

National Library of Australia Persistent Identifier (PI) Model for Newspapers

Author: Bronwyn Lee (Business Analyst – ANDP)

Version: 1.0

Date: 27 September 2007

1. Background

The Australian Newspaper Digitisation Program (NDP) commenced on 1st March 2007. Initially a pilot is being undertaken to digitise 50,000 pages by September 2007. The next steps will see an additional 500,000 pages digitised by mid- 2008 and 3 million pages digitised in 4 years.

A persistent identifier is a name for a resource which will remain the same regardless of where the resource is located. Therefore links to the resource will continue to work even if it is moved. The National Library has implemented its own managed naming scheme for digital resources and coupled this with an internal resolver service that seamlessly directs a request for an item to the current storage location of that item. The National Library's naming scheme is published at <http://www.nla.gov.au/initiatives/nlapi.html> .

In August 2007 the Newspaper Digitisation development team recommended that PIs for newspapers depart significantly from the existing model. They particularly recommended a move away from meaningful PIs based on descriptive metadata. This was because metadata such as ISSN, issue date and page sequence number, can change, for instance when inaccuracies or missing pages are found.

A new principle was agreed that the PI should not be based on metadata. Instead numbers will be used which have no meaning, no fixed length and no leading zeros. However it was agreed that PIs can have structure which facilitates repository or delivery system management e.g. collection prefixes may be retained and indicators may be used for classes of objects which would be treated in different ways by a delivery system (eg page, article etc).

Another principle agreed was that particular renditions of pages, articles and illustrations e.g. thumbnails, would not have their own PIs with copy-roles, as they do in the existing scheme and the current delivery system, but could be requested by a URL (at least until there is a standard protocol for requesting certain behaviours) containing the PI of the requested item.

2. PIs and the newspaper delivery system

In the National Library's current delivery systems the URL <http://nla.gov.au/> followed by the PI resolves to a default 'contextual display' eg <http://nla.gov.au/nla.pic-vn3579778> displays a 'view' copy of a picture with descriptive metadata from the catalogue record and links to other items in the set.

The default contextual displays for newspapers will probably include:

Title:

- metadata from the Newspaper Digitisation management system such as ISSN, title name, notes

- metadata from catalogue
- list or links to issues available

Issue:

- metadata from Newspaper Digitisation QA system: ISSN, date, notes
- list or links to editions, supplements, sections, pages and articles available
- possibly image/s of one or more pages

Page:

- metadata from Newspaper Digitisation QA system: ISSN, date, edition, supplement, section, page number, notes etc
- page-image derivative/s (derived from greyscale TIFF file)
- list or links to articles available

Article:

- metadata from Newspaper Digitisation QA system about page/s the article occurs on: ISSN, date, edition, supplement, section, page number/s, notes etc
- metadata provided by Apex CoVantage (derived from a METS XML file) about the article: title, subtitle, author, abstract, notes?
- article-image/s (derived, probably dynamically, from the page-image derivatives using coordinates from the ALTO XML file (provided by Apex) which in turn is derived from bitonal TIFF page-image)
- link to article-text (derived from ALTO XML file)
- possibly link to illustrations if these are delivered separately in addition to within an article

Illustration:

- metadata from provided by Apex about the article the illustration occurs in (including page and article metadata).
- metadata about the illustration i.e. the caption if any
- illustration-image (derived, probably dynamically, from the page-image derivatives using coordinates from the ALTO XML file)

Therefore PIs are needed in the delivery system for Title, Issue, Page, Article and Illustration.

This document does not at this stage address PIs at the file level in the preservation repository.

3. PIs and the resolver service

The following forms of PI are proposed: nla.news-titlen, nla.news-issuen, nla.news-pagen, nla.news-articlen, nla.news-illustrationn where n is a number with no meaning, no fixed length and no leading zeros.

PIs for Titles

For the number the Newspaper Digitisation QA unique system id for the title will be used (A title is defined by an ISSN). The issues that a title PI resolves to may change as issues are added or deleted.

PIs for Issues

For the number the Newspaper Digitisation QA unique system id for the issue will be used. (An issue is defined by an ISSN and a date.) The pages and articles that an issue PI resolves to may change if the dates in the pages' metadata changes. It's highly improbable an issue will be left with no pages after it is published but if it did a note would be put in the issue's metadata which would explain where the pages went.

PIs for Pages

For the number the Newspaper Digitisation QA unique system id for the page image (called the image sequence number) will be used. A page is defined as the page image; a page is NOT defined by its metadata (ISSN/date/page number, which can change). The page image is the bitonal provided by Pascoe's since this is what is QA'd and OCR'd. It is assumed that the greyscale is the same content unless it doesn't match the bitonal when run through an inhouse program called MatchTIF. A page PI would be associated with the various files that are derived from the page image: bitonal, greyscale, derivatives of these, the ALTO XML file.

Pages with the same metadata ('duplicate' pages) have different image sequence numbers. However non-preferred duplicates are not OCR'd, derivatives are not created for them and they will not be delivered; hence non-preferred duplicate pages will not need a PI.

If a frame on a microfilm is re-scanned, the image will be treated as a new image and given a new image sequence number and a new page PI. If it replaces (as the preferred duplicate) another image which has already been published in the delivery system, a record of the old image is kept so that if a user requests the old image's PI, the user can be redirected to the new image which has the same page metadata.

PIs for Articles

What constitutes an article depends on how Apex have drawn the article boundaries. If a page is re-zoned and re-OCR'd for any reason, the boundaries could completely change. If a page is re-OCR'd and all the articles replaced, a record would be retained of the old (deleted) article PIs. These PIs would resolve to a note saying that the article has been removed but that it appeared on such-and-such a page with a link to the page (or the first page) the article appeared on. If the page image the old article was on has also been replaced, the delivery system would need to check the deleted-PI table to find the old page's metadata and re-direct to a new page with the same metadata.

For the number system ID for the article will be used.

PIs for Illustrations

An illustration (which includes the caption if any) can be part of an article or can constitute the whole article; however the same comments about boundaries apply as for articles, and an illustration will not necessarily belong to the same article if the page is re-zoned. A strategy similar to that for articles would be employed if a page was re-zoned and re-OCR'd.

For the number the system ID for the illustration will be used.

4. The form of newspaper PIs

Newspaper PIs will consist of:

- Collection prefix :nla.news-
- Indicator for class of object: title, issue, page, article or illustration
- Number which is the unique system ID of the title, issue, page image, article or illustration. (System ID is a meaningless running number with no fixed length and no leading zeros.)

	Example	Example PI
Title	The Argus (ISSN 18339719)	nla.news-title13
Issue	The Argus, 1920-01-01	nla.news-issue48003
Page	Page 1 of above issue	nla.news-page406561
Article	None yet	nla.news-article1234567
Illustration	None yet	nla.news-illustration1234

5. Particular 'rendering' of pages, article and illustrations

Particular renditions of pages, articles and illustrations will not have their own PIs but may be requested by a URL which will take the form:

- PI followed by a slash and a string (which may be meaningful) for the particular rendition.

Examples:

<http://nla.gov.au/nla.news-page1234567> resolves to default contextual display

<http://nla.gov.au/nla.news-page1234567/view> resolves to the image in the default display, without the context

<http://nla.gov.au/nla.news-page1234567/print> resolves to the PDF copy of the page

<http://nla.gov.au/nla.news-page1234567/thumbnail> resolves to the thumbnail

<http://nla.gov.au/nla.news-page1234567/metadata> resolves to xml format descriptive metadata about identified page

(the following might be used for example to publish a page image in an online exhibition):

<http://nla.gov.au/nla.news-page1234567/segment?zoomLevel=4&x=3&y=2> resolves to the tile at X-coord 3, Y-coord 2, at a zoom level of 4. Minimum zoomLevel is 1 and the maximum depends on what is available. In the Australian Newspapers Digitisation Program six zoom levels are proposed; zoomLevel=1 will be 5% and zoomLevel=6 will be 100%.

<http://nla.gov.au/nla.news-page1234567/view?gd=200> resolves to a resized version of the identified image with a greatest dimension of 200 pixels.