Digidaily – Inter-Agency Mass Digitisation of Newspapers in Sweden

Heidi Rosen; the National Library of Sweden. Torsten Johansson; the National Library of Sweden. Henrik Johansson; the National Library of Sweden. Mikael Andersson; the Swedish National Archives/MKC

Abstract

The National Library of Sweden (KB) and the Media Conversion Centre (MKC), a part of The National Archives of Sweden, have created Digidaily, a unique inter-agency development project for mass digitisation of newspapers. The goal of the project is to create efficient methods and processes that allows for high quality mass digitisation of newspapers. In Digidaily, we jointly created an atmosphere that allows us to openly discuss, test, and evaluate both the processes and the technical specifications necessary for a development project of this size.

At the core of the workflow is the Workflow System, being developed by MKC in close cooperation with the KB. The Workflow System is used to control all parts of the workflow, and it needs to track both the location and the image capture of individual newspapers and pages, their physical condition, the resulting image files and OCR output, the collection of metadata, and the delivery of the digital files.

Digidaily – a newspaper collaboration

Digitising cultural heritage is currently a topical issue for many cultural institutions. Many countries began this work several years ago, but the large amount of material usually means that for reasons of costs and handling, only parts of collections or small collections can be digitised. These are some of the reasons why the National Library of Sweden has waited until now. The project Digidaily is a development project and collaboration across authority borders, in which the Swedish National Archives and the National Library of Sweden are developing rational methods and processes for digitising newspapers. Once the project is completed, we are hoping to transfer to a permanent operation and start digitising our entire collection of 122 million newspaper pages.

Background

After having itself managed and also participated in several projects relating to digitisation of large volumes of newspapers, KB soon found that mass digitisation of newspapers did not fit within KB's walls, either physically or organisationally.

At the same time, discussions began with the National Archives, which had set up a digitisation factory, Media Conversion Centre, "MKC", in Fränsta, Sweden, mainly to digitise church records. The operation is the largest of its kind in Europe, and the capacity is around 100 000 scanned images per 24 hours. The discussions formed the basis for an application to the Swedish Agency for Economic and Regional Growth for funding from the EU's structural funds. The application was approved and the project Digidaily started in April 2010.

The collection and the selection

The collection of newspapers at KB amounts to around 122 million newspaper pages. The collection consists of the so-called "Official National Copies", which are to be preserved forever. There is also a large collection of duplicates, and it is mainly these that will be used in the digitisation. The duplicates will afterwards be destroyed to give space for new incoming newspapers. In those cases where the Official National Copies are in a poor state, the duplicate will replace or supplement the torn National Copy. KB is thus taking the opportunity to take stock of and consolidate the collections.

The unique aspect of KB's collection is the large number of duplicates, which distinguishes the collection from many other library collections around the world. But having more than one copy to consider poses challenges to the project and raises a lot of questions, for instance: When should one mend an existing torn newspaper? When should a supplementary copy be looked for? How much time should be spent searching for alternative material? How should the handling of defects issues/pages be set up between the MKC and KB? How should supplementary material be handled in the metadata (file naming, etc.)? What is the borderline for rejection, how much can be allowed to be torn, what shall KB and its end users accept?

But the benefits prevail. The use of duplicates permits more efficient procedures, as the preservation aspect does not need to be considered when handling the material. For example, bound material can be cut open and separated. Scanner types that are not very delicate in their handling can be used, etc. And because most of the material is destroyed, the cost of return freight is lower. In the end, the use of duplicates will be noticeable in the overall price.

In the Digidaily project, we are mainly working with two well-known Swedish newspaper titles, Aftonbladet (1830–2010) and Svenska Dagbladet (1884–2010). The newspapers belong to Schibsted Media Group, which is also co-financing the project. If, at the end of the project, there is any spare capacity, the project group discusses and then decides on the newspaper titles that suit both KB's needs and MKC's production. KB takes into account the state and volume of the newspaper, whether it uses antique or Gothic font type, whether it has been microfilmed, whether the newspaper has a living owner and what demand there is from research, etc.

In addition, the project team tries to choose materials that are consistent with MKC's wishes in terms of categories of material:

- Category 1 - bound, torn, fragile paper, the biggest format size.
- Category 2 - bound, where most can be taken apart and only a few are kept still bound, fair paper quality.
- Category 3 - tabloids stapled but not bound.

- Category 4 – Official National Copies

The different categories also form the basis for the calculation of the end price of the digital page, as the different categories are treated differently in various ways during preparation and scanning.

The collaboration

Our experience to date from the collaboration between our two public authorities has been positive. Cultural differences and the physical distance (450km) between KB and the MKC are, however, aspects that must be considered. In the project Digidaily, we have worked hard at getting closer to each other, and we try to meet once a month for joint project meetings. We have also tried letting staff from each authority work at the other authority, with very good results. We carry out study visits together and in between times we keep in contact by telephone, email and meetings held via Skype. But meeting in person is quite clearly the most effective and rewarding way. In other words, a generous travel budget is of significance for a well-functioning project run at a distance.

An important part of the project is information. The project has a common platform for sharing information and documents called Projectplace. In this way, everybody can stay updated and read memos, time plans, requirement specifications and other important project documents. Projectplace also provides an opportunity to share desks during telephone conferences and, for example, review production and time plans.

Cultural differences are more difficult to overcome. KB is an academic public authority, which often works in a project format, while MKC is a highly efficient production unit, so collisions of culture do occur. But, meeting often and discussing can prevent misunderstandings.

In summary, it could be said that there are lots of positive aspects of working in a development project with another public authority. We have had time to work out and discuss a model that suits both authorities. The wish to maintain high quality in combination with keeping costs down permeates both authorities' attitude to the project, which is an important starting point for a successful collaboration.

Requirement specification

As the project is a development project, changes to the requirement specification during the course of the project are permitted. However, now that we are halfway through the project, any changes must be of such a nature that they entail significant improvement to the project in order to be taken into account. Too many changes, or large-scale changes, would have a negative effect on the project and the time plan for the project would be greatly disrupted.

For example, KB initially chose to save both an archive file and a display file, both in grey scale. Just over a year into the project, KB changed its mind, and chose to save only one file, an archive file. The amount of data KB saves in this way means that the newspaper can now be saved in colour (8 bits/channel). Saving all images in colour provides great added value for end users, for example because supplements are all colour publications.

In short, the end product is a colour page, with segmentation at article level. Manual segmentation or correction of automatic segmentation is not carried out, as the project is striving to use processes that are as automated as possible. Using rules, the CCM software can be adjusted to suit the newspaper it is segmenting. To a large extent, it is the skill of the operator that determines how accurate the segmentation is in the end.

The requirement specification states that the file shall be at most 300ppi, unless this has a negative effect on readability. KB wants the end user to be able to print out a page of acceptable quality. The general view is also that the resolution should be around 300ppi in order to get an optimal OCR result. MKC has commissioned Mid Sweden University in Sundsvall to look at and document how resolution and image manipulation impact on the OCR result.

The files are saved in jpg2000 according to a KB-specific specification. On behalf of KB, Karl-Magnus Drake at the National Archives has investigated the jpeg2000 standard's fulfilment of criteria for the static image format for long-term storage [1].

Production

Workflow system

The workflow system is the unifying tool that supports and directs production and processes within Digidaily; the workflow system could be called the spine of the project. The workflow system is constructed in modular form and is developed by a local team of developers at MKC. The process flow is a sequential flow, with status changes that drive the flow onwards.

The workflow system has the following functions:

- To be a database for information about the bundles, issues and pages of the material.
- To add metadata during the course of the production.
- To keep track of and initiate the next process in the flow with the aid of status codes.
- To collect data about the production and create documentation for planning and follow-up.

KB also has modules for its part of the operation. The material is registered already at KB with basic data, which will then follow the newspaper until the digitisation is complete. KB makes an export from its newspaper database, which is entered into the workflow system with basic information about the name of the newspaper, the start and end dates of the bundles and comments on supplements, editions, condition, etc.

It is also intended that KB shall be able to enter the system and trace the progress of the material, see the image files, extract statistics, etc.

Delivery

Using an annual delivery plan as the basis, MKC collects material from KB in Bålsta outside Stockholm. Special transport boxes have been developed for the transport of sensitive materials, such as Official National Copies. For material that is slightly tougher, KB's ordinary transport trolleys are used.

Documentation in the form of delivery notes and registered items in the workflow system accompany the deliveries. The Official National Copies have further identification, with documentation and receipts to safeguard their controlled return.

Archiving

The material collected is set up in MKC's incoming archive and the archive location is registered in the workflow system.

Preparation

The operator goes through the bundle issue-by-issue, page-by-page, and assesses the condition of each individual page. Three levels of divergent condition can be registered in order to communicate to KB that the condition of the page will affect the end result.

- **Level 1** - Light damage. Pale printing, small areas of loss, impact and/or small marks/stains in limited areas. The context of the article can be understood.
- **Level 2** - Severe damage. Areas that cannot be read even with the eye, and/or parts missing from the page, so that the article cannot be understood.
- **Level 3** - No original. The whole or at least half of a page or issue is missing.

KB will receive reports on the level of rejects via the workflow system and can then research whether any better original exists. In order for the flow of the material not to be disrupted, the damaged copies continue in the production chain. If a better copy is delivered from KB, it is scanned and then replaces the less good copy.

The operator supplements the information in the workflow system with information about, for example:

- The name of the supplement and section.
- The subject area of the supplement.
- The genre it belongs to – supplement, placard or section.
- The edition of the issue.
- The number of the issue.

Once the bundle has been gone through, an assessment is made whether it is suitable to separate the bundle into loose pages.

Image capture

For the moment there are two methods of capturing images – book scanning and wide format sheet feed scanning. For book scanning, the scanner models Zeutchel OS 14 000 A1 and A0 are used. For wide format sheet feed scanning, the scanner model SUPAG Mediascan 880c is used.

Technical quality control

In order to safeguard the quality of the file, a fully automated technical control is carried out on all files. For the technical control, data is extracted from the file and registered in the workflow system as background documentation for METS. If the file diverges from the quality requirements set, it is sent back to the image capture process for rescanning.

A performance file will be saved in the final package of the image file in order to guarantee quality. The performance file includes the latest measurement data from the quality measurement of the image capture equipment. Software manufactured by KB, Colorite [2], will be used for this purpose.

Creating jpeg

In order to ensure convenient handling, MKC creates jpeg copies of all files.

- **High-resolution jpeg**
  A high-resolution jpeg copy is created for the OCR software.
- **Low-resolution jpeg**
  A low-resolution jpeg copy is created for use in ocular control.

Ocular quality control

An ocular quality control is carried out to ensure the image capture has been satisfactory. The decision point for approving an issue initiates the start of two flows; the flow of the digital file and the continued flow of the physical issue.

Flow of the file

KB's archive file shall be in the format jpeg2000 and is created from the TIFF file, according to specifications from KB.

OCR interpretation process

The OCR result is the primary result of the digitisation. The resolution of the image is based on the quality of the OCR result. Testing of how the resolution impacts on the OCR result is being carried out by Mid Sweden University on behalf of the project. The project has not yet found an adequate way of setting requirements for the correctness of the OCR result.

The OCR process consists of three subsidiary processes – OCR, quality control and export. The workflow system creates a log file that directs the regulatory framework for different materials.

- **OCR**
  The originals are interpreted according to a regulatory framework.
- **Quality control**
  An audit is made of the OCR result for a statistical sample of the images.
- **Export**
  Once a satisfactory quality level has been achieved in the interpretation, ALTO files are exported from the program.

Creating METS

The METS file is created based on the data collected relating to the material in the workflow system. KB has clearly specified XML-schema for the METS-file. In collaboration with the National Archives, KB has produced a comprehensive metadata specification, which can also be useful for other digitisation projects within the cultural heritage sector.

Packaging deliveries

A SIP package is created for each issue. Each packet shall contain the following parts:

- jpeg2000/page
- METS/issue
- ALTO/page
- Performance file/issue

Delivering packets

The completed packages are sent via ftp to KB's storage system. Receipts are sent back to confirm the correctness of the packages.

Flow of the physical issue

Material to be returned is sent back to KB for further handling and archiving, and material not to be returned destroyed as instructed by KB.

Summary

The strength and success of the project lies in KB's and MKC's project groups being focused and having the same objective: ensuring quality can be allied to a competitive price. By working together, we can also benefit from the joint competences of the staff in an effective way. The chance of trying it out, in terms of both technology and procedures, has also resulted in an efficient workflow. And last, but not least, the spine of the entire project, the workflow system.

As an experienced production unit, MKC has detailed knowledge about all the subsidiary costs of the flow, which gives us a good tool for further efficiencies. For example, currently the average cost for preparation absorbs around 47% of total costs, and scanning around 39%.

Depending on the scope and condition of the newspaper material, the cost of a digitised page including OCR will be around SEK 3–8 (EURO 0.34–0.89).

For further information, please visit our blog Digidaily.

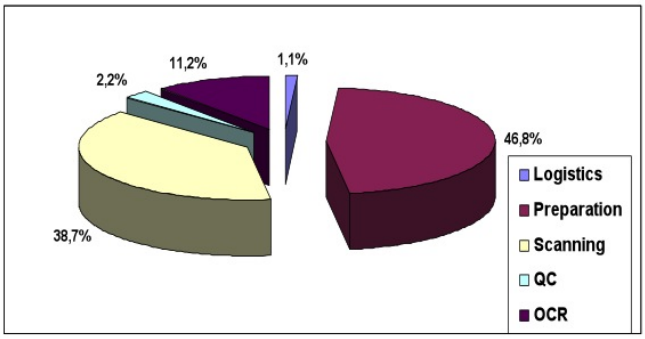

***Figure 1**. Relative shares of the cost*

References

[1] Karl-Magnus Drake, jpeg2000 – utredningsrapport [Investigative Report] version 2011-03-24

[2] Henrik Johansson, Automatic Image Quality Analysis of Arbitrary Targets with Colorite (IS&T, Salt Lake City, Utah, 2011).

Author Biography

Heidi Rosen has studied Graphic Arts Technology at KTH, Royal Institute of Technology in Stockholm. She is currently working as a project manager at the Newspaper Division at the National Library of Sweden, where she is responsible for the library's part of Project Digidaily.

.http://digidaily.kb.se/