IFLA WLIC 2013
SINGAPORE

Yes please, both : large-scale digitization and legal deposit of newspapers in Norway

Svein Arne Brygfjeld

Head of Digital Library Development, National Library of Norway, Oslo, Norway
sveinarnebrygfjeld@nb.no

Mona Løkås

Head of Periodical Section, National Library of Norway, Oslo, Norway
mona.lokas@nb.no

Copyright © 2013 by **Svein Arne Brygfjeld and Mona Løkås**. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License:
http://creativecommons.org/licenses/by/3.0/

Abstract:

The National Library of Norway (NLN) is going through a massive digitization programme. This programme aims at digitization of the entire analog collection in the library, digital legal deposit, digital long term preservation and digital library user experiences.

Digitisation of newspapers is performed both based on paper originals as well as microfilm. Parts of this activity is out-sourced, others are done in-house. At the time of writing, some 12 million pages are digitized.

To reduce the need for digitization, NLN seeks to establish digital legal deposit also for newspapers published on paper. PDF's in print quality are delivered to the library based on automatic file transfers during the night. Such digital legal deposit is established for 15 newspapers, aiming at about 50 newspapers by the end of 2013.

All digital long term preservation as well as user access is based on JPEG2000 file formats. This choice of format also gives the users immediate access to the highest quality in the digital images. The text content from all digital sources are derived and stored in a large-scale search platform, giving the users the opportunity to do full-text search through the entire digital collection.

Keywords: digitization, legal deposit, preservation, user experience, JPEG2000

Yes please, both : large-scale digitization and legal deposit of newspapers in Norway

1 OVERVIEW

The National Library in Norway (NLN) is going through a massive digitization programme aiming at a complete digital version of the entire collection at NLN. The collection includes most information carriers and media, also in various digital formats. At the moment of writing NLN carries out digitizing of media like books, photograps, journals, radio broadcasts, moving images and newspapers.
Wherever possible according to legislative frames NLN gives users access based on modern Internet/WWW based services. More than 25% of the library staff is currently directly involved in the digitization programme.

The collection is being digitized in accordance with the requirements enabling NLN to carry out long-term preservation as well to give as advanced user services. The Library is establishing the required standards for this in collaboration with a number of international organisations. The digital objects are enriched with metadata and sustainable identifiers which will increase the opportunities for archiving, use and reuse over the next millennium. NLN facilitates diverse and varied use of the collection's content. The content is published in an attractive format.

NLN's digitization programme includes large-scale digitization and legal deposit of newspapers. The newspaper collection contain about 55 000 bound item, 25 000 boxes and 71 000 microfilms/microfiches. All these newspapers will be digitized during some years.

2 LEGAL FRAMEWORK

NLN has the rights to perform digitization for in-house long-term preservation, and in-house user access to the digitized collection is permitted. At the same time NLN establish agreements with publishers, giving NLN the opportunity to give user access throughout all libraries in Norway. Such agreements normally also includes a cost-share model for the digitization activity, and the publisher normally get a digital copy for their own use.

3 PRODUCTION

Digitization is performed as a regular stream-lined production activity. Some parts are out-sourced, others are done in-house. Various variants of the production are in place, based on the original media. Those are

1) Microfilmed newspapers
2) Bound items from the collection
3) Rare/fragile items
4) Legal deposit items
5) Digital legal deposit

3.1 General framework

All handling of digital newspapers is based on a common framework. This framework includes among other things requirements for

- Formats, both for preservation and access: PDF, JPEG2000, METS/ALTO
- Qualities: 400 dpi, 24 bit color
- Naming: URN
- Metadata: DC/MODS in METS
- OCR, structure analysis: stored in METS/ALTO

This common framework makes later handling and maintainance of the collection bettery in any aspect.

3.2 Microfilm

Digitization of microfilms is out-sourced. About 1,5 million pages is scanned every year by partners. Such digitization normally includes scanning, OCR and structure analysis/zoning.
Before the microfilms are sent for scanning we have to do some samples and quality control to make sure that the microfilms is suitable for scanning. The microfilming in NLN is made through several

years, some of the films are really old and not suitable for scanning. It's also necessary to get the reduction ratio for the newspapers, so we can get the results from the scanning as exactly the original as possible.

The filenaming of the digital microfilms also have to be correct, we tell the vendor that the title of each newspapers have to be XX. With every shipment we made a list of filename on each title so we can get the output from the scanning directly into our storingsystem, without any renaming of files.

3.3 Bound items and Legal Deposit items

Newspapers items received as Legal Depoit are bound in temporary bindings made for large-scale digitization. Both these and already bound items are scanned on two large-scale automatic turn-page scanners (4DigitalBooks DL 3003). After scanning they go through heavy post-processing, including OCR and structure analysis. Metadata are produced semi-automatically and stored in our in-house catalogue as well as with the digital object in the digital preservation facility. NLN has a monthly production of approx. 150.000 pages through these scanners.

3.4 Rare/fragile items

Some of the items in the collection are rare and/or fragile. To reduce risk for damages, those are run through manual scanners. The rest of the processing is common with those above.

3.5 Digital Legal Deposit

To reduce the need for digitization, NLN has established digital legal deposit for several newspapers based on agreements with the publishers. Since 2007 we have received print-quality PDF files from publishers, specially the largest newspaper publishers in Norway. Nightly processes harvest PDF-files from the publisher's ftp-servers. Automatic quality and integrity checks are performed on arrival, and files are named based on NLN's schema. To make sure that the files are "correct", no pages missing, attachment is enclosed , the filenaming is ok., we test all the edition by running a check script. We also test that the pdf is valid.

4 OCR AND STRUCTURE ANALYSIS

Scanning just produce a large set of images based on the newspaper pages. OCR is performed on all pages. Generally without any form of proof-reading, but it may be performed in some cases based on agreements with publishers. NLN also perform semi-automatic structure analysis/zoning to detect core elements like articles, images, headlines and ingresses. The output from these processes are kept in standardized formats in METS/ALTO files.

5 LONG-TERM PRESERVATION

5.1 Common platform

Long-term preservation is carried out based on the same requirements, principles and infrastructure as any other digital object in NLN's collection. All digital object are stored within one common infrastructure, currently containing a net of 3 Petabytes of data in three copies.

5.2 Formats

The scanned images are stored in lossless JPEG2000 format. These tile and resolution layers within these files are prepared for access as well. OCR, structure information and various metadata are kept as METS/ALTO files with the scanned images. This collection of files defines a digital object, and the are bundled and stored separately from other digital objects.

6 USER ACCESS

User access is based on WWW-based services only. Through a sophisticated user interface, the user can search the content based on OCR text, and it is also possible to zoom into the finest details of the digital document.

6.1 Formats

As for preservation, access is based on high-quality JPEG2000 representations. These files have the same resolution as the preservation version, but they have a lossy JPEG2000 compression allowing smaller sized files. The image delivery service converts from high-quality JPEG2000 to user-specified quality JPEGs for client-side rendering. Newspaper access shares the same service platform used for book access. This service, based on Open-Source software, currently contain roughly 56 million images.

7 SEARCH

All metadata and OCR text, or text from PDF files, are stored in a shared in-house search platform. Thus the user can search across the entire digital collection at NLN, or in a single newspaper item if preferred. Search is performed on OCR or text as is, no proof-reading is performed.

8 CURRENT STATUS

At the time of writing, 11 996 377 (17 %) pages are digitized. The whole newspaper collection is estimated to 70 000 000 pages. 15 newspapers are harvested as print-quality PDF. This will increase during 2013, how much will depend on the agreements with the publishers. As a result of good cooperation with the publishers, we assume that we have legal deposit established for all the 254 Norwegian newspapers within 4-5 years.