



# Australian Newspapers

## Statement of Work Specification for Scanning of Microfilm

23 May 2008

# CONTENTS

1	INTRODUCTION.....	3
1.1	Background.....	3
1.2	Scope of Work.....	3
1.3	Source Material.....	3
1.4	Title Permission.....	3
1.5	Ownership of Data.....	3
2	DIGITISATION PROCESS.....	4
2.1	Task Orders.....	4
2.2	Transport.....	4
2.3	Preparation and Handling of Microfilm.....	4
2.4	Work Report.....	4
2.5	Tape Dump Reports.....	4
2.6	Quality Assurance.....	5
2.7	Re-work.....	5
2.8	Invoicing.....	5
2.9	Delivery Media.....	5
2.10	Deletion of Data.....	6
2.11	Maintenance of Equipment.....	6
2.12	Software and Hardware Upgrades and Changes.....	6
2.13	Changes to Processes or Outputs.....	6
3	DIGITAL IMAGE OUTPUTS.....	7
3.1	General.....	7
3.2	File Format, Bit Depth and Compression.....	7
3.3	Image Quality and Manipulation.....	7
3.4	Metadata.....	8
3.4.1	Tiff tags.....	8
3.4.2	Metadata in Work reports.....	8
3.5	Filenames and Directories.....	9
3.6	Format of Tapes.....	10
	ATTACHMENT A.....	11
	Assessing Microfilm Quality Prior to Digitisation.....	11

# 1 INTRODUCTION

## 1.1 Background

The National Library of Australia (the Library) requires master or second generation microfilm reels of Australian newspaper titles from 1803-1954 to be digitised as part of the Australian Newspaper Digitisation Program (NDP), as per the below specifications. The resulting digital images will be used in Content Analysis and Optical Character Recognition (OCR) processes, resulting in outputs that enable delivery of digital newspapers online to users in a searchable and browsable format. Any variations to this specification will be agreed to in writing by the microfilming bureau and the Project Officer.

## 1.2 Scope of Work

- Minimum of XX page images per week
- High quality 400 dpi raw greyscale tiff images for each newspaper page
- Converted to high quality manipulated bi-tonal tiff images (for OCR)
- Work report for each microfilm reel

## 1.3 Source Material

The source material is master negative microfilm reels, or second generation copies where available. No digitisation must take place from user or positive copies. Some of the master negatives are permanently stored at the microfilming bureau (who also owns them); others will be sourced and sent to the scanning bureau by the Library. The microfilm will be of a suitable quality to ensure high quality digital images. This will be assessed by the microfilm bureau once the microfilm has been made available and before digitisation takes place (Attachment A). If the microfilm is not of suitable quality the microfilm bureau will contact the Library to obtain advice on how to proceed.

## 1.4 Title Permission

The Library will seek written copyright permission from the owner of the master microfilm before work is requested from the microfilm bureau.

## 1.5 Ownership of Data

The Library retains ownership of all digital images created.

## 2 DIGITISATION PROCESS

### 2.1 Task Orders

The order of titles for digitisation will be defined by the Library. The order of work, including title, ISSN and date period will be conveyed to the microfilm bureau through use of the shared wiki. The microfilm bureau must consider each new title requested and discuss the viability of the title with the Library before commencing work. Information on the number of reels within a title, and the quality of the microfilm must be provided to the Library before work commences.

### 2.2 Transport

Many of the microfilm masters are permanently stored at the microfilm bureau premises. If they are not, the Library will arrange and pay for packaging and delivery of items to and from the microfilm bureau.

The microfilm bureau will arrange and pay for the packaging and delivery by courier of LTO2 data tapes containing the digital images to the Library. These costs will be charged back to the Library by invoice. The microfilm bureau must have adequate insurance to cover master microfilm reels they do not own whilst at their premises for digitisation.

### 2.3 Preparation and Handling of Microfilm

The microfilm will be handled appropriately as outlined below:

- wear clean white cotton gloves or equivalent when handling film to avoid scratching
- only touch the edges of the film
- use winders to wind or re-wind the film
- wind or re-wind film gently and slowly

In addition microfilm should be in good condition before scanning. To assist removal of dust the rollers on the scanner will be regularly cleaned as the microfilm is digitised.

### 2.4 Work Report

The microfilm bureau must return to the Library an electronic Work Report for each completed microfilm reel. The Library will supply an xls template for this purpose. The Work Report will provide information about the reel; sequence numbers used and LTO2 tape number. The microfilm bureau may also identify in the Work Report any anomalies or irregularities regarding the content of the microfilm reel. The Work Report will be copied to the LTO2 tape with the files.

### 2.5 Tape Dump Reports

When the digital images are copied to tape a 'tape dump report' will be generated. This report will be automatically e-mailed to the Library to flag that a tape has been produced and so that the Library can cross-check the content of the tape when loaded to the Library's server.

## 2.6 Quality Assurance

The microfilm bureau will quality assure work before it is sent to the Library. It is a requirement that:

- 100% of files are named accurately (section 3.5)
- 100% of files meet file format (section 3.2)
- 100% of files meet metadata specification (section 3.4)
- 100% of images meet imaging specification (section 3.3)
- 100% of the delivery media is readable.

The Library will perform quality checks on all scanned material as follows:

- confirmation of correct filenames (automatic check when data is loaded)
- confirmation of correct metadata (by tiff tag check program)
- confirmation of image pairs and size (by match program)
- images in original sequence, and meet image specifications, including frame split, cropping, deskew, rotate, despeckle (by QA tool)
- .xls work report is present.

If the Library identifies images which do not meet the work specification the microfilm bureau must re-work the individual images or the entire reel as appropriate. Acceptance of work by the Library will be a pre-requisite for payment.

## 2.7 Re-work

Any work that does not meet the Library's work specification must be re-worked. The microfilm bureau must deliver to the Library the re-worked images/reels with a Work Report.

Where an entire reel is re-worked after it has been loaded to the Library's server, the re-worked files will be named with the next sequence number for that ISSN.

If a LTO2 data tape has errors on it and cannot be successfully loaded to the Library's server another tape must be generated by the microfilm bureau. In this case the file names would remain the same.

## 2.8 Invoicing

The microfilm bureau will provide weekly invoices to the Library's Project Officer. The invoice must identify the following:

- number and name of reels,
- tape number supplied,
- number of images, and
- number of Work Reports.

Invoices will be paid only upon acceptance of work.

## 2.9 Delivery Media

LTO2 tapes with a barcode will be supplied by the Library to the microfilm bureau for the copying and transfer of the digital images. The tapes will be couriered back to the Library with a hard copy of the invoice using the provided secure transport containers.

Each tape must be created with 'write protect' status set. Contents of a single microfilm reel must not be split across tapes; the complete contents of a reel must be on a single tape.

## **2.10 Deletion of Data**

Under normal circumstances the Library will load data tapes to the server within 1 week of receipt from the microfilm bureau and after this point the data can be deleted from the microfilm bureau server.

No longer than 3 months after delivery of the digital files the microfilm bureau will destroy all backup and duplicate copies of the digital files.

## **2.11 Maintenance of Equipment**

The microfilm bureau must keep relevant equipment well maintained and serviced, and be able to show a record of this. Maintenance of equipment may affect image output or quality so significant equipment maintenance activities should be recorded in relevant Work Reports.

## **2.12 Software and Hardware Upgrades and Changes**

The microfilm bureau must report in writing to the Library any software/hardware upgrades or changes, as this may effect the work specification (e.g. upgrades to scanning software).

## **2.13 Changes to Processes or Outputs**

The microfilm bureau must report in writing to the Library as soon as they are aware of any changes to digitisation processes or expected weekly outputs.

## 3 DIGITAL IMAGE OUTPUTS

### 3.1 General

- Microfilm will be assessed by the microfilm bureau before digitisation takes place to ensure that it is suitable for digitisation (see section 1.3).
- All frames on the microfilm reel are to be captured, saved and supplied to the Library, including blank pages and missing issue/page targets, with the exception of microfilm start and end reel targets.
- One digital image should represent one newspaper page. (In many cases the microfilm will have 2 pages per frame which will need to be split into 2 digital images).
- A raw greyscale tiff image is to be supplied for each page.
- A bi-tonal tiff image optimised for OCR through the use of software filters, is also to be supplied for each page.
- Each newspaper page has a 'pair' of images (a greyscale and a bi-tonal).
- The 'pair' should match exactly in size, co-ordinates and cropping.
- Some image manipulation is to take place such as cropping, de-skew, rotation to reading view, and enhancement of bi-tonal files.
- Image manipulation is to be automated wherever possible within the scanning software e.g. auto-frame split, crop, rotate, de-skew, de-speckle.
- No image manipulation is to take place manually or using image manipulation programs other than the agreed scanning software without the prior agreement of the Library.
- Any discrepancies in page numbers, missing pages, or queries/comments about the reel can be noted in the Work Report.
- A minimum of XX page images should be processed per week.

### 3.2 File Format, Bit Depth and Compression

- All images scanned to greyscale 8 bit, lossless compressed Lempel Ziv Welch (LZW) tiff files.
- A copy of this file made, manipulated and saved as Bi-tonal 1 bit, Group 4 compression tiff file.

### 3.3 Image Quality and Manipulation

- Scan page images to 400 dpi greyscale.
- Frame splitter to be used so that one newspaper page is one image.
- Image auto de-skewed to within  $\pm 1$  degree from parallel.
- Images cropped to the edge of the page with a 1-2 mm border. (If pages have been filmed on top of other newspaper pages crop using the scanning software to the edge of the smallest page).
- All images should be of consistent size and dimension or as close to this as possible.
- Image saved as raw greyscale tiff.
- A copy image has filters applied to improve text for OCR (e.g. de-speckle, smooth edges) and saved as bi-tonal tiff (but no further cropping must take place on this image). It is essential that file co-ordinates e.g. edge and centre of page match on page file pairs.
- Images should have the same polarity as the original material (usually, dark text on light background).
- Variations in film densities should be corrected where possible.

### 3.4 Metadata

#### 3.4.1 Tiff tags

The following tiff tags **must** be included in the tiff file. The operator should ensure that none of the tags are overwritten or deleted. Most if not all of these tags may be automatically generated by the scanning software. The tags will be checked by the Library as part of the quality assurance process.

Tag	Name	Example Data
256	Image width (in pixels)	5184
257	Image length (in pixels)	7016
258	Bits per sample	1 for bitonal images and 8 for Greyscale images
259	Compression	Group 4 Fax
262	Colour space	0=bitmap, 1=greyscale, 2=RGB
271	Make of capture device	nextScan,Inc
272	Model of capture device	Eclipse, SN# 415003
273	StripOffsets	460
274	Orientation	Top/left
278	Rows per strip	7016
279	StripByte Counts	903728
282	X Resolution	400
283	Y Resolution	400
296	Resolution Unit	inch
305	Capture Software	nextStar v1.01
306	DateTime	2007-07-02 17:01:56 (YYYY:MM:DD HH:MM:SS)
315	Artist	Pascoe

Additional tags may also be supplied.

#### 3.4.2 Metadata in Work reports

Microfilm reel details	Example Data
Title	The Argus (Supplied by NLA)
Reel name	36
Dates of coverage	
From	DD-MM-YYYY (where possible)
To	DD-MM-YYYY (where possible)
Filmed by	(Microfilming Bureau)
Filmed for	(Microfilm requester)
Date microfilmed	DD-MM-YYYY where possible
Reduction Ratio	18x
ISSN	18339719 (Supplied by NLA)
Scanning details	
Scanned by	(Scanning Bureau)
Date of scanning	DD-MM-YYYY
Checked by	(First name Surname)
Assigned File Number sequence Range	63393-64578

<b>Comments/notes</b>	(Optional) scanner maintenance, condition of microfilm, changes to software etc.
<b>Delivery Details</b>	
Tape number	PA0052 ( barcode supplied on tape)
Date sent to NLA	DD-MM-YYYY

### 3.5 Filenames and Directories

The microfilm bureau is to assign filenames for digital images and deliver these in an arrangement of directories as outlined below.

#### General:

- All file names will be unique.
- File 'pairs' will have identical names except for the file type g or b (g for greyscale files and b for bi-tonal files) (A 'pair' is an image of the same page saved as greyscale and bi-tonal).
- The sequence numbering should be continued from the end of one reel to the start of the next reel for each individual title, so that all file names are unique for a title.
- Files will be named sequentially in the order they appear on the microfilm.
- Targets are to be treated as newspaper pages and will be named accordingly, except the microfilm start and end targets which are to be ignored. See section 3.1
- The Library will supply the correct ISSN for each title to the microfilm bureau.

#### NLA NDP File naming convention

nla.news-issnxxxxxxxx-s#-b and  
nla.news-issnxxxxxxxx-s#-g

where

- **x** stands for International Standard Serial Number (8 digits long - but note that the final character, which is a check digit, may sometimes be an "x" rather than a number. Any hyphens are ignored);
- **s** stands for sequence;
- **#** is a running number with no leading zeroes;
- **b** stands for bi-tonal; or **g** stands for greyscale;
- all letters are lowercase.

Each file is to have a name consisting of the base "**nla.news-issn**", followed by the ISSN for the publication (8 numeric characters, sometimes "x" as the final character, with no hyphen), followed by a unique sequence number for that page starting with "**-s**", then by "**-g**" for the greyscale image or "**-b**" for the bi-tonal image, and ending with an extension for the file type i.e. ".tif".

For example, if the ISSN for a title is 1234-567X, the base file name for this publication is

- nla.news-issn1234567x-s# (where # is the sequence number, there is no hyphen in the ISSN number in the filename and the "x" in the ISSN is converted to lower case)

The filename of the first image for this publication would therefore be

- nla.news-issn1234567x-s1-g.tif (for the greyscale image)
- nla.news-issn1234567x-s1-b.tif (for the bitonal image)

The second image would be

- nla.news-issn1234567x-s2-g.tif (for the greyscale image)
- nla.news-issn1234567x-s2-b.tif (for the bitonal image)

All letters are lowercase and there are no leading zeroes, i.e. do not include “0” in the filename sequences. For example,

- nla.news-issn1234567x-s1
- **NOT** nla.news-issn1234567x-s01 and **NOT** nla.news-issn1234567X-s1

### **Directory Structure:**

The files will be organized into folders by title according to their originating microfilm reel number, and each folder will contain:

- a sub folder for the set of bitonal images,
- a sub folder for the set of greyscale images, and
- an EXCEL spreadsheet containing the work report.

For example:

The Argus

Reel 32

Bitonal  
Greyscale  
Work report

Reel 33

Bitonal  
Greyscale  
Work report

## **3.6 Format of Tapes**

LTO2 Tapes should be written using GNU TAR format. The TAR should start at the beginning of the tape (Fileset 0). The TAR should be written using a blocksize of 32768 bytes ( 32kb). There should be no more than 1 TAR fileset written to each cartridge.

## ATTACHMENT A

### Assessing Microfilm Quality Prior to Digitisation

Microfilmed newspapers must be of the standard and quality outlined below in order to be considered acceptable for digitisation. It is the responsibility of the microfilm bureau to assess the quality of microfilm prior to digitisation when they have the masters available for scanning, and to discuss quality with the Library prior to digitisation if they have concerns.

#### Content/condition of original hard copy papers filmed:

- Complete run
- Preferably unbound, but if bound no serious obscurement of text in gutters
- Clean copies (no blotching, dirt, fading, text is readable)
- Even inking throughout
- No excessive bleedthrough
- No excessive damage, tears, holes
- No folds or other obscurement of content.

#### Microfilm quality:

- The microfilm is as complete as possible - more than 99% of available pages have been included
- There is no cropping of pages or text in more than 98% of the microfilm frames
- The original page reduction ratio should not be more than 20x
- The microfilm Dmax should be between 0.9 to 1.3, if possible<sup>1</sup>
- No excessive skew (where excessive is more than  $\pm 4$  degrees from parallel) in more than 98% of frames on the microfilm
- The smallest text on 100% of the pages should be legible
- There is minimal inclusion of foreign objects e.g. clips
- Any joins in the microfilm are not be across actual content
- The text is in focus.

#### Microfilm Condition:

- Clean (no excessive dirt, mould or dust)
- 98% of frames on the microfilm have no scratches/marks or other extensive damage
- Age (at present any age microfilm is being scanned but this may be reviewed).

---

<sup>1</sup> A measure of the darkest shadow which can be scanned by a scanner.