Microfilming for Digitisation and Optical Character Recognition

Supplement to Guidelines for Newspaper Preservation Microfilming

Introduction

The Guidelines for the preservation microfilming of newspapers were published by IFLA in 1996 and mounted on the IFLA website.

The Newspapers Section (formerly the Round Table on Newspapers) has planned to issue a supplement to these Guidelines in recognition of the need to ensure the best possible future access to texts captured on microfilm, when digitisation is undertaken.

This work has received some points of significant impetus recently:

- the rapid growth of the Internet, which enhances the means of communicating newspaper texts, both present and past.
- the greatly increased ability of the machines available, which, together with enhanced software packages, have the capacity to secure good quality results in rapidly capturing texts in digital form.
- the reduction of the costs of storing digital data, together with the means of storing far larger quantities of digital data than previously.

At a time when technical developments change our perceptions of what is possible, year by year, the Newspapers Section considered it appropriate to issue this guidance document: Microfilming for Digitisation and Optical Character Recognition. Microfilming remains for many libraries the preferred medium for the preservation of texts. Optimising access to microfilms may be the most suitable means to open up access on a worldwide basis, at some future point, to the huge quantities of newspapers, some of which are now held on microfilm already, and some of which have yet to be microfilmed.

Digitisation has increased our means to make newspapers more readily available by adding advanced search features to access texts. Because of their heavy use, newspapers are seen as an important part in the libraries' digitisation programmes. Digitisation puts special requirements on the quality of the microfilm.

Digitisation of newspapers by using microfilm as the intermediary, means higher quality requirements. In the Nordic TIDEN project Helsinki University Library, together with the Royal Library of Sweden, has tested the impact of microfilm quality on digitisation and optical character recognition. Since the results might also be interest to a broader audience, the project participants (in addition to the two libraries mentioned also the National Library of Norway and the State Library of Aarthus, Denmark, participate in the project) have contributed by drafting guidance for the IFLA Round Table on Newspapers for "Preservation Microfilming for

Digitisation and OCR". The Nordic TIDEN project has been digitising Nordic newspapers mainly in Gothic but also in Roman font, from 1640 to 1900 by using microfilm as the intermediary. In addition to the comments of members of the Round Table collaboration with the Bibliothèque nationae de France and the British Library has to be mentioned in particular. The results will be incorporated in the "Guidelines for Newspaper preservation Microfilming", published already earlier by the IFLA/Round Table on Newspapers.

Scope

Improved access to data stored on electronic media and the long-life expectancy of microfilm are a good combination for preservation and access. Microfilm can last up to 500 years, which secures the archival safety of texts, which have been transferred from brittle paper. Digitisation also offers easy access. But microfilm can be used as more than an archival medium. It is also as an efficient and economically favourable intermediary in a digitisation programme if this aspect is considered during the microfilming process.

The aim of the guidance is to show how to use microfilm as a platform for future digitisation and to enable full text search in the digitised collections. because the development is very rapid, it has to be emphasized that the suggestions in the IFLA guidance "Preservation Microfilming for Digitisation and OCR" reflect the present situation. New methods will certainly emerge.

Microfilming and digitisation can be combined in several ways. It is possible to microfilm first and to digitise from film. But it is also possible to digitise first and to preserve the contents on film by using the COM method (Computer output on microfilm). There is also a possibility of using hybrid microfilm-digitisation cameras for simultaneous microfilming and digitisation. For newspapers the first option, microfilming first, is by far the most used and it is also the method, which is discussed here more in detail.

It has to be stressed that the quality of the microfilm is crucial for the quality of the digitised images. Advanced search facilities in digitised newspaper files require high quality results from the OCR treatment, which can be achieved only with the help of high technical quality. In other words, the output depends on the input.

Some suggestions for enhanced quality

Libraries have to decide whether a microfilming program based on higher quality demands has to be part of the library's preservation and digitisation policy. The level of quality is to be chosen according to the library's needs. It is most important to follow the standards set for microfilming and to use e.g. the IFLA "Guidelines for Newspaper Preservation Microfilming". If a library would wish to implement a largely automatic, quality-based digitisation and full text search in its microfilming program, this can also be done.

Features of the microfilm

High-contrast microfilm is a good feature especially comparing the scanning result of the originals and the microfilms of grey text on brittle and brownish paper. The text is clearly distinguishable from the base. Smaller discolorations, spots or wrinkles in the paper are eliminated. This offers full text search capability when optical character recognition (OCR) software is used on bitonal images. The resolution is quite sufficient from low to 16x reduction rates favours a Quality Index of 18. When reduction rates exceed 20x, OCR may not be possible.

However, photographic material will never be as good as digitising in greyscale from the original. There is also an obvious difference between the different kinds of microfilm copies, while the negative direct duplicate produces the poorest digitisation from original photographs printed in newspapers.

Results from the Nordic TIDEN project

In the TIDEN - The Nordic Newspaper Digitisation project-Helsinki University Library and the Royal Library of Sweden compared the results of OCR treatment of newspapers from the early 19th century until the late 20th century which were digitised from 235mm microfilms. Four factors we found, which influenced the success or failure of OCR-reading in particular:

- The quality of the text of the original; complex layouts and paper quality
- The reduction ratio of the camera
- The font and the font size (Roman fonts give good results whereas the older Gothic text demands skill and extra effort. Multiple fonts including Gothic text might result in incorrect text interpretation).
- The language (Swedish gives better OCR-results than Finnish. The OCR of multiple languages mixed in the same are or sentence is harder or impossible to interpret in comparison with the interpretation of one language only).

The influence of the brand and resolution of the camera, the polarity and the generation of the film (first and second generation) was marginal. It should be noted however, that old microfilms made in the 1950's and 1960's may not be sufficient for OCR-reading, while films from the 1980's based on international standards are good enough for OCR.

Optimising scanning from microfilm

It is possible to get equally good full text search from texts, which are scanned from microfilm as from scanning from the originals. However, the potential of the film is not always exploited in the best way of lack of good laboratory practice during preparation and microfilming procedures. Improving the overall standard needs a microfilming program that can be implemented throughout the production.

1. The quality of the originals is crucial both for microfilming and digitisation. That is why best possible originals should be chosen. Extremely brittle material should be handled under the supervision of conservation staff. The newspapers should be unbound. Bound volumes should be filmed on a book cradle under glass (in this case the margin should be filmed on a book cradle under glass (in this case the margin should be 1.5cm in width from the spine gutter to the text).

2. Microfilm scanners may have problems with identification of varying page sizes on the film. This is because the scanners usually recognize a new frame at place indicated, at its left corner or margin. In this case identification could be helped by marking the left side of the frame area of the camera table with a tape strip. This strip could further be developed to a bar carrying machine readable information about the images, like newspaper issue/number (page), new edition, supplement, page repetition etc. The other option is to implement computer software to deal with the questions mentioned.

3. When microfilming for digitisation is carried out, even lighting within each frame and throughout the film is important and should be within the range 0.2 within each image and the film. Any change of density within the frames will need special handling from the scanning operator, or the quality of the OCR-reading may be diminished. The varying density between exposures requires manual adjustment of the scanning parameters for each image. This is more costly and time consuming than continuous automatic film scanning.

4. In older films the film-area between the frames can be too narrow for the scanner to separate the frames. Equal film-areas should be used.

To sum up the suggestions, the international and national standards for preservation microfilming are the basis for a successful digitising from film. Filming from unbound volumes and the reduction ratio of the microfilm are most important points to be kept in mind in striving for OCR-reading reproducing the quality of the original. Density values should be narrow enough to allow for the automatic scanning process. A white tape strip or bar should indicate the beginning of each image. Blipping, or machine-readable indexing by barcodes for the newspaper issue/number (pages) and the microfilm targets could be used to further automate the whole process from digitisation to mounting the digitised texts on a database etc. Computer software might also handle processes of scanning and indexing.

Economic Issues

Implementation of the quality program for preparation of the publications for microfilming and the microfilming itself will need additional financing in the beginning. This is needed for planning and implementation of both parts of the process both in the library and the microfilming agency. Planning has to produce a microfilming program, which is the basis for future digitisation programmes.

Higher financing might be needed e.g. by changing to lower reduction ratios and unbinding the newspapers. If barcodes are used, changing the barcodes will require special attention of the camera operator and might slow down the speed. Setting new density readings may rise the quality demands and cause re-filming decisions. New equipment might be needed for the filming of bound volumes on book cradle under glass. Many of these changes have, however, already implemented in a number of libraries and can be combined with ongoing routines.

Even if the initial cost may be higher, the high-quality microfilm along with machine readable information allows for easier and cheaper digitisation process. Savings will eventually be recurrent because technological advancement and the demands of the users may require re-digitisation at higher resolution/bit depth than is possible at the moment. having produced a high-quality microfilm there is no need to go back to the original which, in the worst case, may no longer exist. Do it once, do it right.

Technology will surely remove some, if not all, of our present obstacles. If a quality program for microfilming is established as a part of the library strategy, it will make things easier for the future digitisation activities. Microfilming should be considered as a part of the library's digital program.

December 2002