

## Use of METS and ALTO for the Australian Newspapers Digitisation Program at the National Library of Australia

Author: Bronwyn Lee (Business Analyst – ANDP)  
Version: 1.1  
Date: 7 May 2008

### Introduction

This document is for NLA staff, NDP stakeholders and anyone else interested in how METS and ALTO is being applied to newspaper digitisation at NLA. NLA has looked at other digitisation projects and has tried to follow common practices while adapting them to the NDP's particular needs. This document does not recommend how METS and ALTO should be used but simply describes what has been decided for the NDP. Some familiarity with METS and OCR processing is assumed.

### The Australian Newspapers Digitisation Program

The Australian Newspapers Digitisation Program is developing a service which will provide Web-based access to a range of digitised out-of-copyright Australian newspaper titles. In the initial phase one major newspaper from each Australian state and territory will be digitised, amounting to 3.5 million pages of content. State and territory libraries and other owners of quality master microfilm will provide microfilm versions of relevant newspapers for the digitisation process. The microfilm will be scanned and digital images created. The digital images will be converted into full text searchable files through the use of Optical Character Recognition (OCR) technology. Content analysis by human operators will identify articles composed of page segments (zones) and will create metadata for the articles. OCR and content analysis is being undertaken by Apex CoVantage LLC.

Apex will provide

1. page-level image files to which the OCR results apply. These files are derived from the page images provided by NLA by rotating them to eliminate visual skew
2. article-level image files derived from the page images in 1.
3. page-level PDF files
4. an xml file for each page, conforming to the ALTO schema and containing the results of OCR
5. an xml file for each issue, conforming to the METS schema and containing most of the human supplied metadata for each issue.

### METS and ALTO as metadata exchange formats

METS and ALTO are being used to transfer metadata from Apex to NLA. NLA will extract metadata from the ALTO and METS files and store them in the repository database's internal format. It has not been decided whether or how long the METS and ALTO files will be stored in the repository. METS and ALTO may be used to exchange metadata with other systems in the future.

## METS

NLA is using METS as follows:

METS document	<p>There is one METS document for each issue of a newspaper.</p> <ul style="list-style-type: none"> <li>In the NDP, an issue is defined as a particular date on which a newspaper was published. An issue may have editions and/or supplements and/or sections. This is to allow page sequence numbers, used in delivering pages to users, to be unique within an edition/ supplement/ section. Currently edition and section are not being used (edition number and section number are both zero).</li> </ul>
METS header <metsHdr>	<p>Contains information about the METS document itself including agent and software that created it and date of creation.</p>
Descriptive metadata <dmdSec>	<p>Contains the bibliographic metadata describing the content of the newspaper issue.</p> <ul style="list-style-type: none"> <li>There is one &lt;dmdSec&gt; for the issue and a &lt;dmdSec&gt; for each article in the issue. There is also a &lt;dmdSec&gt; for each edition, supplement or section if any.</li> <li>Each &lt;dmdSec&gt; contains a MODS record.</li> <li>The issue &lt;dmdSec&gt; includes publication date and volume and issue number.</li> <li>The article &lt;dmdSec&gt; includes title, subtitle, author, abstract.</li> <li>The edition, supplement or section &lt;dmdSec&gt; includes part number and part name (e.g. Late Edition) if any.</li> </ul>
Administrative metadata <amdSec>	<p>Contains technical and digital provenance metadata about the files.</p> <ul style="list-style-type: none"> <li>There is only one &lt;amdSec&gt; per METS document.</li> <li>Contains a &lt;techMD&gt; for each file. Each &lt;techMD&gt; contains a PREMIS Object record.</li> <li>Contains a &lt;digiprovMD&gt; for each page-level TIFF file. Each of these &lt;digiprovMD&gt; contains a PREMIS Event record with the parameters used in deskewing of the source TIFF.</li> <li>Contains one &lt;digiprovMD&gt; containing a PREMIS Agent record describing the deskewing software associated with all the PREMIS Events.</li> <li>&lt;rightsMD&gt; and &lt;sourceMD&gt; are not being used.</li> </ul>
File section <fileSec>	<p>Lists all the files which are being transferred and to which the metadata in this METS document pertains.</p> <ul style="list-style-type: none"> <li>Contains a file group &lt;fileGrp&gt; for each type of file being delivered, i.e. page-level TIFF, page-level PDF, page-level ALTO XML file, article-level TIFF.</li> <li>Each &lt;fileGrp&gt; contains a &lt;file&gt; element for each file in the group.</li> <li>&lt;file&gt; includes file size, checksum, and location of the file. There is also a link to the administrative metadata in &lt;amdSec&gt; about the file.</li> </ul>
Structural map <structMap>	<p>The structural map shows how the files fit together so that a system can reconstruct and deliver the digital objects.</p> <ul style="list-style-type: none"> <li>There are two structural maps: a 'physical' one which lists the pages and a 'logical' one which lists the articles.</li> <li>Each page has an order number and pointers to page-level files in &lt;fileSec&gt;.</li> <li>Articles do not have an order number as articles are only 'ordered' in newspapers according to the pages they occur on and their position in the pages.</li> <li>Each zone within an article has an order number which</li> </ul>

	<p>specifies the reading order of the zones within the article.</p> <ul style="list-style-type: none"> <li>• Each article has a pointer to an article-level file in &lt;fileSec&gt; and pointers to parts within page-level files in &lt;fileSec&gt;.</li> <li>• Each zone has pointers to parts of page-level files in &lt;filesec&gt;.</li> <li>• Coordinates give the article or zone position in the page-level image files</li> <li>• ID references give the position of the article or zone text in the page-level ALTO XML file.</li> </ul>
--	--

## ALTO

NLA is using ALTO as follows:  
There is one ALTO file for each page.

<Description>	The container element containing information about the ALTO file. Contains: <sourceImageInformation> <OCRProcessing>
<Styles>	The container element for style information in the OCR file. Contains: <TextStyle/> <ParagraphStyle ID="" ALIGN="">
<Layout>	The container element for the content information in the OCR file. Contains a <Page> element which contains: <TopMargin> <LeftMargin> <RightMargin> <BottomMargin> <PrintSpace>
<PrintSpace>	Contains <ComposedBlock> elements.
<ComposedBlock>	The top-level instance of <ComposedBlock> is used to contain the content for a single article on the page. Subordinate instances of <ComposedBlock> within the article-level <ComposedBlock> represent each article zone within the page. Each zone-level <ComposedBlock> element will contain <TextBlock> elements to contain paragraph text. Additionally, a single <ComposedBlock> will be used to associate an illustration and its caption. This <ComposedBlock> will contain nested <ComposedBlock> elements for the illustration and the caption.  <ComposedBlock> elements may contain: <TextBlock> <Illustration> <ComposedBlock>
<TextBlock>	Contains <TextLine> elements which represent a single line of text within the paragraph: <TextLine> contains: <String> (represents a single string of characters within a line of text) <SP> (represents white space within a line of text)

## Links

Sample METS file:

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

Sample ALTO file:

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