanded also to other types of digital publications. Work is already going on to adapt the system to support digital deposit of radio and television. The number of nationwide radio and television broadcast channels has risen significantly over the last years, and thus the need for a support system is evident.

The next step may be to adapt the system to support digital deposit of journals. We have already made agreements on digitization of historical issues with a few publishers of Norwegian journals. Digital deposit is the obvious next step.

5. Web preservation

Preserving the Norwegian cultural heritage on the web requires a completely different set of tools. The National Library of Norway has been working actively with solutions for harvesting and preservation of web contents for 15 years. The first full domain harvest of the .no domain was carried out in 2001. Up to 2003, the Nordic national libraries collaborated on this issue in the Nordic Web Archive project. In 2003, the Nordic countries joined an international initiative, and formed the International Internet Preservation Consortium together with other National Libraries and the Internet Archive. IIPC (2014) has grown to be the international driving force within web archiving, and is to day counting 49 member institutions from 25 countries, including national, university and regional libraries and archives, as well as some service providers.

A lot of tools have been developed by IIPC members, and most of them are described on the IIPC website. Quite a few of the tools are available as Open Source. The most important tools are Heritrix for the harvesting of web pages, and OpenWayback for giving access to web archives. Both of these tools have been developed by the Internet Archive, with contributions from other IIPC members. Currently, IIPC is in the process of taking more of the responsibility for future development of these critical tools in the context of web archiving.

Currently, the Norwegian web archive counts more than 8 billion files. We have not performed full domain harvests since 2008, due to a political discussion on privacy issues that still is not concluded. Approximately 1 000 sites are at the time being harvesting very frequently. The home pages are harvested every hour, while the rest of the sites are harvested at regular intervals. To do this, we have to inform the site owner before starting to harvest their site.

The already mentioned expected revision of the Norwegian legal deposit law is also expected to clarify how privacy should be handled in the web archiving effort. Thus, the National Library is currently preparing to be able to once again scale up to full domain harvest of the Norwegian web domain. Also, the National Library has signed agreements with several site owners on both preservation of their sites and on giving access to the web archives of their sites. Consequently, the National Library will establish a limited open web archive giving access to the web archives of these sites along with the web archive of the government sites.

6. Concluding remarks

The National Library uses a lot of human resources to handle the legal deposit of paper materials. Unfortunately, when digital deposit of printed materials is implemented, the paper versions of the publications still also have to be handled. Therefore there are not any staff to reallocate to the handling of digital deposit. Even so, it has to be acknowledged that digital deposit also requires human resources.

By introducing a new tool for the handling of digital deposit of newspapers, a large step is taken towards being able to scale up the digital deposit significantly. Still, there will be a need to closely monitor the digital deposit both from an ICT systems maintenance perspective, and from a legal deposit quality assurance perspective.

Until the legal deposit law is revised to include the digital born version of paper publications, there is also a need to focus on negotiating and making agreements with the publishers to be able to get the digital versions.

By scaling up the digital deposit significantly, we can reallocate our digitization resources to focus further on the historical materials. Thus, our goal to become a digital national library gets closer!

Also, it is important to preserve at least the most important parts of the national web domain. Already, the short history of the world wide web shows that the web part of a nations cultural heritage is typically very volatile and short lived. Therefore the National Library of Norway will continue to play an active part in IIPC to support an international focus on the harvesting and preservation of cultural heritage on the world wide web, as well as giving access to the web archives for research and documentation purposes.

References

Bookshelf agreement (2012) *Contract regarding the digital dissemination of books (Bokhylla/The Bookshelf)*. [Online] Available from: http://www.nb.no/content/download/997/16468/file/Bookshelf Contract.pdf. [Accessed 22 July 2014]

docWorks (2014) *Content Conversion Specialists*. [Online] Available from: http://content-conversion.com/?lang=en#docworks-2. [Accessed 27 July 2014]

IIPC (2014) *International Internet Preservation Consortium (IIPC)*. [Online] Available from: http://netpreserve.org/. [Accessed 22 July 2014]

JHOVE (2014) *JHOVE Free Softwaredownloads at SourceForge.net*. [Online] Available from: http://sourceforge.net/projects/jhove/. [Accessed 25 July 2014]

JPEG2000 (2007) *JPEG2000*. [Online] Available from: http://www.jpeg.org/jpeg2000/. [Accessed 25 July 2014]

Mavis (2011) *Feenyx*. [Online] Available from: http://www.feenyx.com.au/. [Accessed 27 July 2014]

METS (2014) *Metadata Encoding and Transmission Standard (METS) Official Web site*. [Online] Available from: http://www.loc.gov/standards/mets/. [Accessed 25 July 2014]

Solbakk, S.A. (2012) Digital preservation at the National Library. *Meta*, 3, p.17-21. Available from: https://www.notur.no/sites/notur.no/files/publications/pdf/meta_2012_3.pdf. [Accessed 22 July 2014]

XML/ALTO (2014) *ALTO: Technical Metadata for Optical Character Recognition (Standards, Library of Congress)*. [Online] Available from: http://www.loc.gov/standards/alto/about.php. [Accessed 25 July 2014]