

Outline of Workflow Processes

Author: Rose Holley (Manager – ANDP)

Version: 2.0

Date: 20 August 2008

Text to accompany diagram on final page.

1. Microfilm is sourced

Newspaper microfilm is sourced from State and Territory Libraries and sent for scanning.

All State and Territory Libraries are contributors to the program. W & F Pascoe Ltd also own microfilm master.

2. Creation of digital images from microfilm (Scanning Contractor)

Digital images are created by scanning microfilm.

W. & F. Pascoe Ltd. - a microfilm Bureau based in Sydney - are providing scanning services for the Program. The hardware being used is NextScan Microfilm Eclipse scanners with NextStar software. The digitised images are then copied to LTO2 data tapes and are sent to the National Library for quality assurance. Microfilm is returned to libraries.

3. Quality assurance work - part 1 and 2 (Library)

The QA (quality assurance) process is a very important aspect of the Program. It is carried out at the National Library of Australia, and includes automated and manual processes as outlined below:

Part 1:

- Addition of metadata (title of newspaper, issue date, page numbers, notes)
- Sequencing of pages

Part 2:

- Identifying missing newspaper pages and issues and creating targets for these
- Removing duplicate pages
- Grouping digital images into batches of 2000 for OCR processing

Once the images have been quality assured, they are then ready to be sent to the OCR Contractor in India. If there are any issues with digital images they will be returned for reprocessing by the contractor.

4. Zoning, categorising, OCR, rekeying text, and metadata (OCR Contractor)

Apex CoVantage are contracted to complete zoning, categorisation, OCR and rekeying work which takes place in their production facility in India.

Zone articles:

- Each newspaper page is zoned into separate articles
- The coordinates of the zones on the page are recorded in the Alto/METS file

Categorise articles:

Each article is assigned a category from the list below (which will assist with excluding/including items in searching):

- News
- Advertising
- Family Notices
- Detailed lists, results, guides

In addition illustrated articles are categorised as:

- Illustrations
- Photographs
- Cartoons
- Maps
- Graphs

Optical Character Recognition (OCR) on articles

Each article is converted into a full text file by having Optical Character Recognition (OCR) software automatically convert characters in the image into full text searchable words.

Re-keying of text in articles

The titles, subtitles, authors and first four lines of text in the articles are rekeyed manually to achieve 99% accuracy.

Metadata

An XML file in ALTO format is created for each page and a METS file for each issue is created by Apex. The ALTO file contains the results of OCR including position of zones and words on the page. The METS file (one per issue) contains the re-keyed data (title, abstract etc) and structural information about the pages and articles.

5. Quality acceptance work - part 3 (Library)

The processed pages and articles are quality assured by National Library staff to check that they meet the quality acceptance criteria % as defined in the Contract. They are either accepted by the Library or sent back to the contractor for re-processing. The XML files are ftp'd to the Library. The Library creates derivative images from the master greyscale TIFF images for use in the public search and delivery system (JPEGs, and PDFs and thumbnail images).

6. Public Search and correction of data

The data is loaded into the search and delivery system and is available to the public. Public may correct the OCR text and if they do so the changes are saved in the database and available to others. This improves the accuracy and therefore the searching of articles for everyone and is a valuable contribution the public can easily make.

Workflow Process Diagram – Australian Newspapers Digitisation Program

