

Use of METS and ALTO in the Australian Newspapers Digitisation Program (ANDP) at the National Library of Australia (NLA)

May 2010

Introduction

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

The Australian Newspaper Digitisation Program

The Australian Newspaper Digitisation Program has developed a service which provides Web-based access to a range of digitised out-of-copyright Australian newspaper titles. State and Territory libraries and other owners of quality master microfilm provide microfilm versions of relevant newspapers for the digitisation process. The microfilm is scanned and digital images created. Digital images may also be created by scanning hard (print) copies of newspaper pages where microfilm is unavailable or unsuitable.

The digital images are converted into full text searchable files through the use of Optical Character Recognition (OCR) technology. Content analysis by human operators identifies articles composed of page segments (zones) and creates metadata for the articles. Apex Publishing LLC was responsible for OCR and content analysis in the initial phase. Subsequently a panel of OCR and content analysis providers was established to cater for the expanding Program.

OCR contractors process page-level image files provided by NLA. These images may be rotated to eliminate visual skew by the OCR contractor's software.

OCR contractors provide:

1. an xml file for each page, conforming to the ALTO schema and containing the results of OCR
2. 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 OCR contractors to NLA. NLA will extract metadata from the ALTO and METS files and store them in the repository database's internal format. The METS and ALTO files themselves may be stored in the repository for an indefinite period.

METS

NLA is using METS as follows:

METS file	There is one METS file for each issue of a newspaper. <ul style="list-style-type: none">• In the ANDP, an issue is defined as a particular date on which a newspaper title 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. Where edition / supplement / section
-----------	---

	are not being used (as is the case with most titles), ANDP sets the edition /supplement / section numbers to zero.
METS header <metsHdr>	Contains information about the METS file itself including agent and software that created it and date of creation.
Descriptive metadata <dmdSec>	<p>Contains the bibliographic metadata describing the content of the newspaper issue.</p> <ul style="list-style-type: none"> • There is one <dmdSec> for the issue and a <dmdSec> for each article in the issue. There is also a <dmdSec> for each edition, supplement or section if any. • Each <dmdSec> contains a MODS record. • The issue <dmdSec> includes publication date and volume and issue number. • The article <dmdSec> includes elements re-keyed or corrected by the OCR contractor, usually title, subtitle, author, abstract. • The edition, supplement or section <dmdSec>, if any, includes a mandatory sequence number and a name (e.g. Late Edition, Melbourne edition). Content analysis specifications for contractors should contain advice on how edition, supplement or section metadata is to be applied in particular titles.
Administrative metadata <amdSec>	<p>Contains technical and digital provenance metadata about the files.</p> <ul style="list-style-type: none"> • There is only one <amdSec> per METS document. • Contains a <techMD> for each file. Each <techMD> contains a PREMIS Object record. • Contains a <digiprovMD> for each page-level IMAGE file. Each of these <digiprovMD> contains a PREMIS Event record with the parameters used in deskewing of the source image. • Contains one <digiprovMD> containing a PREMIS Agent record describing the deskewing software associated with all the PREMIS Events. • <rightsMD> and <sourceMD> are not being used.
File section <fileSec>	<p>Lists all the source files and deliverable files, to which the metadata in this METS document pertains.</p> <ul style="list-style-type: none"> • Contains a file group <fileGrp> for the page-level source image files and for the page-level ALTO XML files. • Each <fileGrp> contains a <file> element for each file in the group. • <file> includes file size, checksum, and location of the file in the directory structure relative to the METS file and the file name. There is also a link to the administrative metadata in <amdSec> about the file.
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"> • There are two structural maps: a 'physical' one which lists the pages and a 'logical' one which lists the articles. • Each page has an order number and pointers to page-level files in <fileSec>. • 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. • Each article has at least one article-part. An article which runs over two or more pages will have an article-part for

	<p>each page that the article occurs on. Each article-part has an order number which specifies the reading order of the parts of the article e.g. an article could start on the last page (the back page) and continue on the second-last page.</p> <ul style="list-style-type: none"> • Each article-part has at least one article-zone. Each article-zone has an order number which specifies the reading order of the zones within the article-part. • Each article-part and article-zone has pointers to parts of page-level files in <filesec>. • Coordinates give the article part or zone position in the page-level image files • ID references give the position of the article part and zone text in the page-level ALTO XML file.
--	--

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>	A 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)

An ALTO file must contain at least one <String> containing at least one character of text.

Links

Information about the Australian Newspaper Digitisation Program is at <http://www.nla.gov.au/ndp/>

Sample METS and ALTO files can be found linked to the ANDP Project Details page at http://www.nla.gov.au/ndp/project_details/

APPENDIX 1 : ANDP METS specification

METS specification

The file name must begin with "issue-" (lower case) and end in ".xml"

A sample XML METS/MODS file at Issue Level can be found linked to the ANDP Project Details page at http://www.nla.gov.au/ndp/project_details/:

Note: Specification and ingest system were developed for the versions current at the time:

- Metadata Encoding and Transmission Standard (METS), Version 1.6
- Metadata Object Description Schema (MODS), Version 3.2

Contractors wishing to use later versions should provide a pilot sample for testing before proceeding to production. ANDP has accommodated and will continue to accommodate later versions of the schemas where possible.

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	"http://www.w3.org/2001/XMLSchema-instance"
xmlns:mets	"http://www.loc.gov/METS/"
xmlns:mods	"http://www.loc.gov/mods/v3"
xmlns:premis	"http://www.loc.gov/standards/premis"
xmlns:xlink	"http://www.w3.org/1999/xlink"
Xsi:schemaLocation	"http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-2b.xsd http://www.loc.gov/standards/premis/v1 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 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>	
<p><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.</p> <p>Multiple <mets:dmdSec> elements are presented in the following sequence. This sequence does not define the structural hierarchy of the issue. Instead, structural hierarchy is defined in the Structural Map elements.</p> <p>Issue Edition Supplement Section Article</p>	
<u>Occurrence:</u> mandatory, repeatable	
<u>Attributes:</u>	
ID	First occurrence MUST contain issue XML filename without the “.xml” extension e.g. if the issue file has the name issue-slv.news-issn08136017_18750121.xml, the first <mets:dmdSec> should be <mets:dmdSec ID="issue-

	slv.news-issn08136017_18750121">
	All occurrences of this element must have an ID. Except for the first occurrence the ID can be anything.

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> <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> <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”> <mods:titleInfo> <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"
<u>Occurrence:</u> mandatory, non-repeatable

Volume Number

<u>Element Name:</u> <mods:part> <mods:detail type="volume"> <mods:number>	
<u>Description:</u> 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.	
<u>Occurrence:</u> optional, non-repeatable	
<u>Attributes:</u>	
type (in <mods:detail>)	Always "volume"

Issue Number

<u>Element Name:</u> <mods:part> <mods:detail type="issue"> <mods:number>	
<u>Description:</u> 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.	
<u>Occurrence:</u> optional, non-repeatable	
<u>Attributes:</u>	
type (in <mods:detail>)	Always "issue"

Edition Elements

<p><u>Element Name:</u> <mods:mods> <mods:titleInfo> <mods:partName> <mods:partNumber></p>	
<p><u>Description:</u> 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) and is mandatory. The <mods:partNumber> element contains the edition sequence number</p>	
<p><u>Occurrence:</u> <mets:dmdSec> optional , repeatable <mods:mods> non-repeatable</p>	
<p><u>Attributes:</u></p>	
<p>id (in <mets:dmdSec>)</p>	<p>e.g. "modsedition#", where "#" is the edition sequence number</p>

In the following example there were two sets of pages published on the same date: pages 1-24 in an unnamed main edition and pages 1-2 in a "Special Edition". Both editions have their own <mets:dmdSec>:

```
<mets:dmdSec ID="modsedition1">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods xmlns="http://www.loc.gov/mods/v3">
        <mods:titleInfo>
          <mods:partName></mods:partName>
          <mods:partNumber>1</mods:partNumber>
        </mods:titleInfo>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
<mets:dmdSec ID="modsedition2">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods xmlns="http://www.loc.gov/mods/v3">
        <mods:titleInfo>
          <mods:partName> SPECIAL EDITION.</mods:partName>
          <mods:partNumber>2</mods:partNumber>
        </mods:titleInfo>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
```

Supplement Elements

<u>Element Name:</u> <mods:mods> <mods:titleInfo> <mods:partName> <mods:partNumber> <mods:originInfo> <mods:dateIssued>	
<u>Description:</u> 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.	
<u>Occurrence:</u> optional , repeatable	
<u>Attributes:</u>	
id (in <mets:dmdSec>)	e.g. “modssupplement#”, where “#” is the supplement sequence number

Section Details

<u>Element Name:</u> <mods:mods> <mods:titleInfo> <mods:partName> <mods:partNumber>	
<u>Description:</u> 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.	
<u>Occurrence:</u> optional , repeatable	
<u>Attributes:</u>	
id (in <mets:dmdSec>)	e.g. “modssection#”, where “#” is the section sequence number

Article Elements

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

Article Container Element

<u>Element Name:</u> <mods:mods>

<u>Description</u> : This is the container element for article metadata.	
<u>Occurrence</u> : optional , repeatable	
<u>Attributes</u> :	
id (in <mets:dmdSec>)	e.g. “modsarticle#”, where “#” is a sequential number for each article in the issue

Article Title and Subtitle

<u>Element Name</u> : <mods:titleInfo> <mods:title> <mods:subTitle>	
<u>Description</u> : 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.	
<u>Occurrence</u> : <mods:title> mandatory, non-repeatable <mods:subTitle> optional, repeatable	

Article Authors

<u>Element Name</u> : <mods:name type="personal"> <mods:namePart> <mods:role> <mods:roleTerm type="text">creator</mods:roleTerm>	
<u>Description</u> : 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”..	
<u>Occurrence</u> : optional , repeatable	
<u>Attributes</u> :	
type (in <mods:name>)	Always “personal”
type (in <mods:roleTerm>)	Always “text”

Article Abstract

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

Article Type

<u>Element Name</u> : <mods:genre>article <mods:genre type="articleCategory">
<u>Description</u> : 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.
<u>Occurrence</u> : mandatory, non-repeatable

Administrative Metadata Elements

Administrative metadata is provided for each deliverable file and the page level IMAGE 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 IMAGE file to which the OCR coordinates apply. The order of appearance for these elements is as follows:

All page-level IMAGE image files

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 IMAGE Image</u> : Contains "IMAGE" <u>Page-level OCR file</u> : Contains "XML ALTO"
<formatVersion>	<u>Page-Level IMAGE Image</u> : Contains "IMAGE 6.0" <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 IMAGE Image</u> : Contains corresponding source image filename <u>Page-level OCR file</u> : Contains corresponding page-level IMAGE image filename
<relatedObjectSequence>	Always contains “0”
<relatedEventIdentifierType>	Always contains “National Library of Australia”
<relatedEventIdentifierValue>	This optional element is present only for page-level IMAGE images and contains the filename of the corresponding source image, prefixed with the label “deskew-”, e.g. “deskew-nlaImageSeq-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 IMAGE image. These elements contain the de-skew information for each page-level IMAGE 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-nlaImageSeq-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. “filebase,003.tif,skew, -50,src.9029ae11cc7bca72672eb5e3d00cfd36,check, f259a8df44800646a0cd75988eee1389”
<linkingAgentIdentifierType>	Always contains “National Library of Australia”
<linkingAgentIdentifierValue>	Always contains “DeskewingSoftware”
<linkingObjectIdentifierType>	Always contains “National Library of Australia”
<linkingObjectIdentifier>	Contains the filename of the corresponding

Value>	source image
--------	--------------

NOTE RE DESKEW ANGLE AND CROPPING AFTER DESKEWING

This should be expressed in degrees multiplied by 100. If the rotation is anticlockwise, the number is positive; if the rotation is clockwise the number is negative. Degrees should always be less than 180 degrees (in practice it is usually less than 1 degree). For example:

- a rotation of 1 degree anticlockwise is expressed as "100"
- a rotation of 1 degree clockwise is expressed as "-100"
- a rotation of 359.45 degrees anticlockwise is expressed as "-55"

The ANDP ingest system assumes, that the contractor rotates the image at the centre point, and then does a centre crop on the rotated images to derive word and zone coordinates in the METS and ALTO files.

For Center cropping, the cropping steps are:

1. The original image (Image1) had dimensions h1 & w1 (height & width)
2. The deskewed image (Image2) had the dimensions h2 & w2.
3. Understandably h2>h1 & w2>w1.
4. The cropping has to be done on all 4 sides.
5. Let $h3 = (h2-h1)/2$ & $w3 = (w2-w1)/2$
6. From the right side of the image remove a rectangular strip of height h2 & width w3.
7. From the left side of the image remove a rectangular strip of height h2 & width w3.
8. After this step, the image dimensions would be h2, w1.
9. From the bottom side of the image, remove a rectangular strip of height h3 & width w1.
10. From the top side of the image, remove a rectangular strip of height h3 & width w1.
11. After this step, the image dimensions would be h1, w1.

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	<u>Page-Level IMAGE files:</u> Contains “IMAGEpage” <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	<u>Page-Level IMAGE files:</u> Contains “image/tif” <u>Page-level OCR files:</u> Contains “text/xml”
SIZE	Contains size of file in bytes
CHECKSUMTYPE	Contains “MD5” or “SHA1” as type of checksum used
CHECKSUM	Contains 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 IMAGE 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 may be supplied by the Library as source metadata records or may be determined by the contractor during processing, depending on the work order specification. 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"> <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

In the following example there were two sets of pages published on the same date: pages 1-24 (image sequence 2802043 to 2802066) in an unnamed main edition and pages 1-2 (image sequence 2802067-2802068) in a "Special Edition". There are separate <div>s for each edition:

```
<mets:structMap ID="structmap1" TYPE="physical">
  <mets:div TYPE="issue" DMDID="issue-nla.news-issn03126323_19160603">
    <mets:div TYPE="edition" ORDER="1" DMDID="modsedition1">
      <mets:div ID="divpage1" TYPE="page" ORDER="1">
        <mets:fptr FILEID=" nlaImageSeq-2802043-b.tif"/>
        <mets:fptr FILEID=" nlaImageSeq-2802043-b.xml"/>
      </mets:div>
      .....
      .....
      <mets:div ID="divpage24" TYPE="page" ORDER="24">
        <mets:fptr FILEID=" nlaImageSeq-2802066-b.tif"/>
        <mets:fptr FILEID=" nlaImageSeq-2802066-b.xml"/>
      </mets:div>
```

```

</mets:div>
<mets:div TYPE="edition" ORDER="2" DMDID="modsedition2">
<mets:div ID="divpage25" TYPE="page" ORDER="1">
  <mets:fptr FILEID=" nlaImageSeq-2802067-b.tif"/>
  <mets:fptr FILEID=" nlaImageSeq-2802067-b.xml"/>
</mets:div>
.....
.....
  <mets:div ID="divpage26" TYPE="page" ORDER="2">
    <mets:fptr FILEID=" nlaImageSeq-2802068-b.tif"/>
    <mets:fptr FILEID=" nlaImageSeq-2802068-b.xml"/>
  </mets:div>
</mets:div>
</mets:div>
</mets:div>
</mets:structMap>

```

Page Divisions

<u>Element Name:</u> <mets:div TYPE="issue"> <mets:div ID="divpage#" TYPE="page" ORDER="#">	
<u>Description:</u> Subordinate <div> elements correspond to each page.	
<u>Occurrence:</u> mandatory, repeatable	
<u>Attributes:</u>	
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.
LABEL	Optional attribute which may be used to provide further information about the page if required, e.g. "duplicate page", "front cover", or when the printed page number is non-numeric or otherwise different from the ORDER number e.g. "S1"

Page File Pointers

<u>Element Name:</u> <mets:div TYPE="issue"> <mets:div TYPE="page" ORDER="#"> <mets:fptr FILEID="imagepage#" />

<code><mets:fptr FILEID="altopage#" /></code>	
<u>Description</u> : The <fptr> elements have “FILEID” attributes indicating the files identified in the File Group elements for each page.	
<u>Occurrence</u> : mandatory, repeatable	
<u>Attributes</u> :	
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 for the page on which the article starts. The structural hierarchy of the pages may be supplied by the Library as source metadata records or may be determined by the contractor during processing, depending on the work order specification.

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.

NOTE: Positional coordinates must be expressed in pixels.

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"> <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

In the following example there were two sets of pages published on the same date: pages 1-24 (image sequence 2802043 to 2802066) in an unnamed main edition and pages 1-2 (image sequence 2802067-2802068) in a "Special Edition". There are separate <div>s for each edition:

```
<mets:structMap ID="structmap2" TYPE="logical">
  <mets:div TYPE="issue" DMDID="issue-nla.news- issn03126323_19160603">
    <mets:div TYPE="edition" DMDID="modsedition1" >
      <mets:div ID="divarticle1" TYPE="article" DMDID="modsarticle1">
        <mets:div ID="divarticle1-1" TYPE="article-part" ORDER="1">
          <mets:fptr>
            <mets:area FILEID=" nlaImageSeq-2802043-b.tif"
SHAPE="RECT" COORDS="297,701,1059,4373"/>
          </mets:fptr>
          <mets:fptr>
            <mets:area FILEID=" nlaImageSeq-2802043-b.xml"
BETYPE="IDREF" BEGIN="ART1"/>
          </mets:fptr>
          <mets:div ID="artzone1-1" TYPE="article-zone">
            <mets:fptr>
              <mets:area FILEID=" nlaImageSeq-2802043-b.tif"
SHAPE="RECT" COORDS="297,701,1059,4373"/>
            </mets:fptr>
          </mets:div>
        </mets:div>
      </mets:div>
    </mets:div>
  </mets:structMap>
```


Article Divisions

<u>Element Name:</u> <mets:div ID="divarticle#" TYPE="article" DMDID="modsarticle#">	
<u>Description:</u> 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.	
<u>Occurrence:</u> mandatory, repeatable	
<u>Attributes:</u>	
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

<u>Element Name:</u> <mets:div ID="divarticle#-#" TYPE="article-part" ORDER=" #">	
<u>Description:</u> 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.	
<u>Occurrence:</u> mandatory, repeatable	
<u>Attributes:</u>	
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

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

Article Part File Areas

<u>Element Name:</u> <mets:area FILEID=”page image filename” SHAPE=”RECT” COORDS=” x1,y1,x2,y2”/> <mets:area FILEID=”page OCR filename” BETYPE=”IDREF” BEGIN=” ART#”/>	
<u>Description:</u> 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. 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.	
<u>Occurrence:</u> mandatory, repeatable	
<u>Attributes:</u>	
FILEID	<u>Page image file:</u> Contains page image filename <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"> <mets:div ID="artzone#-#" TYPE="article-zone" > <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

<u>Element Name:</u> <mets:area FILEID="page image filename" SHAPE="RECT" COORDS=" x1,y1,x2,y2"/> <mets:area FILEID="page OCR filename" BETYPE="IDREF" BEGIN=" ZONE#-#" />

<p><u>Description:</u> The first instance of this element is contained within a <mets:fptr> 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 <mets:fptr> element and provides a reference to the block ID of the zone-level text in the ALTO file.</p>	
<p><u>Occurrence:</u> mandatory, repeatable</p>	
<p><u>Attributes:</u></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.

Describing pages, which do not have a corresponding ALTO file, in the METS structural map:

1. **Missing issue**
2. **Missing page**
3. **Technical target**
4. **Blank page**
5. **Duplicate page**
6. **Other**

All supplied image files should be represented in METS. If an image is not OCR'd and does not have a corresponding xml file, a specific LABEL attribute must be included so NLA's system knows how to deal with it and can reconcile the number of images with the number of expected ALTO xml files.

Note that the ingest process will check all file locations listed in the METS document and if the expected files are not found the ingest will fail. Therefore all file locations must be accurate. (For example the system must know where to find technical targets, in order to suppress them and replace them with targets with the contributor's name and logo.)

1. Missing issue

Each missing issue will have its own METS file. There will be only a single occurrence of <mets:dmdSec> describing the issue level. Since there are no editions, supplements, sections or articles there will be no other occurrences of <mets:dmdSec>.

1.1 where the missing issue is represented by a technical target on the microfilm. The technical target will be suppressed from the delivery system and will be replaced by a target image with name and logo of the contributing library. There should be no xml file for the technical target. The LABEL is "missing issue target". There is no div element at the page level. (However the target generated by NLA will be given a page sequence number of "1" in the delivery system.)

The structural map expected is:

```
-<mets:structMap ID="structmap1" TYPE="physical">
--<mets:div TYPE="issue" DMDID="id_of_dmdsec_element" LABEL="missing
issue target">
---<mets:fptr FILEID="filenameprefix-snumber-b.tif"/>
```

Note that where a technical target lists multiple missing issues there should be a METS file for each missing issue listed. Each of these METS files will point to the same image (which is the technical target) i.e. they should have the same filename in FILEID. The ingest system will create a target image for each missing issue.

1.2 a missing issue which is not represented by a technical target is not expected to be identified by OCR contractors. However if contributors or contractors do wish to identify and report missing issues which are not represented by a file, they must use a blank file location (<fileSec> is mandatory in METS so the file reference cannot simply be omitted).

Example of fileSec and structMap for a missing issue with no technical target:

```
<mets:fileSec>
<mets:fileGrp USE="IMAGEpage">
<mets:file ID="anything" MIMETYPE="image/tif">
<mets:FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="" />
</mets:file>
</mets:fileGrp>
</mets:fileSec>
<mets:structMap ID="structmap1" TYPE="physical">
<mets:div TYPE="issue" DMDID="id_of_dmdsec_element" LABEL="missing
issue">
<mets:fptr FILEID="anything"/>
</mets:div>
</mets:structMap>
```

2. Missing page

2.1 where the missing page is represented by a technical target. The technical target will be suppressed from the delivery system and will be replaced by a target image with name and logo of the contributing library. There should be no xml file for the technical target. The missing page has a page sequence number in the ORDER attribute.

Example: Extract from METS structural map for an issue which is missing page 1

```

-<mets:structMap ID ="structmap1"    TYPE ="physical">
--<mets:div TYPE ="issue"    DMDID ="id_of_dmdsec_element">
---<mets:div TYPE ="page"    ID =" whatever "    ORDER ="1"
LABEL="missing page target">
----<mets:fptr    FILEID ="filenameprefix-snumber-b.tif" />
---<mets:div TYPE ="page"    ID =" whatever "    ORDER ="2">
----<mets:fptr    FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr    FILEID ="filenameprefix-snumber-b.xml" />

```

2.2 where the missing page is not represented by a technical target but was found to be missing (by the OCR contractor for example). A target image with the name and logo of the contributing library will be generated for the delivery system.

Similar to 2.1 but without the fptr element for the missing page and with a different LABEL i.e. "missing page"

Example: Extract from METS structural map for an issue which is missing page 1

```

-<mets:structMap ID ="structmap1"    TYPE ="physical">
--<mets:div TYPE ="issue"    DMDID ="id_of_dmdsec_element">
---<mets:div TYPE ="page"    ID =" whatever "    ORDER ="1"
LABEL="missing page">
---<mets:div TYPE ="page"    ID =" whatever "    ORDER ="2">
----<mets:fptr    FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr    FILEID ="filenameprefix-snumber-b.xml" />

```

3. Technical target

where the scanned image is of a technical target on the microfilm, the image is suppressed from the delivery system. There should be no xml file for the image. The structural map looks like this:

```

---<mets:div TYPE ="page"    ID ="whatever"    ORDER ="0"
LABEL="technical target">
---- <mets:fptr    FILEID ="filenameprefix-snumber-b.tif" />

```

4. Blank page

4.1. where the blank page is not a newspaper page but a piece of paper inserted into the microfilming for some reason, commonly used to signify start / end of a reel, this is treated in the same way as a microfilm target. The image is suppressed from the delivery system and there should be no xml file for the image.

```

---<mets:div TYPE ="page"    ID ="whatever"    ORDER ="0"
LABEL="technical target">
---- <mets:fptr    FILEID ="filenameprefix-snumber-b.tif" />

```

4.2 where the blank page is a blank page that was actually printed in the newspaper. These pages usually have the number printed at the bottom and are clearly a real newspaper page rather than a blank piece of paper. The blank page image will be delivered as is. The blank page has a page sequence number and the LABEL is "blank page". There should be no xml file for the page.

Example: Extract from METS structural map for an issue where page 1 in the printed newspaper was blank.

```

-<mets:structMap ID ="structmap1" TYPE ="physical">
--<mets:div TYPE ="issue" DMDID ="id_of_dmdsec_element">
---<mets:div TYPE ="page" ID ="whatever" ORDER ="1" LABEL="blank
page">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
---<mets:div TYPE ="page" ID ="whatever" ORDER ="2">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr FILEID ="filenameprefix-snumber-b.xml" />

```

Note: If there were a target on the microfilm indicating that the following page was blank in addition to the blank page itself, then the target would be treated as a technical target and the blank page treated as in 4.2.

5. Duplicate page

where duplicate pages are identified only one of the pages (the 'preferred' duplicate) should be OCR'd and have an xml file (and no LABEL attribute). Non-preferred duplicates should not have an xml file and should have LABEL= "duplicate page". The ORDER attribute may contain the page number or may contain "0" if the page number was not recorded.

Example: Extract from METS structural map for an issue where page 1 had two duplicates.

```

-<mets:structMap ID ="structmap1" TYPE ="physical">
--<mets:div TYPE ="issue" DMDID ="id_of_dmdsec_element">
---<mets:div TYPE ="page" ID ="whatever" ORDER ="1" >
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr FILEID ="filenameprefix-snumber-b.xml" />
---<mets:div TYPE ="page" ID ="whatever" ORDER ="1"
LABEL="duplicate page">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
---<mets:div TYPE ="page" ID ="whatever" ORDER ="1"
LABEL="duplicate page">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
---<mets:div TYPE ="page" ID =" whatever " ORDER ="2">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr FILEID ="filenameprefix-snumber-b.xml" />

```

Example: Extract from METS structural map for an issue where there were two duplicate pages whose page numbers were not recorded.

```

-<mets:structMap ID ="structmap1" TYPE ="physical">
--<mets:div TYPE ="issue" DMDID ="id_of_dmdsec_element">
---<mets:div TYPE ="page" ID ="whatever" ORDER ="1" >
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr FILEID ="filenameprefix-snumber-b.xml" />
---<mets:div TYPE ="page" ID =" whatever " ORDER ="2">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
----<mets:fptr FILEID ="filenameprefix-snumber-b.xml" />
---<mets:div TYPE ="page" ID ="whatever" ORDER ="0"
LABEL="duplicate page">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
---<mets:div TYPE ="page" ID ="whatever" ORDER ="0"
LABEL="duplicate page">
----<mets:fptr FILEID ="filenameprefix-snumber-b.tif" />

```

6. Any other image that is not OCR'd for whatever reason

Where an image does not have a corresponding xml file because the image was not OCR'd for any other reason, if none of the above cases apply then order should be "0" and LABEL should be "other". Contractors should discuss cases which fall into this category with NLA before proceeding if possible.

The image will be suppressed from the delivery system and no target will be generated

```
---<mets:div TYPE ="page" ID ="whatever" ORDER ="0" LABEL="other">  
---- <mets:fptr FILEID ="filenameprefix-snumber-b.tif" />
```

APPENDIX 2: ALTO Specification

File name of alto file must have the same name as the image file used for ocr except for the .xml extension

Note: Specification and ingest system were developed for the versions current at the time:

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

Later versions of ALTO may be acceptable. Contractors provide a pilot sample for testing before proceeding to production.

A sample ALTO file can be found linked to the ANDP Project Details page at http://www.nla.gov.au/ndp/project_details/

Positional coordinates (HPOS, VPOS, WIDTH, HEIGHT) must be expressed in pixels.

Schemas and default namespaces: The ingest system expects there to be either a no namespace schema against which the file may be validated or it expects a default namespace. (The default namespace and no namespace schema can't co-exist.)

Example 1. with a default namespace (**in bold**), and without a no namespace schema. The ALTO file won't be validated.

```
<alto xmlns="http://schema.ccs-gmbh.com/ALTO"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://schema.ccs-gmbh.com/ALTO alto.xsd">
```

Example 2. without a default namespace, but with a no namespace schema (**in bold**). The ALTO file will be validated.

```
<alto xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://schema.ccs-gmbh.com/metae/alto-1-4.xsd"
xmlns:xlink="http://www.w3.org/1999/xlink">
```

<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 "pixel"

<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, “ART#” for article-level blocks “ZONE#-#” for article-zone level blocks ”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
STYLEREF:	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.