TITLE: Experiences from Digidaily – Inter-Agency Mass Digitisation of Newspapers in Sweden

The project Digidaily is a development project and collaboration across authority borders, in which the Swedish National Archives and the National Library of Sweden, "KB", 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.
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 amounts can be digitised. These are some of the reasons why the National Library of Sweden has waited until now with our digitising.

Background

KB has during the years both managed and 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 in 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 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 to get 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 to share information. The project has a common online platform for sharing information and documents called Projectplace. In this way, every participant in the project can stay up to date 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.

The collection

Swedish newspaper publishers deliver three legal deposit copies of all Swedish newspapers printed. One copy stays at KB, one goes to Lund University Libraries and one goes to a company called A2D for microfilming. KB has 31 600 meters of newspapers or approx. 122 million pages out of these are 70 million pages on microfilm.

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 for digitisation. KB is taking the opportunity to take stock of and consolidate the collections, so 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 and will therefore be kept.

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 meta data (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 outweigh. 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 those cases the official national copies need to be used, KB now considers if it is possible to allow a "gentle dismantling" of the bound material. This would allow a more rational production and therefore a lower price. Strict rules when dismantling isn't allowed must however be followed.

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 more than 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 an example KB chose to change a major requirement a year into the project.
KB initially chose to save both an archive and a display file, both in grey scale. After a lot of considering, KB changed its mind, and chose to save only one file, an archive file. The amount of data KB saves this way means that the newspaper pages now can be scanned in colour (8 bits/channel) instead and saved in jpeg2000. Saving all images in colour provides great added value for end users as for example most supplements are colour publications.

In short, the end product is a colour page, with segmentation at article level. Manual segmentation or correction of automatic segmentation will **not** be 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 specific 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. Correction of for example headlines and other text blocks will not be done either. Here as well we strive to have an accurate and automatic work flow.

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]

Regarding the metadata, a Swedish METS profile has been created. The METS profile is a result of collaboration between the National Library of Sweden, the National Archives of Sweden and other Swedish archives. The basis is a choice of metadata standards such as METS, MODS, PREMIS, MIX and ALTO. This METS profile can also be useful for other digitisation projects within the cultural heritage sector.

---

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

MKC will create a package, one for each issue, containing all the object files and one METS file with metadata about the package. All packages will be delivered to the National Library by FTP and stored in Mimer (new technical platform for handling digital collections of the national library). Mimer is still under development.

Comment [t1]: Paketet innehåller mer

Overall planning

In the Digidaily project, we are mainly working with two well-known Swedish newspaper titles, Aftonbladet (1830–) and Svenska Dagbladet (1884–). 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 groups from KB and MKC will discuss the matter and then decide on a suitable newspaper title that will suit both KB's needs and the MKC's production.

KB will take into account the state, size and volume of the newspaper. Whether it uses antique or Gothic font type, if it has been microfilmed, the legal rights of the newspaper and what demand there is from research.

Comment [t2]: Jag hanger inte med vad är det vi vill tala om här?

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

Delivery plans are worked out between KB and MKC in order to fulfil to the needs of both organisations as much as possible.

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.

Also KB has modules for its part of the operation. The material is registered already at KB with basic data, which then will 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.

Both MKC and KB will be able to enter the workflow 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.

Each case is followed by a printed packing list from the Workflow system. 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. Official national copies are kept in the transportation cases in the incoming archive to minimize handling of the material and ensure secure archival keeping.

Preparation

The preparation process consists of two sub-processes - *Go through* and *Take apart*. The process has two purposes. One is to capture and record metadata in the bundle, issue and page, and the second is to prepare the material for an efficient image capture.

Go through

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** - Lightly damaged. 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** - Severely damaged. 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 or the whole or at least half of a page is missing.

KB will receive reports of the level of rejects via the workflow system and KB can then decide whether or not to search for any better copies.

In order for the flow of the material not to be disrupted, the damaged copies continue in the production chain. If a better page/issue is delivered from KB, it is scanned and then replaces the less good page/issue.

The operator also decides which scanning line type is appropriate for the material and checks the pre-registered information in the Workflow system in terms of date and number of the issue, and how many pages each part contains.

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

- Name of supplement and/or section
- The genre it belongs to – Supplement, news bill 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. Destruction or return copies can be taken apart if the text information can be assured afterwards.

Take apart

The operator takes apart the bundles with great caution. Knives are used to divide the bundle into smaller piles, and then an electrical cutter is used to get an even and smooth cut surface.

Image capture

For the moment there are two methods of capturing images – book scanning and wide format sheet fed duplex scanning
For book scanning, the scanner models Zeutchel OS 14 000 A1 and A0 are used. The model allows image capture of two pages at the same time, which can be divided into separate images.
For wide format sheet feed scanning, the scanner model SUPAG Mediascan 880c is used. The model allows image capture of double-sided pages and the entire spread. Front and back pages are scanned in one feeding. Spreads are being divided into four separate images. A function to number and organize files on 4-page scanning is also available.

Technical quality control

In order to safeguard the quality of the file, a fully automated technical control is carried out on all files. From that process metadata are lifted out of the data file and recorded in the Workflow system as a basis for METS.
In case of deviation from the established quality requirements, the file will be returned to the image capture process for re-scanning.

A performance file will be saved in the final package 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 file is created for use in the ocular quality control process.

Ocular quality control

An ocular quality control is carried out to ensure the image capture has been satisfactory. The operator examines 100 per cent of the images visually. A future development of the ocular control aims to use statistically methods to extract significant number of images that will be examined visually. If the quality control shows that the quality of the image does not meet established quality levels, the image is reported for re-scanning.

Approval of image capture

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.
For the digital workflow the creation of a *Jpg2000* file is initiated. For the physical issue flow the delivery process for the *Return* copy and *official national copy* is initiated.

File flow

KB's archive file shall be of format JP2000. The JP2000 copy is created from TIFF, 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 content conversion software used for the segmentation and OCR interpretation comes from the Norwegian company Zissor.

The OCR process consists of three subsidiary processes – *OCR, quality control* and *export*. The Workflow system creates a log file that controls the layout analysis set up of the different materials.

---

[2] Final Program and Proceedings – Archiving 2011. Automatic Image Quality Analysis of Arbitrary Targets with Colorite

- **OCR**
  According to the layout analysis set up, the images are being interpreted. No manual article segmentation is made in project Digidaily.

- **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 for each issue is created from the data collected about the material in the Workflow system.
As mentioned before, KB has clearly specified XML-schema for the METS-file.

Packaging deliveries

A SIP package is created for each issue. Each packet should 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. A delivery receipt with information if the delivery is approved or disapproved is sent to MKC.

Flow of the physical issue after digital delivery

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

Re-delivery to KB

For the copies that are to be returned to the KB, the period of retention at MKC should be kept as short as possible. Material that has been taken apart are packed into capsules before labelling and delivery. A pack-list is being extracted from the Workflow system and delivery notes are attached to the packages. Official national copies have additional registration and receipts to ensure the delivery procedures.
According to the established delivery schedule, material is returned to KB for further registration and archiving. The re-delivery to KB initiates the start to reform and strengthens the collection. Better copies will either replace damaged official national copies or the decision will be made to keep both it they are in a very poor state. The material will be stored for final archiving.

Destruction Process at MKC

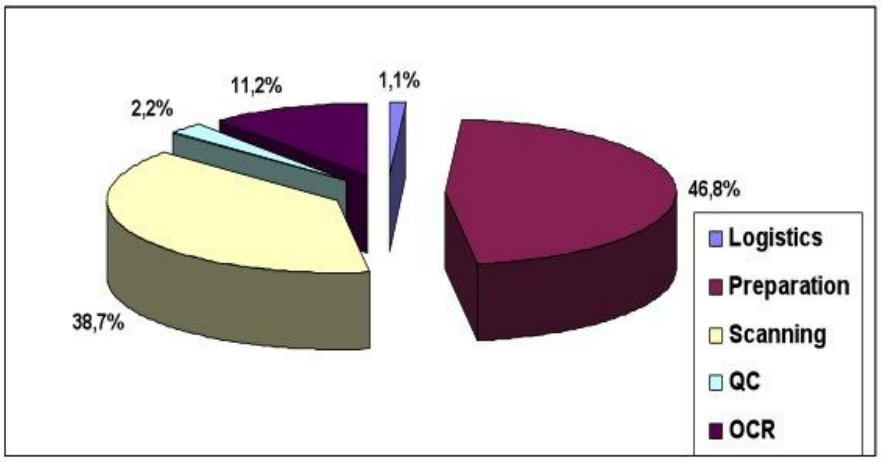
Once delivery of digital images has been approved the destruction process begins. To ensure that the right material is sent for destruction, strict rules and protocols must be followed. A recycling company will collect the material for further destruction.

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 that 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 lastly, but not least, the spine of the entire project, the workflow system, which makes it possible for both MKC and KB to keep track of every single page.

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 39%. The better quality of the material, the lower preparations costs and therefore also a lower total cost.

Relative shares of the cost



Depending on the scope and condition of the newspaper material, the cost of a digitised page including OCR, will be from around SEK 2,30 (Category 3 - tabloids stapled but not bound), SEK 3,40 (Category 2 - bound, where most can be taken, fair paper quality) up to about SEK 10 (Category 1 - bound, torn, fragile paper, the biggest format size A0).
In Euros the price span will be around € 0.25–1.11.

For further information, please visit our blog Digidaily.
http://digidaily.kb.se/