



# Australian Newspapers

Statement of Work Specification for  
Content Analysis and Optical Character  
Recognition (OCR)

23 May 2008

# CONTENTS

|       |  |    |
|-------|--|----|
| 1     | INTRODUCTION .....   | 3  |
| 1.1   | Background.....  | 3  |
| 1.2   | Scope of Work.....   | 3  |
| 1.3   | Copyright Clearance .....                                      | 3  |
| 1.4   | Source Material .....  | 3  |
| 1.5   | Deliverables.....  | 3  |
| 1.6   | Ownership of Data.....   | 3  |
| 2     | WORKFLOW PROCESS .....   | 4  |
| 2.1   | Batch definition.....  | 4  |
| 2.2   | Work Schedule.....   | 4  |
| 2.3   | Monthly Tracking Report .....                                  | 4  |
| 2.4   | Transport of Source.....                                       | 4  |
| 2.5   | Processing .....   | 4  |
| 2.6   | Delivery Notification.....                                     | 5  |
| 2.7   | Delivery of Deliverables .....                                 | 5  |
| 2.8   | Quality Assurance by Contractor .....                          | 5  |
| 2.9   | Quality Acceptance by the Library .....                        | 5  |
| 2.10  | Acceptance/Rejection Notification .....                        | 7  |
| 2.11  | Rejection of Batches, Rejection Report and Re-processing ..... | 7  |
| 2.12  | Invoicing .....  | 7  |
| 2.13  | Deletion of Data .....   | 7  |
| 2.14  | Data Security and Management.....                              | 7  |
| 2.15  | Software Upgrades and Changes.....                             | 7  |
| 2.16  | Changes to Workflows and Processes .....                       | 7  |
| 2.17  | Changes to Work Specification .....                            | 8  |
| 2.18  | Project Management.....  | 8  |
| 2.19  | Communication .....  | 8  |
| 2.20  | Time Zones.....  | 8  |
| 2.21  | Processing of Missing and Duplicate Pages.....                 | 8  |
| 3     | CONTENT ANALYSIS AND OCR SPECIFICATIONS .....                  | 9  |
| 3.1   | Library Supplied Source .....                                  | 9  |
| 3.2   | Contractor Supplied Deliverables .....                         | 11 |
| 3.3   | Batch Naming Convention .....                                  | 12 |
| 3.4   | Directory Structure.....                                       | 13 |
| 3.5   | File Naming.....   | 13 |
| 3.5.3 | Batch Check File.....  | 14 |
| 3.6   | Content Analysis - Zoning .....                                | 14 |
| 3.7   | Content Analysis - Categorisation.....                         | 15 |
| 3.8   | Re-keying Specifications .....                                 | 16 |
| 3.9   | Treatment of Graphics/Illustrations.....                       | 18 |
| 3.10  | Content Analysis and OCR Generalities.....                     | 18 |
| 3.11  | Colophon – Definition.....                                     | 19 |
| 3.12  | Masthead – Definition .....                                    | 20 |
| 4     | XML METS/MODS FILE SPECIFICATION .....                         | 21 |

|     |   |    |
|-----|---|----|
| 5   | XML ALTO SPECIFICATION .....                  | 44 |
| 6   | EXAMPLE FILES.....                            | 48 |
| 6.1 | Sample XML METS/MODS file at Issue Level..... | 48 |
| 6.2 | Sample XML ALTO (OCR) file at Page Level..... | 48 |

## 1.1 Background

As part of the Australian Newspaper Digitisation Program (ANDP) the National Library of Australia (the Library) requires the Content Analysis and Optical Character Recognition (OCR) Contractor (the Contractor) to process TIFF digital page images of selected Australian newspapers from 1803-1954 into XML files using Content Analysis and OCR techniques, and as per the below specifications. The resulting XML files will be used for delivery to end users via a web interface.

Any variations to this specification will be agreed to in writing by the Contractor and the Library.

## 1.2 Scope of Work

- Minimum/maximum of XX page images per week
- Different newspaper titles
- Different time periods

## 1.3 Copyright Clearance

The Library will ensure that source material is copyright cleared or out of copyright before sending to the Contractor.

## 1.4 Source Material

The Library will create and supply the following source material to the Contractor on LT02 tapes:

- Bi-tonal 400 dpi TIFF images or grayscale LZW compressed page TIFF images
- Page Level metadata (e.g. title, date, page number) in a CSV file.

## 1.5 Deliverables

The Contractor will:

- Create Issue Level metadata (volume and issue);
- Identify article zones,
- Apply a category to every article;
- Provide OCR text for every article; and
- Rekey title, subtitle, abstract and author text.

The Contractor will return to the Library:

- XML METS/MODS file for each newspaper issue
- XML ALTO file for each newspaper page
- Batch manifest file for each batch of pages

Deliverables will be provided to the Library via ftp.

## 1.6 Ownership of Data

The Library retains ownership of all XML files created.

### 2.1 Batch definition

Pages for processing will be grouped into batches defined by the Library. The Contractor will use Library defined batches and batch names. Batch sizes may vary and will be determined by the Library but will usually be around 2500 pages per batch. Batches will be processed in the order received unless the Library requests otherwise.

### 2.2 Work Schedule

The order of processing batches will be defined by the Library. The number of pages to be processed per week will be agreed by the Contractor and Library. Timeframes for delivery, re-processing, payment, and penalties for non or late delivery will be defined in the Contract.

The Contractor will provide the Library with access to a weekly work schedule detailing batches received by Contractor, number of pages and expected processing completion dates.

### 2.3 Monthly Tracking Report

For each batch in process the Contractor will provide the Library with a monthly tracking report detailing:

- Batch number
- Image type
- Title and Year of newspaper
- Total Number of Pages in batch
- Total Number of Articles in batch
- Date batch received at Contractor
- Contractor processing status
- Date processing complete notification sent to Library
- Library QA status
- Number of delivery rounds
- Library rejection/acceptance date
- Specification/software version used on batch
- % of articles categorised as News/Advertising/Family Notices/Detailed Lists

The tracking report will be in Word format within a table and e-mailed from the Contractor Project Manager to the Library Project Manager within 3 days of the first of the month.

### 2.4 Transport of Source

The Library will send LT02 tapes containing the original source by tracked Courier to the Contractor. The tapes are not required to be returned to the Library.

### 2.5 Processing

The Contractor will process the source in accordance with the specification and workflows outlined in this document.

## **2.6 Delivery Notification**

The Contractor will send an automated notification to the Library to advise that batches have been processed/re-processed and are available for download from the Contractors ftp site.

## **2.7 Delivery of Deliverables**

XML deliverable files will be transferred to the Library by ftp.

## **2.8 Quality Assurance by Contractor**

The Contractor will perform quality assurance checks on the deliverables before they are supplied to the Library, in particular the batch level checks that require 100% accuracy.

## **2.9 Quality Acceptance by the Library**

The Library will perform both automated and manual Quality Acceptance checks on the deliverables through the Library's Content Management System.

Automated checks on criteria that have an acceptance level of 100% will be performed on all deliverables. Manual checks for criteria that have an acceptance level less than 100% will be performed using random sampling in line with ISO 2859 Procedures for Inspection by Attributes (Part 0-4).

e.g.

On an average batch size of 2500 pages:

Batch Level acceptance criteria – all batches will be automatically checked.

Article Level acceptance criteria - 315 articles of a total of 25,000 articles will be manually checked = 1 %

Issue Level acceptance criteria - 50 issues of total 310 issues will be manually checked = 16 %

**Batch-level Acceptance Criteria (automated checks)**

| <b>Criteria</b>                  | <b>Description</b>  | <b>Acceptance Level</b> |
|----------------------------------|---|-------------------------|
| Inventory reconciliation         | Number of source pages equals the number of pages delivered   | 100%                    |
| Conformity to XML Schema         | XML text file conformity to XML schema as defined   | 100%                    |
| File and directory name accuracy | File names and directory names are accurate   | 100%                    |
| Batch-level metadata accuracy    | Metadata accuracy for Newspaper title, ISSN, Edition/Supplement/Section information retained as per source metadata records | 100%                    |

**Issue-level Acceptance Criteria (manual checks within batches)**

| <b>Criteria</b>                  | <b>Description</b>                                    | <b>Acceptance Level</b> |
|----------------------------------|---|-------------------------|
| Volume and issue number accuracy | Volume number and issue number captured as per source | 99%                     |

**Article-level Acceptance Criteria (manual checks within batches)**

| <b>Criteria</b>                     | <b>Description</b>  | <b>Acceptance Level</b> |
|-------------------------------------|---|-------------------------|
| Page image skew accuracy            | The page on which the article appears is within skew specifications   | 95%                     |
| Zoning accuracy                     | Zoned article boundaries include all content belonging to an article; no part of an article is cut off by a zone boundary | 99%                     |
| Illustration image linking accuracy | Illustrations within an article are linked according to specifications.   | 97%                     |
| Split article linking accuracy      | Articles which are split across a single page or continued across several pages are linked accurately                     | 99%                     |
| Article Category Accuracy           | Article Category accurately reflects the subject matter of the article  | 92%                     |
| Metadata accuracy                   | Title, subtitle, author names and abstract metadata elements are populated with the correct data                          | 99%                     |
| Re-keyed field accuracy             | Title, subtitle, author names and abstract fields are accurate to source  | 99.50%                  |

If the Library identifies batches which do not meet the work specification or required acceptance level the Contractor must re-process the batches as appropriate. Acceptance of work by the Library will be a pre-requisite for payment.

## **2.10 Acceptance/Rejection Notification**

The Library will send an automated e-mail to the Contractor to notify the Contractor that batches have either been accepted or rejected.

## **2.11 Rejection of Batches, Rejection Report and Re-processing**

If batches are rejected the automated e-mail will also contain a report outlining the reason (s) for rejection. Any batches that do not meet the Library's work specification must be re-processed and re-delivered to the Library within 7 working days of notification.

## **2.12 Invoicing**

The Contractor will provide monthly invoices to the Library. For all batches processed and accepted within the month the invoice must detail the following:

- Batch Number
- Total Number of Pages in the batch
- Total Number of articles in the batch
- Page image format (bitonal or grayscale)

Invoices will be paid only upon acceptance of work.

## **2.13 Deletion of Data**

The Contractor will retain Library source data and accepted delivered data on its server for 12 months. After 12 months the Contractor will destroy all backup and duplicate copies of the digital files.

## **2.14 Data Security and Management**

The Contractor must have current certification for ISO 9001:2000 Quality management systems – Requirements, and ISO 27001 Information Security Management. The Contractor must have in place and provide to the Library relevant documentation outlining the disaster recovery and backup plan for power outages.

## **2.15 Software Upgrades and Changes**

The Contractor must provide to the Library in writing at the commencement of the Contract details of software, including the version number, being used to process and report on the Library's data. The Contractor must report in writing to the Library any intended software upgrades or significant software changes prior to implementation. If software is upgraded or changed it must be implemented on new batches, not implemented on batches that are part way through processing.

## **2.16 Changes to Workflows and Processes**

The Contractor must report in writing to the Library if at any time agreed workflows cannot be met. The Contractor is invited to make suggestions for workflow and/or process improvements to the Library in writing at any time.

## **2.17 Changes to Work Specification**

If the Library or Contractor require changes to any work specification this must be done in accordance with the Contract. It is likely that changes to the work specification may also require changes to software, workflows and processes. Adequate notice must be given by either Party. The Contractor must not change the specification without the Library's written agreement. If the specification is changed it must be implemented on new batches, not implemented on batches that are part way through processing.

## **2.18 Project Management**

The Contractor must appoint a suitable and experienced project manager to manage the work requested by the Library. All day to day communication from the Contractor to the Library on operational issues will be through the Project Manager. If the Contractor is required to change the Project Manager written agreement is required from the Library. The Library will appoint a suitable and experienced project manager to manage the work requested by the Library. All day to day communication from the Library to the Contractor on operational issues will be through the Project Manager.

## **2.19 Communication**

Communication between the Library and Contractor will be via e-mail, reports, and letter. As and when necessary or requested by either Party teleconferences will be held during Australian business hours.

## **2.20 Time Zones**

All timeframes for supply and delivery of data will be in Australian Eastern Standard Time or Australian Eastern Summer Time (depending on the time of the year). Teleconferences between the Contractor and the Library will take place during Australian business hours which are generally between the hours of 9:00am and 5:00pm.

## **2.21 Processing of Missing and Duplicate Pages**

The Library may require 'found' missing pages and duplicate pages to be processed. This will happen within normal workflows. These pages will not duplicate file names previously supplied to the Contractor.

**3.1 Library Supplied Source**

**3.1.1 Batches**

Source material will be provided in batches. Each batch will contain a single newspaper title with multiple issues. Issues will never be split across batches. Batches will be processed in the order received which will generally be in sequential batch number order. When multiple batches are received on the same day they will be processed in batch number order.

All page images in a batch will be uniformly grayscale or bitonal. A batch will never contain a combination of bitonal and grayscale images.

Batches will be assigned a batch number by the Library and all source page images and source metadata files will be contained within a batch-level directory named as per the batch number.

|   |  |
|---|--|
| batch#/<br>batch#*.tif<br>batch#/pagelist.csv | batch-level directory, batch number is <u>not</u> zero-padded<br>source page images<br>source metadata records |
|---|--|

**3.1.2 Digital Page Images**

- Page images will be formatted as 400 dpi bitonal CCIT G4 or grayscale LZW TIFF files
- Each image will contain one physical page, i.e. no 2-up images
- The entire physical page will be represented in the image file, without cropping of any detail.
- Text on the page will be visible in the image.
- In rare circumstances the Library may send page fragments for processing, but generally pages will be sent in their entirety.
- Generally page images will be free of excessive skew ( ±1 degree of skew)
- Page images will be presented in natural reading orientation
- A single page will be provided as either bitonal or grayscale image
- The Library will have assigned metadata to each image at page level and this will be provided in the relevant CSV file (see next section).
- The Library will have assigned flags to duplicate, blank and missing pages and these pages will be ignored and not processed by the Contractor (see next section).
- The Library will have assigned virtual page sequence numbers to pages and these will be included in the file name and in the metadata file (see next section).
- Each page will have a unique file name as outlined below:

**nlImageSeq-#-t.tif**

e.g. **nlImageSeq-900000-b.tif**

where:

|             |   |
|-------------|---|
| nlImageSeq: | literal image filename prefix   |
| #:          | unique image number (starts at 1, is not zero padded, incrementing number, sequential within project) |

|    |   |
|----|---|
| t. | image format indicator, bitonal (b) and grayscale (g) |
|----|---|

### 3.1.3 Source Metadata Records and CSV file

The Library will provide descriptive metadata records with each batch. The format of the metadata records will be comma-separated ASCII (CSV) file named as “**pagelist.csv**”. Each record in the file will pertain to a single source page image. The metadata will not be altered by the Contractor. The metadata will be used in the METS/MODS issue file. The sequence of pages and supplements will not be changed by the Contractor or in the deliverables.

Metadata records will detail the following:

| Field Name                 | Contents (Mandatory Fields)  |
|----------------------------|--|
| Newspaper title            | The title of the newspaper.  |
| ISSN of newspaper          | The International Standard Serial Number of the newspaper.   |
| Issue publication date     | The date of publication of the issue in “yyyymmdd” format.   |
| Page sequence number       | The virtual page sequence number (assigned by the Library). The page sequence number is not unique in an issue, but is unique when combined with edition, supplement and section numbers.                                |
| Page image filename        | The filename of the source page image.   |
| Field Name                 | Contents (Additional fields if relevant)   |
| Edition sequence number    | A number indicating the sequential order of the edition.   |
| Edition name               | The name or title of the edition.  |
| Supplement sequence number | A number indicating the sequential order of the supplement   |
| Supplement name            | The name or title of the supplement  |
| Supplement date            | The publication date of the supplement in “yyyymmdd” format. Only supplements will have publication date information. If a supplement date is not provided the issue publication date will be used as a supplement date. |
| Section sequence number    | A number indicating the sequential order of the section  |
| Section name               | The name or title of the section   |

|             |  |
|-------------|--|
| Target flag | A Boolean indicator. If set to “y”, then the page image is a placeholder image which does not require any processing (for blank, duplicate or target pages). |
| Notes       | Notes regarding the page image. (Generally to be ignored by the OCR Contractor).   |

Example of pagelist.csv file:

The Canberra Times,01576925,19290913,1,0,,0,,0,,,nlalImageSeq-24537-b.tif,

The Canberra Times,01576925,19290913,2,0,,0,,0,,,nlalImageSeq-24538-b.tif,

The Canberra Times,01576925,19290913,3,0,,0,,0,,,nlalImageSeq-24539-b.tif,

The Canberra Times,01576925,19290913,4,0,,0,,0,,,nlalImageSeq-24540-b.tif,

The Canberra Times,01576925,19290913,5,0,,0,,0,,,nlalImageSeq-24541-b.tif,

### 3.2 Contractor Supplied Deliverables

Four deliverable types are required as follows:

- XML METS/MODS file for each issue
- XML ALTO (OCR) file for each page
- Batch check file
- Batch manifest file

#### 3.2.1 Creation of Metadata at Issue Level

The Contractor will create the following metadata at issue level, which will supplement the Library supplied metadata at page level and will be included in the METS/MODS file for each issue:

| Field Name: | Contents (Mandatory if present on source)   |
|-------------|---|
| Volume      | A number indicating the volume of the newspaper title, transcribed exactly as it appears (Roman or Arabic). |
| Issue       | A number indicating the issue of the newspaper title, transcribed exactly as it appears (Roman or Arabic).  |

The volume and issue number appear directly below the masthead on the front page of an issue. If only a single number is present this should be used as the issue number. If any of the characters are illegible a [?] should be used to replace them.

#### 3.2.2 XML METS/MODS file at Issue Level

The content will be as below. For the detailed specification see section 4.

Descriptive Metadata:

- Newspaper title (Library provided. See section 3.1.3)
- Issue publication date (Library provided. See section 3.1.3)
- Edition, supplement and section information (Library provided. See section 3.1.3)
- Volume and issue number (Contractor provided. See section 3.2.1)
- Article type (category)
- Article title (re-keyed text)
- Article subtitle (re-keyed text)
- Article authors (re-keyed text)
- Article abstract (re-keyed text)

Administrative Metadata:

- Logical structural map
- Physical structural map i.e. where the articles occur on the pages , the order of pages in an issue, and which files represent which pages
- Manifest of deliverable files
- Image coordinates and de-skew angle

### 3.2.3 XML ALTO (OCR) file at page level

The content will be as below. For the detailed specification see section 5.

- Page Layout (zones, paragraphs, lines, words and fonts recognised by the OCR process and their coordinates on the page)
- Page Content (raw OCR)

### 3.2.4 Batch Check File

An index file for each batch listing the name and MD5 checksum and byte count (in kilobytes) of each deliverable file. Each file delivered for a batch will be listed in this file; the Batch Check File will have a header listing the contents of the batch.

Each newspaper title delivered in the batch will be listed in the header. Each file delivered in the batch will be listed with full path and filename and corresponding checksum. For example:

```
<batch="0012Rn"><issn="01576925"><issuedate="19290913">  
<batch="0012Rn"><issn="18340946"><issuedate="18201118">  
865cba9b34658f765383912e8fb82f1d /nla.news-  
issn18340946/18201118/pages/nlalmageSeq-4092-b.tif  
d92072ca39374505f73c13186de2e9f8 /nla.news-  
issn01576925/19290913/pages/nlalmageSeq-24537-b.tif
```

### 3.2.5 Batch Manifest File

A simple XML file detailing:

- The metadata provided in the source metadata file for each batch
- The page image details for each source page image for each batch.

## 3.3 Batch Naming Convention

Pages are defined into batches by the Library. The Library will generally provide batches of around 2500 pages. The batch naming convention will be:

### 3079-bbbbRn

where:

3079-: literal job number prefix  
bbbb: 4-digit zero-padded batch number.  
R: delivery round prefix  
n: delivery round number (where first delivery = 1)

Example:

3079-0001R1 would be the first delivery of batch 0001

3079-0001R2 would be the second delivery of batch 0001 (i.e. if the batch is rejected and re-processing is required).

## 3.4 Directory Structure

Deliverables will be provided in the following directory structure:

3079-bbbbRn/ batch-level directory  
3079-bbbbRn/3079-bbbbRn.chk batch check file  
3079-bbbbRn/3079-bbbbRn.xml batch manifest file  
3079-bbbbRn/nla.news-issnxxxxxxx/ title-level directory  
3079-bbbbRn/nla.news-issnxxxxxxx/yyyymdd/ issue-level directory  
3079-bbbbRn/nla.news-issnxxxxxxx/yyyymdd/\*.xml issue XML file  
3079-bbbbRn/nla.news-issnxxxxxxx/yyyymdd/pages/ page level directory  
3079-bbbbRn/nla.news-issnxxxxxxx/yyyymdd/pages/\*.xml page OCR files

## 3.5 File Naming

The page images are provided to the Contractor using the naming convention outlined in section 3.1.2. Page images do not need to be renamed by the Contractor or returned to the Library. The Contractor will follow the naming convention below for the XML files created.

### 3.5.1 XML ALTO (OCR) file for each page

Use the page file name but change the extension from .tif to .xml

**nlalimageSeq-#-t.xml**

e.g. **nlalimageSeq-900000-b.xml**

where:

|               |   |
|---------------|---|
| nlalimageSeq: | literal image filename prefix   |
| #:            | unique image number (starts at 1, is not zero padded, incrementing number, sequential within project) |
| t:            | image format indicator, bitonal (b) and grayscale (g)   |

### 3.5.2 XML METS/MODS file for each issue

Using the information from the Library supplied CSV file name as follows:

**nla.news-issnxxxxxxxx\_yyyymmdd.xml**

e.g. **nla.news-issn12345678\_18030418.xml**

where:

|               |   |
|---------------|---|
| nla.news-issn | Literal prefix  |
| xxxxxxxx      | newspaper ISSN (8 digits long, but note that the final character which is a check digit may sometimes be an x rather than a number) |
| yyymmdd:      | issue publication date  |

### 3.5.3 Batch Check File

Use the batch name (as outlined in section 3.3) and use the file extension .chk

3079-bbbbRn.chk

e.g. 3079-0071R2.chk

### 3.5.4 Batch Manifest File

Use the batch name (as outlined in section 3.3) and use the file extension .xml

3079-bbbbRn.xml

e.g. 3079-0071R2.xml

## 3.6 Content Analysis - Zoning

Content on each page is zoned into individual articles.

Article zones are determined by article titles. The presence of an explicit high-level title is the start of an article. Generally individual articles are zoned separately. However two article categories may be zoned as a single article, as outlined below:

- Advertising articles on the same page are always zoned as a single article, regardless of article length.
- Family Notices on the same page are always zoned as a single article, regardless of article length.

In addition:

- The Colophon is always zoned together with the preceding article (see definition of Colophon in section 3.11)
- Page fragments may be zoned with the previous articles on the same page if no title is present, or the last article on the previous page if only a fragment exists for the entire page.

Articles with visual cues indicating continuation (i.e. “Continued on Page #”, “Continued on Next Page”, “Continued from Page #”, “Continued from Previous Page”) are linked

across pages, regardless of type or length. 'Advertising' and 'Family Notices' will not be linked across pages unless there is an explicit visual cue for continuation.

The great majority of articles will appear in standard sequence, that is, the article will begin at the start of an issue and move sequentially towards the end of an issue. However, in some cases, an article may have reverse page sequence. In other words, the article starts at end of an issue and moves "backwards" toward the start of an issue. For example, an article starts on Page 7 and continues on Page 6.

In such cases, the article will be sequenced in "reading order", as in the example given above; the article will start on Page 7 and continue to Page 6. This sequence will be reflected in the article-level elements of the Issue-level XML file.

### **3.7 Content Analysis - Categorisation**

The Library is using article categorisation to support enhanced searching by the user. Categories will enable relevance ranking of result sets and allow users to narrow search results. The Contractor will apply categories and illustration types during the content analysis and OCR process and this metadata will be recorded in the .METS/MODS file. All articles will be assigned one of the following 4 categories:

- News
- Family Notices
- Advertising
- Detailed Lists, Results and Guides

Illustrations appearing in News articles will always have an illustration type applied:

- Illustration (default)
- Photo
- Cartoon
- Map
- Graph

Standalone illustrations are categorised as News. For further specifications on illustrations (see section 3.9)

The full definitions for article categories are:

#### **News**

News articles cover a wide range of subject matter, including current affairs, law courts and crime, official appointments and notices, commerce and business, sport and social news. Obituaries, editorials, letters and correspondence (usually to the editor) and editorial or political cartoons are also categorised as News articles.

Articles with subject matter related to shipping news or intelligence, including arrival and departure information, sailing schedules, and fares and services for ships are categorized as News articles.

Articles with subject matter related to art, literature, music, theatre, comics, shows, gardening, travel, crafts (such as crochet and knitting), stories, fiction and poetry are categorized as News articles.

News is the default category. If an article cannot be clearly identified as any of the specific article types listed below it will be categorized as News.

## Family Notices

Birth, death and marriage notices and related announcements including weddings, anniversaries, in memoriam, bereavement, birthdays, and congratulations will be categorized as Family Notices.

## Advertising

This category contains both display advertising and classified advertising. Display advertising usually contains both text and graphic information such as logos, drawings, or other pictures or photographs. Display advertising is usually large in size spanning multiple columns or single entire columns, with large fonts and is placed outside of the classified advertisements section on whole pages or inserted amongst news items. All advertising in the newspaper masthead although small is display advertising.

Classified advertising generally appears in a specific section of the newspaper and contains text only. Classified advertising may include property notices, items for sale, employment notices, public and personal notices.

## Detailed Lists, Results, Guides

This category contains detailed sporting results, guides, radio and television guides, weather forecasts, election results, education results and courses and stock market lists, crossword puzzles, word games and quizzes.

### 3.8 Re-keying Specifications

In order to improve the quality of data presented in the search and delivery system the Library requires the following parts of articles to be re-keyed.

- Article Title (for News and Detailed Lists and Results categories only)
- Subtitles (for News and Detailed Lists and Results categories only)
- Author information (for News category only)
- First four lines of main article text –abstract (for News category only)

This information must be captured in the order in which it appears in the source and in its entirety.

**The Article Title** is the prominent display heading which marks the start of an article. The Article Title is typically offset from other text through emphasis, larger font size or different font face. An Article can have only one Article Title.

Anecdotes, quotes and/or preamble text appearing before the title are captured as part of the article title.

When articles are grouped under a single group heading, each individual article within the group is treated as a separate article. The article group heading will **not** be captured.

Article titles are captured for News and Detailed Lists, Results and Guides categories only. The article types below will have the article type automatically generated as the article title:

- Advertising
- Family Notices

In some cases an article may not have an identifiable title. For example a standalone photograph or article that starts at the top of a page. In this case the article title will be input as 'No title'.

If an illustration has a caption only it will be given 'no title' (captions are processed as abstracts).

**The Article Subtitle** is all of the text that occurs after the article Title and before the running article text, with the exception of Author information. An article can have multiple Subtitles. Subtitles are captured for News and Detailed Lists, Results and Guides categories only. Subtitles are not captured for Advertising or Family Notices.

**Article authors** are typically indicated with a byline appearing after the article title/subtitle or at the end of the article body. Article author names are captured as they appear on the source.

Author information includes all of the text associated with the author including titles, honorifics and qualifiers. An article may have more than one Author.

Article authors are captured (when available in the source) for News articles only.

**Article Abstract.** If not present on the source, the article abstract is constructed from the first four lines of the article text, immediately following the last piece of identified metadata, e.g. article title, subtitle or authors.

If the fourth line of text ends in a hyphenated word, then the abstract is truncated *before* the hyphenated word.

In standalone illustrations the caption text is to be used as the abstract text.

Article abstracts are captured only for News articles.

### **Hyphens in re-keyed text**

Hyphens occurring at the end of the line are treated as follows:

- If the hyphen is a required hyphen, as in a hyphenated word, the hyphen is retained as per source.
- If the hyphen follows any punctuation symbol, the hyphen is retained as per source.
- If the hyphen is a word wrap hyphen, used to indicate the truncation of a word at the end of a line, the hyphen is removed.

### **Long S in re-keyed text**

Instances of "long s" are converted to the modern "s". The "long s" character was formerly used where 's' occurred in the middle or at the beginning of a word, for example *l̄infulnefs* ("sinfulness"). In many instances, the "long s" appears similar to the letter "f".

### **Small caps in re-keyed text**

Small Caps are captured as ALL CAPS. For example, "THIS IS AN ARTICLE TITLE" is captured as "THIS IS AN ARTICLE TITLE"

## **Illegible Text in re-keyed text**

Illegible characters will be replaced with the placeholder value, [?]. The [?] placeholder will be used to replace illegible single-characters or adjacent characters within the same string.

## **3.9 Treatment of Graphics/Illustrations**

### **Article Illustrations**

Illustrations can fall within the category 'News', 'Family Notices' or 'Detailed Lists, Results and Guides', but not Advertising.

An illustration is defined as a graphical image that does not prominently contain a company logo or advertise a product or service. Illustrations advertising products and services with logos are considered Advertising.

An illustration should be identified as a "photo", "cartoon", "map", "graph", or "illustration" as the default if it is not clearly one of the other types."

Illustrations within articles categorised as 'Advertising' are never zoned as a separate illustration or given an illustration type.

Small decorative stamps e.g. pointing fingers, coats of arms or other graphics which are for decorative purposes only will not be treated as an illustration and will be contained within the main article text zone.

Illustrations and associated captions will be captured as a single zone as part of the article image.

### **Standalone Illustrations**

An illustration that is not associated with an article will be captured as a Standalone Illustration. A Standalone Illustration has all of the properties of a News article and the re-keying rules for title, subtitles, author and abstract apply.

### **Complex Standalone Illustrations**

Standalone Illustrations may have several graphics and text blocks interspersed and overlapping with one another. If a rectangular zone can be constructed around Article Information and Abstract, these elements are captured; otherwise, not.

## **3.10 Content Analysis and OCR Generalities**

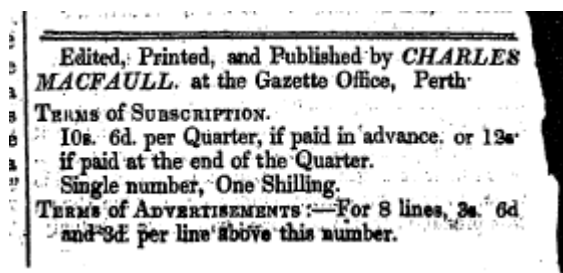
- Page images must be rotated by the Contractor to eliminate visual skew. The de-skew angle applied to the image will be provided as metadata for each page image.
- The newspaper masthead will not be processed for OCR and will not be categorised as an article. Advertising appearing in a masthead, however, will be captured.
- Source page images are in natural reading orientation. However, if a page has content in a different orientation (e.g. landscape table on a portrait page), then a separate zone will be drawn for the content. The zone will be rotated to natural reading orientation (e.g. landscape table on a portrait page would be rotated to portrait orientation) for OCR processing. However, the OCR coordinates and zone coordinates in the deliverables will be "un-rotated" to retain relevance to the original source.
- Abby Finereader is currently being used as the preferred OCR software.

- Both Primary and Custom dictionaries will be used by the Contractor to aid in character recognition.
- The Primary dictionary will be the default English dictionary used by Abbyy FineReader. This is optimal as the Primary dictionary contains a more comprehensive set of word variations and derivatives than a typical Custom dictionary. The Custom dictionary will be created from the “Gazetteer of Australian Place Names” txt file supplied by the Library.
- Unedited OCR text, formatted as ALTO XML, will be captured for each page image.
- Zones on the page will be captured as blocks for each article appearing on the page.
- All text zoned within an article will be OCR’d.
- A single block will consist of one of the following content types:
  - Text only
  - Illustration only
  - Combination of text and illustration
- Block-level position coordinates will be captured.
- Block-level rotation information will be captured.
- Word and character confidence levels will be captured.
- Word-level positional coordinates will be captured.

### 3.11 Colophon – Definition

A colophon is a small paragraph identifying the publisher and place of publication of the newspaper. It may also include subscription details which usually (but not always) appear at the bottom of a column on a front or back page. The colophon should be zoned with the preceding article regardless of article type. The colophon should not be treated as a separate article.

Examples:



rt  
lc  
r

Printed and published by ROBERT N. MYERS, of 205 Lennox st., Richmond, for THE ARGUS AND AUSTRALASIAN LIMITED, at the Registered Office, 365 Elizabeth st. Melbourne.

### 3.12 Masthead – Definition

The Masthead appears at the top of the first page of a newspaper issue. It contains the name of the newspaper (usually in Gothic font), the date of publication, volume and issue details and often advertising and headlines.

To-day's Weather.—Fair or Fine and Warmer.

FEDERAL TAXATION RELIEF FORESHADOWED—Page 13.

THE WHISKY  
**GRANT'S**  
With Flavour  
And Finish  
Real SCOTCH

# The Courier-Mail

THE PUBLIC LIBRARY

Cable and General News,  
Sporting Results, Stories,  
Jokes, Magazine Features,  
Illustrated Throughout

THE  
SUNDAY MAIL

No. 5. (Registered at the General Post Office, Brisbane,  
for transmission by post as a Newspaper.)

FRIDAY, SEPTEMBER 1, 1933.

Phone B2991 (7 lines)

4 SEP 1933

22 PAGES—TWO PENCE

Metadata Encoding and Transmission Standard (METS), Version 1.6

Metadata Object Description Schema (MODS), Version 3.2

### Root Element

|   |   |
|---|---|
| <u>Element Name:</u> <mets:mets>  |   |
| <u>Description:</u> This element is the container element of the Issue XML file |   |
| <u>Occurrence:</u> mandatory, non-repeatable                                    |   |
| <u>Attributes:</u>  |   |
| xmlns:xsi   | Always "http://www.w3.org/2001/XMLSchema-instance"  |
| xmlns:mets  | Always "http://www.loc.gov/METS/"   |
| xmlns:mods  | Always "http://www.loc.gov/mods/v3"   |
| xmlns:premis  | Always "http://www.loc.gov/standards/premis"  |
| xmlns:xlink   | Always "http://www.w3.org/1999/xlink"   |
| Xsi:schemaLocation  | Always<br>"http://www.loc.gov/METS/<br>http://www.loc.gov/standards/mets/mets.xsd<br>http://www.loc.gov/mods/v3<br>http://www.loc.gov/standards/mods/v3/mods-3-2b.xsd<br>http://www.loc.gov/standards/premis/v1<br>http://www.loc.gov/standards/premis/v1/PREMIS-v1-1.xsd " |

### METS Header Elements

#### METS Header Container Element

|  |   |
|--|---|
| <u>Element Name:</u> <mets:metsHdr>  |   |
| <u>Description:</u> This element is the container element of the METS Header information |   |
| <u>Occurrence:</u> mandatory, non-repeatable   |   |
| <u>Attributes:</u>   |   |
| CREATEDATE   | Must be xsd:dateTime compliant. The time zone should be specified as Z (UTC) or (+/-)hh:mm. |
| LASTMODDATE  | Must be xsd:dateTime compliant. The time zone should be specified as Z (UTC) or (+/-)hh:mm. |

## METS Header Agent Element

|  |   |
|--|---|
| <u>Element Name:</u> <mets:agent>  |   |
| <u>Description:</u> Two instances of this element are used in the Issue XML file. The first instance contains the name of the organization responsible for creating the METS record. The second instance contains the name of the software used to create the METS record. |   |
| <u>Occurrence:</u> mandatory, repeatable   |   |
| <u>Attributes:</u>   |   |
| ROLE   | Always "DISSEMINATOR" for name of organization responsible for creating METS record<br>Always "CREATOR" for name of software used to create the METS record |

## METS Header Agent Name Element

|  |  |
|--|--|
| <u>Element Name:</u> <mets:name>   |  |
| <u>Description:</u> This element is a child element of <agent> and contains the name of the organization responsible for creating the METS record when the ROLE="DISSEMINATOR" attribute is present in the corresponding <agent> element. This element contains the name of the software used to create the METS record when the ROLE="CREATOR" attribute is present in the corresponding <agent> element. |  |
| <u>Occurrence:</u> mandatory, repeatable   |  |
| <u>Attributes:</u> none  |  |

## Descriptive Metadata Elements

Descriptive metadata is provided as MODS elements within a <mets:mdWrap MDTYPE="MODS"> wrapper element, e.g.:

```
<mets:dmdSec>
<mets:mdWrap MDTYPE="MODS">
<mets:xmlData>
<mods:mods xmlns="http://www.loc.gov/mods/v3">
...
</mods:mods>
</mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
```

### Descriptive Metadata Container Element

|  |  |
|--|--|
| <u>Element Name:</u> <mets:dmdSec>   |  |
| <u>Description:</u> This element is the container element for descriptive metadata. Multiple occurrences of this element are used to contain descriptive metadata for the issue and any editions, supplements or sections within the issue. Additionally, each article within the issue is contained within separate <mets:dmdSec> elements. |  |
| Multiple <mets:dmdSec> elements are presented in the following sequence. This sequence does not represent the structural hierarchy of the issue. Instead, structural hierarchy is represented in the Structural Map elements (see Sections 0 and 0):   |  |
| <ol style="list-style-type: none"><li>1. Issue</li><li>2. Edition</li><li>3. Supplement</li><li>4. Section</li><li>5. Article</li></ol>  |  |
| <u>Occurrence:</u> mandatory, repeatable   |  |
| <u>Attributes:</u>   |  |
| ID   | First occurrence contains issue XML filename without extension.<br>All occurrences of this element must have an ID |

## Newspaper Title Elements

### Issue Genre

|   |
|---|
| <u>Element Name</u> : <mods:genre>                                  |
| <u>Description</u> : This element always contains “newspaper issue” |
| <u>Occurrence</u> : mandatory, non-repeatable                       |

### Issue Language

|   |                  |
|---|------------------|
| <u>Element Name</u> : <mods:language><br><mods:languageTerm>                            |                  |
| <u>Description</u> : The <mods:languageTerm> element always contains “en” (for English) |                  |
| <u>Occurrence</u> : mandatory, non-repeatable   |                  |
| <u>Attributes</u> :   |                  |
| type  | Always “code”    |
| Authority   | Always “rfc3066” |

### Issue Publication Date

|   |
|---|
| <u>Element Name</u> : <mods:originInfo><br><mods:dateIssued>  |
| <u>Description</u> : The <mods:dateIssued> element contains the issue publication date in “yyyymmdd” format |
| <u>Occurrence</u> : mandatory, non-repeatable   |

### Newspaper Title

|  |
|--|
| <u>Element Name</u> : <mods:relatedItem type=”host”><br><mods:titleInfo><br><mods:title> |
| <u>Description</u> : The <mods:title> element contains the newspaper title.              |
| <u>Occurrence</u> : mandatory, non-repeatable  |

### Newspaper Genre

|  |
|--|
| <u>Element Name</u> : <mods:genre>   |
| <u>Description</u> : This element is a child of the <mods:relatedItem type=”host”> element for newspaper-level information and always contains “newspaper” |
| <u>Occurrence</u> : mandatory, non-repeatable  |

### Newspaper ISSN

|   |
|---|
| <u>Element Name</u> : <mods:identifier>   |
| <u>Description</u> : This element is a child of the <mods:relatedItem type=”host”> element for newspaper-level information. These elements identify the newspaper ISSN prefixed with the label “ISSN” |

Occurrence: mandatory, non-repeatable

### Volume Number

Element Name: <mods:part>  
<mods:detail type="volume">  
<mods:number>

Description: This set of elements is a child of the <mods:relatedItem type="host"> element for newspaper-level information. The <mods:number> element contains the volume number.

Occurrence: optional, non-repeatable

Attributes:

|                         |                 |
|-------------------------|-----------------|
| type (in <mods:detail>) | Always "volume" |
|-------------------------|-----------------|

### Issue Number

Element Name: <mods:part>  
<mods:detail type="issue">  
<mods:number>

Description: This set of elements is a child of the <mods:relatedItem type="host"> element for newspaper-level information. The <mods:number> element contains the issue number.

Occurrence: optional, non-repeatable

Attributes:

|                         |                |
|-------------------------|----------------|
| type (in <mods:detail>) | Always "issue" |
|-------------------------|----------------|

### Edition Elements

Element Name: <mods:mods>  
<mods:titleInfo>  
<mods:partName>  
<mods:partNumber>

Description: This set of elements identifies the edition information and are contained within a separate <mets:dmdSec> container element. The <mods:partName> element contains the name of the edition). The <mods:partNumber> element contains the edition sequence number

Occurrence: <mets:dmdSec> optional , repeatable  
<mods:mods> non-repeatable

Attributes:

|                       |   |
|-----------------------|---|
| id (in <mets:dmdSec>) | Always "modsedition#", where "#" is the edition sequence number |
|-----------------------|---|

## Supplement Elements

|   |   |
|---|---|
| <b>Element Name:</b> <mods:mods><br><mods:titleInfo><br><mods:partName><br><mods:partNumber><br><mods:originInfo><br><mods:dateIssued>  |   |
| <b>Description:</b> This set of elements identifies the supplement information and are contained within a separate <mets:dmdSec> container element. The <mods:partName> element contains the name of the supplement. The <mods:partNumber> element contains the supplement sequence number. The <mods:dateIssued> element contains the supplement date. |   |
| <b>Occurrence:</b> optional , repeatable  |   |
| <b>Attributes:</b>  |   |
| id (in <mets:dmdSec>)   | Always “modssupplement#”, where “#” is the supplement sequence number |

## Section Details

|  |   |
|--|---|
| <b>Element Name:</b> <mods:mods><br><mods:titleInfo><br><mods:partName><br><mods:partNumber>   |   |
| <b>Description:</b> This set of elements identifies the section information and are contained within a separate <mets:dmdSec> container element. The <mods:partName> element contains the name of the section. The <mods:partNumber> element contains the section sequence number. |   |
| <b>Occurrence:</b> optional , repeatable   |   |
| <b>Attributes:</b>   |   |
| id (in <mets:dmdSec>)  | Always “modssection#”, where “#” is the section sequence number |

## Article Elements

Articles appear in the XML within separate <mets:dmdSec> elements

### Article Container Element

|   |   |
|---|---|
| <b>Element Name:</b> <mods:mods>  |   |
| <b>Description:</b> This is the container element for article metadata. |   |
| <b>Occurrence:</b> optional , repeatable                                |   |
| <b>Attributes:</b>  |   |
| id (in <mets:dmdSec>)   | Always “modsarticle#”, where “#” is a sequential number for each article in the issue |

### Article Title and Subtitle

|  |
|--|
| <b>Element Name:</b> <mods:titleInfo><br><mods:title><br><mods:subTitle>   |
| <b>Description:</b> This set of elements is a child of the <mods:mods> container element of the article. The <mods:title> element contains the article title. The <mods:subTitle> element contains the article subtitle. |
| <b>Occurrence:</b> <mods:title> mandatory, non-repeatable<br><mods:subtitle> optional, repeatable  |

### Article Authors

|  |                   |
|--|-------------------|
| <b>Element Name:</b> <mods:name type="personal"><br><mods:namePart><br><mods:role><br><mods:roleTerm type="text">creator</mods:roleTerm>   |                   |
| <b>Description:</b> This set of elements is a child of the <mods:mods> container element of the article. Each instance of these elements contains the name of a single article author, with all associated information. The <mods:namePart> element contains the article author. The <mods:roleTerm> element always contains "creator".. |                   |
| <b>Occurrence:</b> optional , repeatable   |                   |
| <b>Attributes:</b>   |                   |
| type (in <mods:name>)  | Always "personal" |
| type (in <mods:roleTerm>)  | Always "text"     |

### Article Abstract

|  |
|--|
| <b>Element Name:</b> <mods:abstract>   |
| <b>Description:</b> This element is a child of the <mods:mods> container element of the article and contains the article abstract. |
| <b>Occurrence:</b> mandatory, non-repeatable   |

### Article Type

|  |
|--|
| <b>Element Name:</b> <mods:genre>article<br><mods:genre type="articleCategory">  |
| <b>Description:</b> This pair of elements are children of the <mods:mods> container element of the article. The first instance of <mods:genre> always contains "article". The second instance of <mods:genre> has the "type" attribute set to "articleCategory" and contains the article type. |
| <b>Occurrence:</b> mandatory, non-repeatable   |

## Administrative Metadata Elements

Administrative metadata is provided for each deliverable file and the page level TIFF file to which the OCR coordinates apply, as PREMIS elements within the <mets:techMD> and <mets:mdWrap MDTYPE="PREMIS"> wrapper elements. Each deliverable file is sequentially numbered as a "PREMISOBJECT#" in the "id" attribute of the <mets:techMD> element, e.g.:

```
<mets:techMD ID="PREMISOBJECT#">
```

```
<mets:mdWrap MDTYPE="PREMIS">
```

```
<mets:xmlData>
```

```
<mets:xmlData>
```

```
<premis:object xmlns:premis="http://www.loc.gov/standards/premis/v1">
```

```
...
```

```
</premis:object>
```

```
</mets:xmlData>
```

```
</mets:mdWrap>
```

## PREMIS Objects

A set of the following elements are provided for each deliverable file and the page level TIFF file to which the OCR coordinates apply. The order of appearance for these elements is as follows:

1. All page-level TIFF image files
2. All page-level ALTO XML files

| <u>Element Name</u>            | <u>Description</u>  |
|--------------------------------|---|
| <objectIdentifierType>         | Always contains "National Library of Australia"   |
| <objectIdentifierValue>        | Contains the filename of the deliverable  |
| <objectCategory>               | Always contains "file"  |
| <formatName>                   | <u>Page-Level TIFF Image:</u> Contains "TIFF"<br><u>Page-level OCR file:</u> Contains "XML ALTO"  |
| <formatVersion>                | <u>Page-Level TIFF Image:</u> Contains "TIFF 6.0"<br><u>Page-level OCR file:</u> Contains "ALTO schema Version 1.1-04"  |
| <relationshipType>             | Always contains "derivation"  |
| <relationshipSubType>          | Always contains "is derivative of"  |
| <relatedObjectIdentifierType>  | Always contains "National Library of Australia"   |
| <relatedObjectIdentifierValue> | <u>Page-Level TIFF Image:</u> Contains corresponding source image filename<br><br><u>Page-level OCR file:</u> Contains corresponding page-level TIFF image filename                                   |
| <relatedObjectSequence>        | Always contains "0"   |
| <relatedEventIdentifierType>   | Always contains "National Library of Australia"   |
| <relatedEventIdentifierValue>  | This optional element is present only for page-level TIFF images and contains the filename of the corresponding source image, prefixed with the label "deskew-", e.g. "deskew-nlImageSeq-24537-b.tif" |
| <relatedEventSequence>         | Always contains "0"   |

## PREMIS Events

A set of these elements are provided within <mets:digiprovMD ID="PREMISEVENT#"> and <mets:mdWrap MDTYPE="PREMIS"> wrapper elements for each page-level TIFF image. These elements contain the de-skew information for each page-level TIFF image.

| <u>Element</u>                  | <u>Description</u>   |
|---------------------------------|--|
| <eventIdentifierType>           | Always contains "National Library of Australia"  |
| <eventIdentifierValue>          | Contains the filename of the corresponding source image, prefixed with the label "deskew-", e.g. "deskew-nlImageSeq-24537-b.tif"   |
| <eventType>                     | Always contains "deskew"   |
| <eventDateTime>                 | Contains the date and time of deskewing, formatted xsd:dateTime compliant. The time zone should be specified as Z (UTC) or (+/-)hh:mm.   |
| <eventOutcome>                  | Contains the details from skew file (comma separated) e.g.<br><br>"filebase,003.tif,skew, -<br>50,src.9029ae11cc7bca72672eb5e3d00cfd36,check,<br>f259a8df44800646a0cd75988eee1389" |
| <linkingAgentIdentifierType>    | Always contains "National Library of Australia"  |
| <linkingAgentIdentifierValue >  | Always contains "DeskewingSoftware"  |
| <linkingObjectIdentifierType >  | Always contains "National Library of Australia"  |
| <linkingObjectIdentifierValue > | Contains the filename of the corresponding source image  |

## PREMIS Agent

A single set of these elements are provided within <mets:digiprovMD ID="PREMISAGENT#"> and <mets:mdWrap MDTYPE="PREMIS"> wrapper elements. These elements contain the identification information for the deskewing software.

| <u>Element</u>        | <u>Description</u>                              |
|-----------------------|---|
| <agentIdentifierType> | Always contains "National Library of Australia" |

|                        |   |
|------------------------|---|
| <agentIdentifierValue> | Always contains "DeskewingSoftware"                             |
| <agentName>            | Always contains the name and version of deskewing software used |
| <agentType>            | Always contains "deskewing software"                            |

## File Group Elements

Each deliverable file is collected into individual groups based on file type. The following elements provide details for each deliverable file:

### File Group

|   |   |                               |                     |                              |                     |
|---|---|-------------------------------|---------------------|------------------------------|---------------------|
| <u>Element Name:</u> <mets:fileGrp>   |   |                               |                     |                              |                     |
| <u>Description:</u> This is the container element for all deliverables of the same type as specified in the “use” attribute |   |                               |                     |                              |                     |
| <u>Occurrence:</u> mandatory, repeatable  |   |                               |                     |                              |                     |
| <u>Attributes:</u>  |   |                               |                     |                              |                     |
| use   | <table> <tr> <td><u>Page-Level TIFF files:</u></td> <td>Contains “TIFFpage”</td> </tr> <tr> <td><u>Page-level OCR files:</u></td> <td>Contains “ALTOpage”</td> </tr> </table> | <u>Page-Level TIFF files:</u> | Contains “TIFFpage” | <u>Page-level OCR files:</u> | Contains “ALTOpage” |
| <u>Page-Level TIFF files:</u>   | Contains “TIFFpage”   |                               |                     |                              |                     |
| <u>Page-level OCR files:</u>  | Contains “ALTOpage”   |                               |                     |                              |                     |

### File

|   |  |                               |                      |                              |                     |
|---|--|-------------------------------|----------------------|------------------------------|---------------------|
| <u>Element Name:</u> <mets:file>  |  |                               |                      |                              |                     |
| <u>Description:</u> This is a child element of <fileGrp> and contains information for each file within the group. |  |                               |                      |                              |                     |
| <u>Occurrence:</u> mandatory, repeatable  |  |                               |                      |                              |                     |
| <u>Attributes:</u>  |  |                               |                      |                              |                     |
| ID  | Contains the corresponding deliverable filename  |                               |                      |                              |                     |
| ADMID   | Contains the ID value of the corresponding <techMD> element for the file in Administrative Metadata  |                               |                      |                              |                     |
| MIMETYPE  | <table> <tr> <td><u>Page-Level TIFF files:</u></td> <td>Contains “image/tif”</td> </tr> <tr> <td><u>Page-level OCR files:</u></td> <td>Contains “text/xml”</td> </tr> </table> | <u>Page-Level TIFF files:</u> | Contains “image/tif” | <u>Page-level OCR files:</u> | Contains “text/xml” |
| <u>Page-Level TIFF files:</u>   | Contains “image/tif”   |                               |                      |                              |                     |
| <u>Page-level OCR files:</u>  | Contains “text/xml”  |                               |                      |                              |                     |
| SIZE  | Contains size of file in bytes   |                               |                      |                              |                     |
| CHECKSUMTYPE  | Contains “MD5” as type of checksum used  |                               |                      |                              |                     |
| CHECKSUM  | Contains MD5 checksum of file  |                               |                      |                              |                     |

## File Location

|  |  |
|--|--|
| <u>Element Name:</u> <mets:FLocat />   |  |
| <u>Description:</u> This is empty element is a child of the corresponding <file> element and contains file location information. |  |
| <u>Occurrence:</u> mandatory, non-repeatable within <file>   |  |
| <u>Attributes:</u>   |  |
| LOCTYPE  | Always "URL"   |
| xlink:type   | Always "simple"  |
| xlink:href   | Contains path of corresponding file relative to the issue XML file. For the page-level TIFF files which are not being delivered, contains "#". |

## Physical Structural Map Elements

The physical structural map elements provide structural division details related to the pages contained within the issue. These elements are contained within the <structMap id="structmap1" type="physical"> container element.

The structural hierarchy of the pages is obtained from the edition, supplement and section information in the Library supplied source metadata records. Based on the information provided in the Library supplied source metadata records, the hierarchy is structured as follows:

- If page has no edition/supplement/section information, then page is part of the issue
- If page has only edition information, then page is part of the edition within issue
- If page has only supplement information, then page is part of the supplement within issue
- If page has only section information, then page is part of the section within issue
- If page has both edition and section information, then page is part of section within edition within the issue
- If page has both supplement and section information, then page is part of section within supplement within the issue
- If page has both edition and supplement information, then page is part of supplement within the edition within the issue
- If page has edition, supplement and section information, then page is part of section within supplement within edition within issue.

## Issue Division

|  |  |
|--|--|
| <u>Element Name</u> : <mets:div TYPE="issue">                        |  |
| <u>Description</u> : The root <div> corresponds to the entire issue. |  |
| <u>Occurrence</u> : mandatory, non-repeatable                        |  |
| <u>Attributes</u> :  |  |
| TYPE   | Always "issue"   |
| DMDID  | References appropriate IDs of Descriptive Metadata for the issue |

## Edition/Supplement/Section Divisions

|  |   |
|--|---|
| <u>Element Name</u> : <mets:div TYPE="issue"><br><mets:div TYPE="xxx" ORDER="#" DMDID="modsxxx#">  |   |
| <u>Description</u> : When an issue contains edition, supplement or section structures, subordinate <div> elements are used to represent the hierarchical structure.. |   |
| <u>Occurrence</u> : optional, repeatable   |   |
| <u>Attributes</u> :  |   |
| TYPE   | Value relative to corresponding structure type, "edition", "supplement" or "section"      |
| ORDER  | Sequential order of divisions within an issue, starting with "1"                          |
| DMDID  | References appropriate IDs of Descriptive Metadata for the edition, supplement or section |

## Page Divisions

|  |   |
|--|---|
| <b>Element Name:</b> <mets:div TYPE="issue"><br><mets:div ID="divpage#" TYPE="page" ORDER="#"> |   |
| <b>Description:</b> Subordinate <div> elements correspond to each page.                        |   |
| <b>Occurrence:</b> mandatory, repeatable   |   |
| <b>Attributes:</b>   |   |
| TYPE   | Always "page"   |
| ID   | A sequential ID for each page division, formatted as "divpage#", e.g. "divarticle1" |
| ORDER  | Sequential order number of the pages in the issue.                                  |

## Page File Pointers

|  |   |
|--|---|
| <b>Element Name:</b> <mets:div TYPE="issue"><br><mets:div TYPE="page" ORDER="#"><br><mets:fptr FILEID="tiffpage#" /><br><mets:fptr FILEID="altopage#" /> |   |
| <b>Description:</b> The <fptr> elements have "FILEID" attributes indicating the files identified in the File Group elements for each page.               |   |
| <b>Occurrence:</b> mandatory, repeatable   |   |
| <b>Attributes:</b>   |   |
| FILEID   | Contains ID relative to the page file type. |

## **Logical Structural Map Elements**

The logical structural map elements provide structural details related to the articles contained within the issue. These elements are contained within the <structMap id="structmap2" type="logical"> container element.

### **Structural Hierarchy**

The structural hierarchy of the articles is obtained from the edition, supplement and section information in the Library supplied source metadata records for the page on which the article starts. Based on the information provided in the Library supplied source metadata records, the hierarchy is structured as follows:

1. If an article starts on a page with no edition/supplement/section information, then the article is part of the issue
2. If an article starts on a page with only edition information, then the article is part of the edition within issue
3. If an article starts on a page with only supplement information, then the article is part of the supplement within issue
4. If an article starts on a page with only section information, then the article is part of the section within issue
5. If an article starts on a page with both edition and section information, then the article is part of section within edition within the issue
6. If an article starts on a page with both supplement and section information, then the article is part of section within supplement within the issue
7. If an article starts on a page with both edition and supplement information, then the article is part of supplement within the edition within the issue
8. If an article starts on a page with edition, supplement and section information, then the article is part of section within supplement within edition within issue.

### **Article Division Structure**

An article will be represented in the METS Logical Structural Map as follows:

1. The first-level <mets:div> contains the complete article with TYPE=article
2. The second-level <mets:div> represents each article image per page with TYPE=article-part
  - The first <mets:fptr> within the article-part division will point to the page image file with <mets:area> "COORDS" of the article image on the page
  - The second <mets:fptr> within the article-part division will point to the page ALTO file with <mets:area> "BEGIN" attribute pointing to the article-level <ComposedBlock>
3. The third-level <mets:div> is for the zones within an article image with TYPE=article-zone
  - The first <mets:fptr> within the article-zone division will point to the page image file with <mets:area> "COORDS" of the zone image on the page
  - The second <mets:fptr> within the article-zone division will point to the page ALTO file with <mets:area> "BEGIN" attribute pointing to the zone-level <ComposedBlock>

4. Example:

```
<mets:div ID="divarticle1" TYPE="article" DMDID="modsarticle1">
  <mets:div ID="divarticle1-1" TYPE="article-part" ORDER="1">
    <mets:fptr>
      <mets:area FILEID="page image filename" SHAPE="RECT"
        COORDS="coords of article image on page"/>
    </mets:fptr>
    <mets:fptr>
      <mets:area FILEID="page ALTO filename" BETYPE="IDREF"
        BEGIN="ART1"/>
    </mets:fptr>
    <mets:div ID="zone1-1" TYPE="article-zone">
      <mets:fptr>
        <mets:area FILEID="page image filename"
          SHAPE="RECT" COORDS="coords of zone image on
            page"/>
      </mets:fptr>
      <mets:fptr>
        <mets:area FILEID="page ALTO filename"
          BETYPE="IDREF" BEGIN="ZONE1-1"/>
      </mets:fptr>
    </mets:div>
  </mets:div>
```

## Issue Division

|  |  |
|--|--|
| <u>Element Name</u> : <mets:div TYPE="issue">                        |  |
| <u>Description</u> : The root <div> corresponds to the entire issue. |  |
| <u>Occurrence</u> : mandatory, non-repeatable                        |  |
| <u>Attributes</u> :  |  |
| TYPE   | Always "issue"   |
| DMDID  | References appropriate IDs of Descriptive Metadata for the issue |

## Edition/Supplement/Section Divisions

|  |   |
|--|---|
| <u>Element Name</u> : <mets:div TYPE="issue"><br><mets:div TYPE="xxx" ORDER="#" DMDID="modsxxx#">  |   |
| <u>Description</u> : When an issue contains edition, supplement or section structures, subordinate <div> elements are used to represent the hierarchical structure.. |   |
| <u>Occurrence</u> : optional, repeatable   |   |
| <u>Attributes</u> :  |   |
| TYPE   | Value relative to corresponding structure type, "edition", "supplement" or "section"      |
| ORDER  | Sequential order of divisions within an issue, starting with "1"                          |
| DMDID  | References appropriate IDs of Descriptive Metadata for the edition, supplement or section |

## Article Divisions

|  |   |
|--|---|
| <b>Element Name:</b> <mets:div ID="divarticle#" TYPE="article" DMDID="modsarticle#">   |   |
| <b>Description:</b> Each article is contained within a first-level <mets:div> element. Subsequent <mets:div> elements are used to contain the article-part and article-zones |   |
| There are no article file pointers (<mets:fptr>) or article file area (<mets:area>) elements for article divisions.  |   |
| <b>Occurrence:</b> mandatory, repeatable   |   |
| <b>Attributes:</b>   |   |
| TYPE   | For articles divisions, the attribute value is "article"                                  |
| ID   | A sequential ID for each article division, formatted as "divarticle#", e.g. "divarticle1" |
| DMDID  | References appropriate IDs of Descriptive Metadata for the "article"-type <div>.          |

## Article Part Divisions

|  |   |
|--|---|
| <b>Element Name:</b> <mets:div ID="divarticle#-#" TYPE="article-part" ORDER=" #">  |   |
| <b>Description:</b> Each article-part is contained within a second-level <mets:div> and represents an article image on a single page. The article-part division is subordinate to the article-level division. Article parts for articles which span across pages are contained within separate <div> elements. |   |
| Subordinate <mets:div> elements are used to contain the article-zones within an article-part.  |   |
| <b>Occurrence:</b> mandatory, repeatable   |   |
| <b>Attributes:</b>   |   |
| TYPE   | For article-parts, the attribute value is "article-part".   |
| ORDER  | Sequential order number of article parts when article is contained in multiple columns on a single page or spans across pages                               |
| ID   | A sequential ID for each article-part division within an article, formatted as "divarticle#-#", e.g. "divarticle1-2" (second article part within article 1) |

## Article Part File Pointers

|  |
|--|
| <b>Element Name:</b> <mets:div ID="divarticle#-#" TYPE="article-part" ORDER=" #"><br><mets:fptr> |
| <b>Description:</b> This element is simply a container element for <mets:area>.                  |
| <b>Occurrence:</b> mandatory, repeatable   |
| <b>Attributes:</b>   |

## Article Part File Areas

|   |   |
|---|---|
| <b>Element Name:</b> <mets:area FILEID="page image filename" SHAPE="RECT" COORDS=" x1,y1,x2,y2"/><br><mets:area FILEID="page OCR filename" BETYPE="IDREF" BEGIN=" ART#"/>   |   |
| <b>Description:</b> The first instance of this element is contained within a <mets:fptr> element and provides coordinate information for the article image, relative to the page image.<br><br>The second instance of this element is contained within a separate <mets:fptr> element and provides a reference to the block ID of the article part text in the ALTO file. |   |
| <b>Occurrence:</b> mandatory, repeatable  |   |
| <b>Attributes:</b>  |   |
| FILEID  | <u>Page image file:</u> Contains page image filename<br><u>Page OCR file:</u> Contains ALTO XML filename        |
| SHAPE   | Used only for page image file, always "RECT"  |
| COORDS  | Used only for page image file, contains positional coordinates of the article image relative to the page image. |
| BETYPE  | Used only for page-level OCR files, contains "IDREF"  |
| BEGIN   | Used only for page-level OCR files, contains ID of article-level <ComposedBlock> in ALTO XML file.              |

## Article Zone Divisions

|   |  |
|---|--|
| <u>Element Name:</u> <mets:div ID="artzone#-#" TYPE="article-zone" >  |  |
| <u>Description:</u> Each article-zone is contained within a third-level <mets:div> and represents an article image on a single page. The article-zone division is subordinate to the article-level division and the article-part division. Each article-zone represents a single zone image within the article image. |  |
| <u>Occurrence:</u> mandatory, repeatable  |  |
| <u>Attributes:</u>  |  |
| TYPE  | For article-zone, the attribute value is "article-zone".   |
| ID  | A sequential ID for each article-zone division within an article, formatted as "artzone#-#", e.g. "artzone1-2" (second article zone within article-part 1). The article zone number is sequential within the article across pages. |

## Article Zone File Pointers

|  |  |
|--|--|
| <u>Element Name:</u> <mets:div TYPE="issue"><br><mets:div ID="artzone#-#" TYPE="article-zone" ><br><mets:fptr>             |  |
| <u>Description:</u> Within article zone divisions, the <mets:fptr> elements are simply container elements for <mets:area>. |  |
| <u>Occurrence:</u> mandatory, repeatable   |  |
| <u>Attributes:</u> None  |  |

## Article Zone File Areas

|  |   |
|--|---|
| <p><b>Element Name:</b> &lt;mets:area FILEID="page image filename" SHAPE="RECT" COORDS=" x1,y1,x2,y2"/&gt;<br/>&lt;mets:area FILEID="page OCR filename" BETYPE="IDREF" BEGIN=" ZONE#-#"/&gt;</p>   |   |
| <p><b>Description:</b> The first instance of this element is contained within a &lt;mets:fptr&gt; element and provides coordinate information of the zone image, relative to the page image.</p> <p>The second instance of this element is contained within a separate &lt;mets:fptr&gt; element and provides a reference to the block ID of the zone-level text in the ALTO file.</p> |   |
| <p><b>Occurrence:</b> mandatory, repeatable</p>  |   |
| <p><b>Attributes:</b></p>  |   |
| FILEID   | <p><u>Page image file:</u> Contains page image filename</p> <p><u>Page OCR file:</u> Contains ALTO XML filename</p> |
| SHAPE  | Used only for page image file, always "RECT"  |
| COORDS   | Used only for page image file, contains positional coordinates of the article image relative to the page image.     |
| BETYPE   | Used only for page-level OCR files, contains "IDREF"  |
| BEGIN  | Used only for page-level OCR files, contains ID of article-level <ComposedBlock> in ALTO XML file.                  |

Analyzed Layout and Text Object (ALTO), Version 1-1-041

**<Description>**

This is the container element containing information about the ALTO file and the software used to create the OCR text.

**<MeasurementUnit>**

This element contains the unit of measurement used in the ALTO file, expressed as “inch1200”

**<sourceImageInformation>**

This is the container element for the image used as the source for OCR text.

**<fileName>**

This element contains the path and filename of the source image file.

**<OCRProcessing>**

This is the container element for the software information used to create the OCR text.

**<ocrProcessingStep>**

This is the container element for each OCR processing step.

**<processingDateTime>**

This element contains the date and time on which the OCR was processed

**<processingAgency>**

This element contains the name of the agency which performed the OCR processing

**<processingSoftware>**

This is the container element for the OCR processing software information

**<softwareCreator>**

This element contains the name of the creator of the OCR software, i.e. “Abbyy”

**<softwareName>**

This element contains the name of the OCR software, i.e. “FineReader”

**<softwareVersion>**

This element contains the software version number, i.e. “8.0”

**<postProcessingStep>**

This is the container element for post-processing steps.

**<processingStepDescription>**

This element contains a description of the processing step performed.

### **<Styles>**

This is the container element for style information in the OCR file.

### **<TextStyle/>**

This empty element contains a unique ID for each text style used in the OCR file. The “fontsize” attribute contains the size of the font of the text style.

### **<ParagraphStyle ID="PAR1" ALIGN="Left"/>**

This empty element contains a unique ID for each paragraph style used in the OCR file. The “align” attribute contains the alignment value for the paragraph style, i.e. “Left”

### **<Layout>**

This is the container element for the content information in the OCR file.

### **<Page>**

This element identifies the page area of the page in the OCR file. The “id” attribute contains a unique ID for the page and the “height” and “width” attribute contains the height and width measurements of the full page.

### **<TopMargin/>**

This empty element contains the information for the top margin of the page.

This element has the following attributes:

|         |                                       |
|---------|---------------------------------------|
| ID:     | unique ID for the margin element      |
| HPOS:   | Horizontal position upper/left corner |
| VPOS:   | Vertical position upper/left corner   |
| WIDTH:  | Width                                 |
| HEIGHT: | Height                                |

### **<LeftMargin/>**

This empty element contains the information for the left margin of the page.

This element has the following attributes:

|         |                                       |
|---------|---------------------------------------|
| ID:     | unique ID for the margin element      |
| HPOS:   | Horizontal position upper/left corner |
| VPOS:   | Vertical position upper/left corner   |
| WIDTH:  | Width                                 |
| HEIGHT: | Height                                |

### **<RightMargin/>**

This empty element contains the information for the right margin of the page.

This element has the following attributes:

|         |                                       |
|---------|---------------------------------------|
| ID:     | unique ID for the margin element      |
| HPOS:   | Horizontal position upper/left corner |
| VPOS:   | Vertical position upper/left corner   |
| WIDTH:  | Width                                 |
| HEIGHT: | Height                                |

### **<BottomMargin/>**

This empty element contains the information for the bottom margin of the page.

This element has the following attributes:

|         |                                       |
|---------|---------------------------------------|
| ID:     | unique ID for the margin element      |
| HPOS:   | Horizontal position upper/left corner |
| VPOS:   | Vertical position upper/left corner   |
| WIDTH:  | Width                                 |
| HEIGHT: | Height                                |

### **<PrintSpace>**

This element contains and defines the boundaries of the OCR text on the page. This element has the following attributes:

|         |   |
|---------|---|
| ID:     | unique ID for print space                             |
| PC:     | Confidence level of the OCR. A value between 0 and 1. |
| HPOS:   | Horizontal position upper/left corner                 |
| VPOS:   | Vertical position upper/left corner                   |
| WIDTH:  | Width   |
| HEIGHT: | Height  |

### **<ComposedBlock>**

The top-level instance of this element is used to contain the content for a single article on the page.

Subordinate instances of this element within the article-level <ComposedBlock> represent each article zone within the page. Each zone-level <ComposedBlock> element will contain nested <TextBlock> to contain paragraph text.

Additionally, a single <ComposedBlock> will be used to contain nested <ComposedBlock> for illustrations and associated caption text. The illustration and the caption text will be contained within separate zone-level <ComposedBlock> elements within a single parent <ComposedBlock>

This element has the following attributes:

|           |   |
|-----------|---|
| ID:       | Unique ID for the element,<br>"ART#" for article-level blocks<br>"ZONE#-#" for article-zone level blocks<br>"ILLBLOCK#" for blocks containing illustration and associated caption text. |
| ROTATION: | Degree of rotation expressed in CCW°  |
| HPOS:     | Horizontal position upper/left corner   |
| VPOS:     | Vertical position upper/left corner   |
| WIDTH:    | Width   |
| HEIGHT:   | Height  |

### **<TextBlock>**

This element contains the paragraph-level text content or the text of a caption associated within an illustration.

This element has the following attributes:

|            |                                       |
|------------|---------------------------------------|
| ID:        | Unique ID for the element             |
| STYLEREFS: | Reference to paragraph style ID       |
| HPOS:      | Horizontal position upper/left corner |
| VPOS:      | Vertical position upper/left corner   |

WIDTH: Width  
HEIGHT: Height

### **<Illustration>**

This element contains the information for a zone consisting of only an illustration.

When associated within a caption, the <Illustration> is contained within a single <ComposedBlock> and the associated captions are contained with a separate <ComposedBlock>. Both of these, in turn, are contained within a single <ComposedBlock>

This element has the following attributes:

ID: Unique ID for the element  
TYPE: One of the valid illustration types  
HPOS: Horizontal position upper/left corner  
VPOS: Vertical position upper/left corner  
WIDTH: Width  
HEIGHT: Height

### **<TextLine>**

This element contains a single line of text within the paragraph.

This element has the following attributes:

ID: Unique ID for the element  
STYLEREFS: Reference to text style ID  
HPOS: Horizontal position upper/left corner  
VPOS: Vertical position upper/left corner  
WIDTH: Width  
HEIGHT: Height

### **<String>**

This empty element represents a single string within a line of text.

This element has the following attributes:

ID: Unique ID for the element  
CONTENT: The character content of the string  
WC: The word confidence level  
CC: The character confidence level  
HPOS: Horizontal position upper/left corner  
VPOS: Vertical position upper/left corner  
WIDTH: Width  
HEIGHT: Height

### **<SP>**

This empty element represents white space within a line of text.

This element has the following attributes:

ID: Unique ID for the element  
HPOS: Horizontal position upper/left corner  
VPOS: Vertical position upper/left corner  
WIDTH: Width

### **<HYP>**

A hyphenation character. Can appear only at the end of a line.

**6.1 Sample XML METS/MODS file at Issue Level**

[http://www.nla.gov.au/ndp/project\\_details/nla.news-issn18339719\\_19450913.xml](http://www.nla.gov.au/ndp/project_details/nla.news-issn18339719_19450913.xml)

**6.2 Sample XML ALTO (OCR) file at Page Level**

[http://www.nla.gov.au/ndp/project\\_details/nlaImageSeq-33386-b.xml](http://www.nla.gov.au/ndp/project_details/nlaImageSeq-33386-b.xml)