Digitization of copyright protected newspapers in Sweden

Torsten Johansson and Heidi Rosen - National Library of Sweden - Kungliga Biblioteket

There is no doubt, that t media today - television, radio, newspapers, Internet and mobile phones - are representing a large and indispensable part of our everyday live. Many of us have highlighted media immediately reasoning how our society has developed in a certain direction in the public debate. Compared to the development in this area you can see clearly, that there is a close relationship between media and society.

Today media are indeed an inseparable part of the structure of the society in the same way as the surrounding social structure is influencing the individual media structure and content. Therefore the opportunity is incredibly important, that science is free, searchable and integrating this material .It is also very important for the scientific community to have access to a critical mass in order to get a reliable result in their research.

The National Library of Sweden - Kungliga Biblioteket (KB) has managed both during the years and participated in several projects relating to digitization of large volumes of newspapers. The KB realized already in 1998, that digital technique and the transformation of images to machine readable text could provide new facilities to access the newspaper collection. From 1998 to 2004 the KB collaborated with the national libraries of the four North-European countries in the project "*Tiden*" focusing the application of OCR and search engines for newspaper digitization. When the project ended we had only a small amount of digitized newspapers, but a lot of experience from working with OCR. However the results from the project could not be directly implemented in the daily work. In 2008 we began to develop a small in house digitalization line for newspapers. This endeavour resulted in about 400 000 digitized pages, which can be accessed today on http://magasin.kb.se.

When analysing the result of these projects we realised that mass digitization of newspapers did not fit within the library walls - neither physically nor organizationally. We had to find a partner, who could transform our and their own ideas to a more industrial production, if we were going to achieve our goals.

From 2010 to April 2014 we received economical support from the EU's structural program and the local government Västernorrland to plan and implement a production line specialized for digitalization of newspapers. During this period we established an infrastructure for mass digitization of historical and modern newspapers. The goal of the project was to create efficient methods and processes enabling high quality mass digitization, i.e. finding ways to harmonize quality and quantity.

The collection

The newspaper collections of the KB are built up by legal deposit. All Swedish newspaper publishers have to deliver three legal deposit copies of all printed newspapers. One copy remains at the KB and one goes to Lund University Library. The last sample is being used for reproduction.

The newspaper collection surrounds nearly 40 000 shelf meters with approx. 130 million pages. The main part of the collection consists of the official national copies, which must be preserved "forever". But the collection also consists of a large collection of duplicates, which have to be used for reproducing. The National Library is taking the opportunity to take stock of and to consolidate the collections. As soon as the reproduction has been done, the duplicates are being destroyed to give space for new volumes. In case, that the official national copies are in a poor state, the duplicate will replace or supplement the torn national copy and will be kept. This unique aspect of the collection is distinguishing it from many other library collections worldwide.

A small part of the newspaper collection. Photo: Jens Gustavsson

For many years the library used microfilm to preserve the content of the newspapers. Out of the total amount of 130 million pages approx. 70 million pages are reproduced on microfilm. This means that nearly the half of the collection is microfilmed. A private company carried out this production.

The benefits of having a duplicate, which could be used for reproduction, is permitting us to use more efficient working procedures, because the preservation aspect must not be considered. Bound material can be cut off and separated. Scanners can be used, which are working in a non-preserving way. At last the use of duplicates is noticeable in the overall price. If the official national copies have to be used, the KB allows a "gentle dismantling" of the bound material, which enables to produce more efficiently production and reduces the costs. However there are strong rules in regard of dismantling, if the newspapers could be jeopardized.

The Quality

The continuously growing internet is raising the demand the ability of cultural institutions to deliver digitized materials. Not so long ago the internal production lines managed the demand, but today the demand of making digitized material available is so large, that many institutions no longer can cope with digitizing within their own ranks. In recent years, many institutions got into so called mass digitizing projects to be able to live up to the high demands on the amount of material from the public and the government.

Talking about digital reproduction is talking about photography. This field have changed over the last twenty years when the digital technique has been taking over the field. In the mind of the people the tendency seems to exist, that digitizing materials is just to press the green bottom and after some seconds the perfect image will be available. What has changed?

In the analogue world we had educated photographers carrying out this work and none of us contested their knowledge about light, optics and chemistry. The quality was part of the performance of their work. But under digital perspectives we obviously suggest, that the skills and the physical laws are different. But this is wrong. Since photography still includes the question of light, optics, chemistry and physical laws. Unfortunately it is hard to find skilled photographers with knowledge of photographical theory today. The change is odd and dangerous, when we are talking about digital reproduction; because bad reproductions are influencing findings and research results.

Of course you can digitize very large volumes in a short time. But why you should do it, if the quality is poor or hardly acceptable? And who is deciding, what acceptable quality really is - the customer or the supplier? As a client you often hear, that you shouldn't expect perfect results, if large newspaper volumes have been digitized. As a national library we do have high quality standards - even in the in house digitization procedures. From that background we don't believe, that a high quantity is automatically connected with lower quality. For example a car buyer would never approve, that his brand new car has scratches in the paint, even if the car was made on an assembly line. Of course not, and we won't either. Digitizing newspapers always involves large volumes and inevitable high costs. In our case we may never have the opportunity to digitize the newspapers again. From that it has to be done right straightaway.

It was important to define a level of quality for our reproduced newspapers. The quality level aims to determine the technical parameters for the capture and define the limits, which are varying depending on the material and the chosen quality level. The technical parameters could be e. g. exposure, gain modulation, colour reproduction and blur. The quality level aims to determine the technical parameters for the capture and define the limits of variations e. g. parameters such as proper exposure, gain modulation and colour reproduction. We decided to use three – metamorphosing - quality levels.

1. Items with high demands on colour accuracy
2. Items with standard requirements for colour accuracy
3. Items that are digitized in grey scale

The newspapers are reproduced under level 2. All digitization should, be done in a controlled environment with a calibrated equipment to minimize deviations between the imaging and source documents. Our methods combining quality with quantity are based on procedures automated as much as possible. It is the software and not people that will find the deficiencies of quality.

The process

In April 2014 the development phase ended and the work has moved to real production. Apart from the inevitable initial problems the developed systems run as planned now. Thereby it has been resulted, that KB ended the use of microfilm as a reproduction method in 2014. Today we are digitizing all the newspapers delivered by the legal deposit.

The heart of the process is the workflow system, which is the unifying tool supporting and directing the production processes. The modularized system could be called the spine of the chain. The workflow is implemented sequentially with status changes, that are driving the flow forwards.

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.

Scanning newspapers. Photo: Heidi Rosen

The KB has specific modules as part of the entire process. We make an export from our newspaper database, which is integrated in 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, conditions or remarks, if any part of the paper is missing.

The Access to the digital material

The KB has developed and implemented an interface for presenting the digitized newspapers: http://tidningar.kb.se. Today this presentation contains approximately 10,5 million pages. This is approximately 7% off the total collection (figure 1).

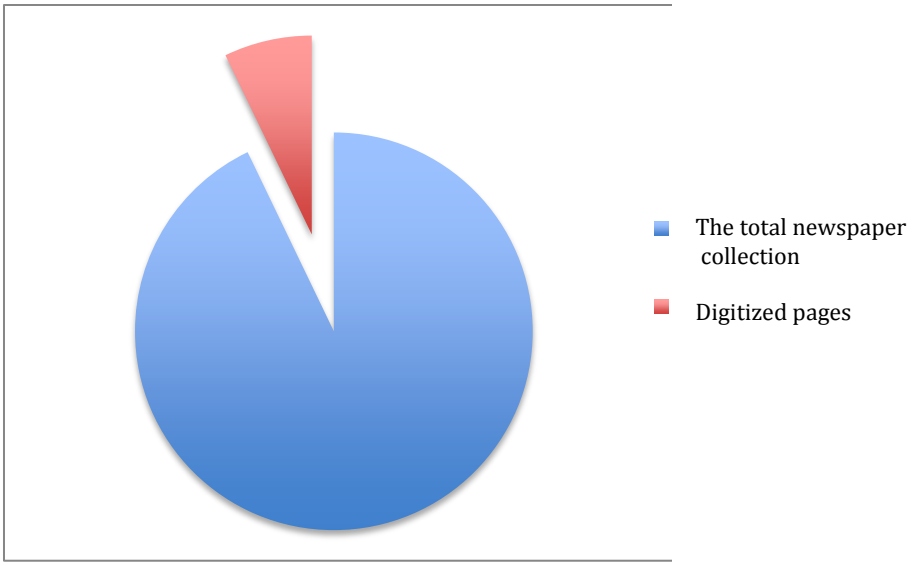

Figure 1: The amount of digitized pages in relation to the total collection

Inside the library building the users have access to all the digitized material, because the library has an exception within the copyright legislation that allows the digitization of collections for preservation issues. But presenting the digital material to the user outside the library building is a hard nut, because of the copyright legislation. Today users outside the library can look at material up to 1899. This is 1,7% of the digitized material (figure 2) and 0,23% of the entire collection (figure 3).

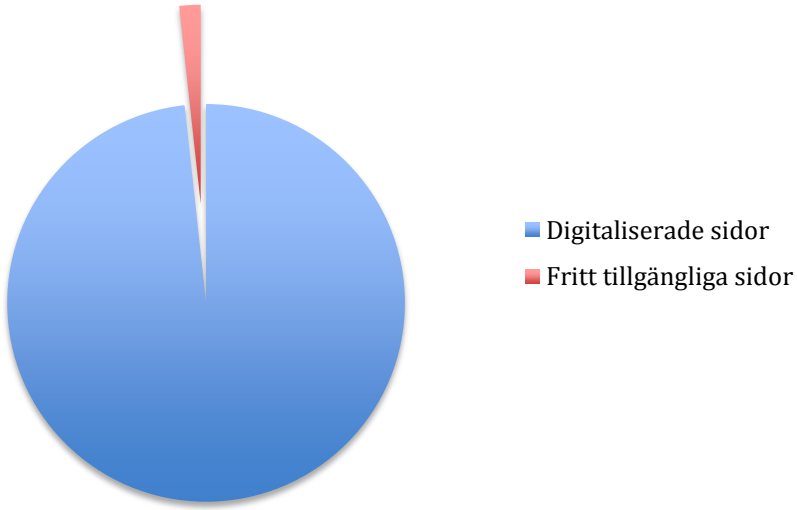

Figure 2: The amount of copyright free material seen in relation to the amount of pages we have digitized.

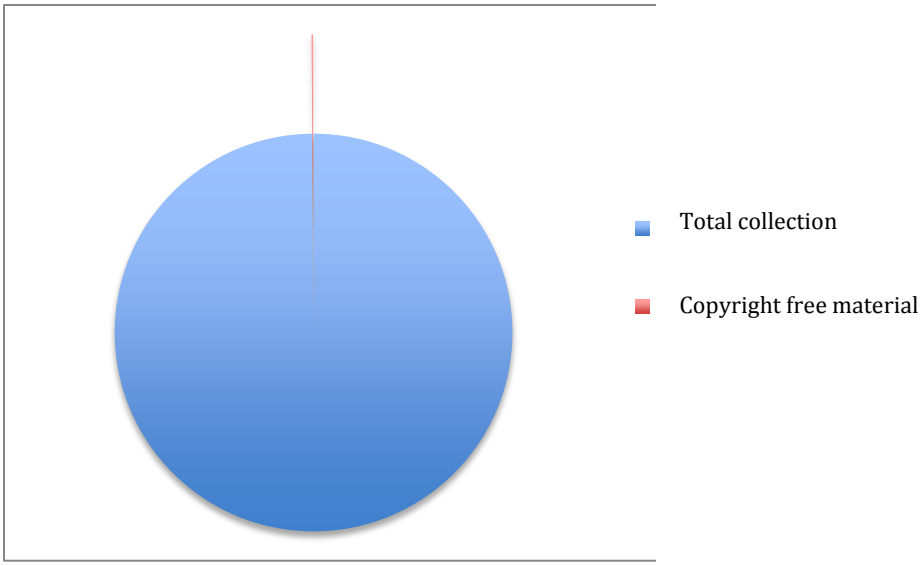

Figure 3: The amount of copyright free material seen in relation to the total collection

Of course the KB is willing to digitize a lager part of the historical newspaper collection, but it is hard to get funds. We are trying to get grants from foundations and through corporation with newspaper companies. But this is difficult. The field of digital newspapers is still young and isn't fully established yet. At the same time the traditional market for printed newspapers is decreasing. Everybody is uncertain of what is coming next. The publishers are afraid of losing market shares, Therefore the situation is stagnating entirely.

References

1. "Colorite: A Flexible Cross-Platform Software Solution for Automatic Image Quality Analysis Using Arbitrary Targets", Henrik Johansson, the National Library of Sweden, Archiving 2011, ISBN / ISSN: 978-0-89208-294-0, p. 199-204, 2011
2. "Digidaily Inter-Agency Mass Digitisation of Newspapers in Sweden"
   Heidi Rosen, Torsten Johansson, and Henrik Johansson, the National Library of Sweden; and Mikael Andersson, the Swedish National Archives/MKC (Sweden); Archiving 2012, ISBN / ISSN: 978-0-89208-300-8, pages 126-129} 2012
3. "Experiences from Digidaily: Inter-Agency Mass Digitization of Newspapers in Sweden", Heidi Rosen, Torsten Johansson, Mikael Andersson and Henrik Johansson; The Memory of the World in the Digital Age: Digitization and Preservation, Edited by: Luciana Duranti and Ehizabeth Shaffer, pages 1 153 — 1 161, 2012
4. "How reproductive is a reproduction? Digital transmission of text-based documents"; Lars Björk, ISBN 978-91-981654-6-3, 2015
5. "To Harmonize Quality and Quantity", Heidi Rosen and Torsten Johansson, The National Library of Sweden; Archiving 2013, ISBN: 978-0-89208-304-6, pages 126-129, 2013