

Australian Newspaper Digitisation Program (ANDP), National Library of Australia (NLA)

Specification of Directory and File requirements for OCR deliverables where issue dates and page numbers are supplied by the OCR contractor.

April 2010

The deliverables must be supplied in accordance with the requirements below so that they can be ingested automatically by ANDP.

FTP

- The deliverables for each Batch must be supplied by transferring a ZIP file and a ZIP file checksum file to NLA's FTP server. TAR files are also acceptable. The Library will provide access details to the OCR contractor.

ZIP file

- The file name must be of the form "[Batch id]R[round number].zip" .
- The Batch id must be the number assigned by ANDP .
- The Round number is 1 the first time the Batch is delivered. If a Batch is re-supplied, e.g. because it was Rejected the previous time, the Round number must be incremented by 1.
- The number in the file name may include leading zeroes e.g. "1108R2.zip" and "01108R2.zip" are acceptable file names for the second time Batch 1108 is delivered.
- The file name may include a prefix followed by a hyphen e.g. "xxx-1108R2.zip" as long as the prefix contains no other hyphens.
- When the ZIP file is unzipped, there must be a single (root) directory which must have the same name as the ZIP file without the .zip extension e.g. "1108R2" for the file "1108R2.zip"

ZIP file checksum file

- Contractors must upload a text file AFTER the completion of the ZIP file upload to NLA's FTP server. Presence of this file will signal the status of "upload complete" to the auto-ingest process polling the server.
- The file must contain only the checksum (MD5 or SHA1 algorithm) of the ZIP file.
- The file name must consist of the ZIP file name and the extension ".md5" or ".sha1" according to the algorithm used e.g. "1108R2.zip.sha1"

Contents of ZIP file

When the ZIP file is unzipped, there must be a single (root) directory.

Directory	Requirements	Comments
Root directory	<ul style="list-style-type: none">• The root directory must have the same name as the ZIP file without the .zip extension e.g. "1108R2" for	

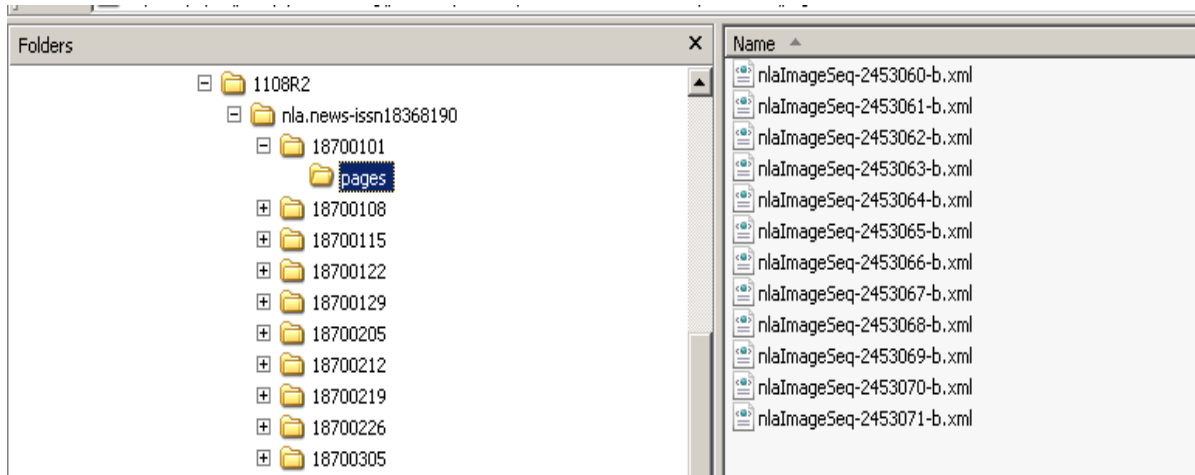
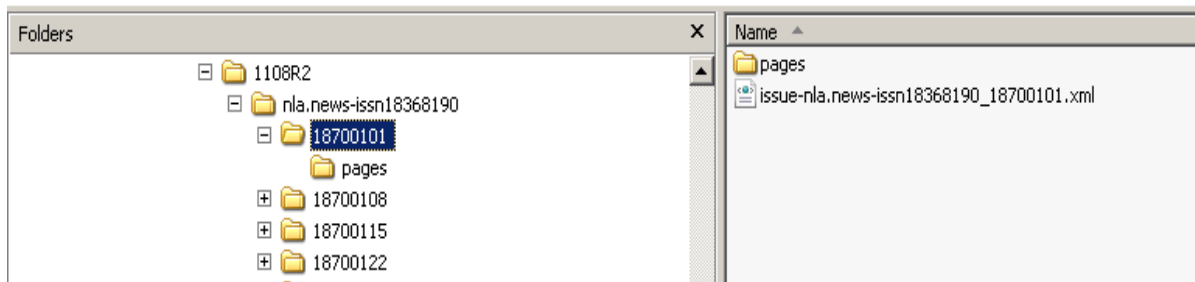
	<p>the ZIP file "1108R2.zip"</p> <ul style="list-style-type: none"> • Must contain a file named "check.csv" • May contain any number of directories but must contain at least one. 	
(Optional) Any number of intermediate directory levels	None	ANDP system does not use. Name and contents of directories ignored
A directory for each issue-date	<ul style="list-style-type: none"> • Must contain a METS file whose name begins with "issue-" and ends in ".xml" • Must contain a single subdirectory for page-level ALTO files. • Directory name may be anything. 	ANDP system identifies directories by looking in the check.csv file for directories <i>containing</i> files whose names start with "issue-" and end in ".xml". Therefore the <i>name</i> of the directory is unimportant.
A directory for page-level ALTO files (for each issue-date)	<ul style="list-style-type: none"> • Must contain an ALTO xml file for each OCR'd page in the issue. • Directory name may be anything. 	ANDP system looks for this directory <i>directly under</i> the issue-date directory. Therefore the name of the directory is unimportant.

Check file

- File name must be "check.csv" (all lower case).
- File is a manifest of files in the ZIP file.
- Each file is listed with the full path (relative to the root directory where check.csv resides) and file name, checksum type (MD5 or SHA1) and checksum. The file path must contain forward slashes (/), not back slashes, because the ANDP system is a Unix system.

Example:





When the file "1108R2.zip" is unzipped, it creates the following structure:

```

1108R2
- nla.news-issn18368190
- check.csv
- - 18700101
- - - pages
- - - - issue-nla.news-issn18368190_18700101.xml
- - - - nlaImageSeq-2453060-b.xml
- - - - nlaImageSeq-2453061-b.xml
- - - - nlaImageSeq-2453062-b.xml

```

Example of part of check file for the above three image files:

```

"nla.news-issn18368190/18700101/pages/nlaImageSeq-2453060-
b.xml",SHA1,8f2d58649dd32f57518352297c3aef823e1ae29e

"nla.news-issn18368190/18700101/pages/nlaImageSeq-2453061-
b.xml",SHA1,5328a1a07636c51c8e895602a3fde0d59c5bd831

"nla.news-issn18368190/18700101/pages/nlaImageSeq-2453062-
b.xml",SHA1,6fe34151c74c1215a10ab2a038a1b6270f80c415

```

Note: The double-quotes are optional as there are no commas within the full path and file name.

METS file

- File name must begin with "issue-" and end in ".xml" e.g. "issue-nla.news-issn18368190_18700101.xml"
- Requirements are in the document *Use of METS and ALTO in the Australian Newspapers Digitisation Program (ANDP) at the National Library of Australia (NLA)* which is available at http://www.nla.gov.au/ndp/project_details . As this document may be revised periodically, please ensure you are viewing the most recent version.
- The above document includes requirements for dealing with cases where a page image has no corresponding ALTO file, usually because the image was not OCR'd. This is important because the ANDP ingest system reconciles the number of ALTO files received with the number it is expecting, and if the number is different, the ingest will fail.

ALTO file

- File name must be the same as the file name of the page image used for OCR except it must have ".xml" as the file name extension instead of the image file name extension e.g. if the image file is named "xxx-b.tif", the corresponding ALTO xml file must be named "xxx-b.xml".
- Requirements are in the document at *Use of METS and ALTO in the Australian Newspapers Digitisation Program (ANDP) at the National Library of Australia (NLA)* which is available at http://www.nla.gov.au/ndp/project_details . As this document may be revised periodically, please ensure you are viewing the most recent version.
- The article, article-part and article-zone IDs must match the values in the BEGIN attributes in the METS file, and the co-ordinates must also match those given in the METS file. (Coordinates in ALTO files are used for word highlighting, whereas coordinates in METS files are for article highlighting.)