

Digitising Newspapers from Hard Copy

Author: Rose Holley (Manager – ANDP)

Version: 1.0

Date: 5 October 2007

Most national newspaper projects have digitised from microfilm rather than hard copy newspapers. This is because preservation of newspapers has been carried out by microfilming and usually once the microfilm is created the original hard copy newspapers are discarded. The newspapers are discarded because of the large storage requirements and difficulty of handling and sourcing for users of hard copy volumes. For significant historic national titles from the 1800's onwards often only the microfilm is available.

Australia has for a number of years been preserving newspaper titles by microfilming through the ANPLAN project www.nla.gov.au/anplan. There are therefore not that many newspaper titles that have not been microfilmed. Digitising from microfilm is much quicker than from hard copy due to the handling of the resource and the scanning equipment used. Up until fairly recently it was not possible to get good enough image results for OCR from scanning hard copy newspapers. This was because of the large size of the newspaper page and the smallness of the text, and there being no scanners capable of reaching the resolution required. No other national newspaper program is currently undertaking large scale digitisation of newspapers from hard copy so if the National Library was to plan for this more research would be needed since there are a number of decisions which would need to be made and factors to take into consideration.

Digitisation from hard copy has been suggested because some of the microfilm is too poor quality to result in good digitisation (this is mainly microfilm from the 1960's and 1970's), or when originally created it was made from incomplete and poor quality originals. If hard copies are still existing they could be re-scanned. There are also some regional titles which have not been microfilmed.

It is possible to:

- Microfilm hard copy newspapers and then digitise (scan) the microfilm
- Digitise (scan) hard copy newspapers and then create a microfilm if required.

There is still a desire to create microfilm since this is the accepted long term preservation format (digital image is not). However this is currently under discussion by ANPLAN with the advancement of digital technologies and digital preservation. Traditionally many libraries through out Australia would buy copies of newly microfilmed newspapers for their users, but in the future once the digital newspapers are freely available via the ANDP there may be no demand from libraries to buy microfilm copies, which means the only purpose for creating them would be for preservation, not access. If the digital image is the accepted preservation media the logical step would be to scan any remaining hard copy newspapers direct to digital format and not create the microfilm as well.

Factors to consider if digitising hard copy newspapers in order to OCR them and then make them available as full text searchable would be:

- Sourcing copies (need the cleanest most complete copies)
- Collating copies (Best copies may not all be at the same location, how to compare and collate copies if stored at different locations? How to do this if they are in off-site storage?)
- Logistics (cost, method, time of transporting large volume of newspapers to scanning facility)
- Handling and storage (unpacking, flattening, sorting, sequencing newspapers)
- Type of scanner (would have to be an overhead colour scanner that could scan at 400-600 dpi resolution to greyscale and bi-tonal tiff file).
- Contractors (Are there any contractors in Australia that can do the volume of work with suitable equipment and storage facilities?)
- Cost (the handling and transport is likely to cost more than the scanning)
- Time (how long would it take to scan a particular title?)
- Quality Assurance (how to check image quality against originals – logistics)
- Image manipulation for OCR (this may need to be a separate stage to scanning depending on the scanning software capabilities. Traditionally image optimization for OCR is built into microfilm scanning software not reflective scanner software).

The National Library is aware that the State Library of New South Wales is currently creating new microfilm from hard copies of the Sydney Gazette, and that the State Library of South Australia has experimented using a hybrid scanner to create microfilm and digital image simultaneously from hard copy newspapers. Useful information has been learnt in both cases. Further investigation and feedback and collating of experiences from State and Territory Libraries around Australia is required before the ANDP can make any recommendations about digitising hard copies of newspaper pages.

NB: If local historical societies or other organizations are already digitising from hard copy the most important things are:

- To flatten out folds and creases as much as possible
- To scan at the highest resolution available (400-600 dpi)
- Scan as greyscale (and save as bi-tonal as well)
- Use unique file name e.g. running number
- To save the file as a tiff for preservation, re-use, participation in the ANDP (store on CD/DVD, portable hard drive if not enough server space).
- Refer to the microfilm scanning specification on the ANDP website for more detail.