THE AUSTRALIAN NEWSPAPERS DIGITISATION PROGRAM:
Helping communities access and explore their newspaper heritage.

Keynote speech given by Rose Holley at the Australian Media Traditions Conference, Charles Sturt University, Bathurst, 23 November 2007.

Rose Holley – Manager Newspaper Digitisation Program
National Library of Australia, Canberra, Australia
rholley@nla.gov.au
02 62621224

ABSTRACT

The paper outlines the work achieved on the National Library of Australia's newspaper digitisation program to date (http://www.nla.gov.au/ndp). It gives an overview of the processes, methods and technologies that are being utilised in the digitisation process and illustrates with screenshots the development of appropriate infrastructure and software to support the program. Software has been developed to support the workflow, content management and delivery of data. Systems infrastructure has been developed within the context of the Library's strategic priorities, so that rapid and easy access to both the Library's collections and other resources can be achieved in a single business model. The service also supports a key objective of the Australian Newspaper Plan (ANPLAN http://www.nla.gov.au/anplan/) "that communities should be able to explore their rich newspaper heritage". This program will greatly improve access for all Australian's to historical newspapers and will give users the ability to rapidly and easily search across the newspapers in a freely publicly accessible system.

INTRODUCTION

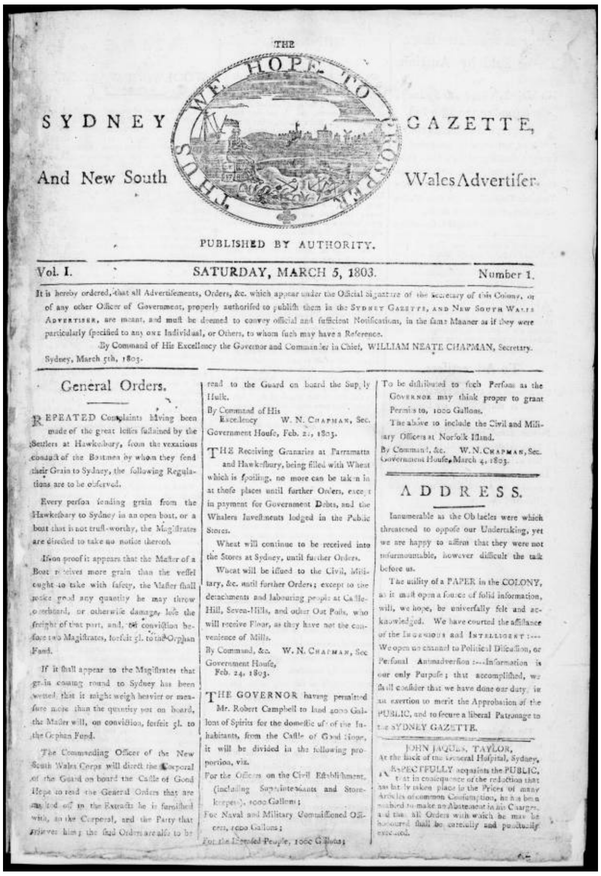
The Australian Newspapers Digitisation Program (ANDP) is of national significance for all Australians and will revolutionise access to newspapers by delivering a single point of access for all Australian newspapers which will be full text searchable and freely available via the internet.

Late last year on 29 November 2006 the Minister for Arts and Sports granted approval for the National Library of Australia to proceed with a National Newspaper Digitisation Program. Prior to this a substantial amount of scoping work had been undertaken to test the viability of such a project, likely cost, and to identify suitable contractors for the digitisation processes.

A budget of $8 million was allocated and approved to digitise 3 million pages over the first four years. Once approval had been gained the Library quickly moved forward and in March 2007 signed up a digitisation contractor who had been sourced through a Request For Tender (RFT) process. In April 2007 I commenced as the Newspaper Digitisation Program Manager.

BACKGROUND AND SCOPE

The newspaper digitisation program is a national effort and newspaper titles from each State and Territory will be included. Initially through collaboration with Australian State and Territory libraries, a single major newspaper title from each Australian State/Territory will be selected, digitised and delivered. It is anticipated that local regional titles will be contributed to the program later. The coverage for the program is for papers published between 1803-1954. On Saturday March 5 1803 the first Australian newspaper was published. This is the Sydney Gazette and New South Wales Advertiser. This was the first paper to be digitised. Following this we are working through papers from the next 150 years up to 1954. After 1954 copyright applies so we can only continue to digitise titles in this date range with the permission of the newspaper publishers.



THE SYDNEY GAZETTE, And New South Wales Advertiſer.

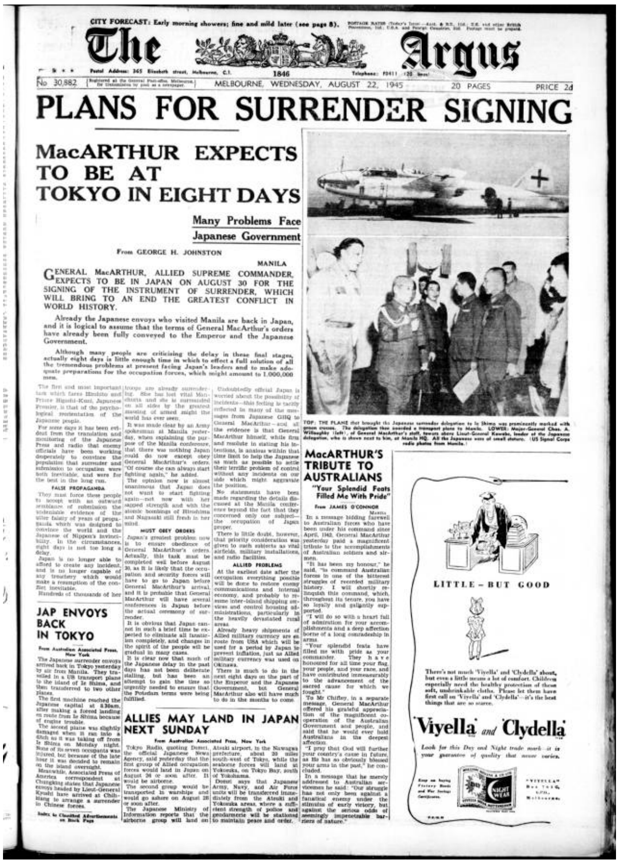
PUBLISHED BY AUTHORITY.

Vol. I. SATURDAY, MARCH 5, 1803. Number 1.

General Orders.

ADDRESS.

Fig 1. Front page of the Sydney Gazette and New South Wales Advertiser, Saturday March 5 1803. Australia's first published newspaper.



The Argus

MELBOURNE, WEDNESDAY, AUGUST 22, 1945 20 PAGES PRICE 2d

PLANS FOR SURRENDER SIGNING

MacARTHUR EXPECTS TO BE AT TOKYO IN EIGHT DAYS

Many Problems Face Japanese Government

JAP ENVOYS BACK IN TOKYO

ALLIES MAY LAND IN JAPAN NEXT SUNDAY

MacARTHUR'S TRIBUTE TO AUSTRALIANS

Viyella and Clydella

Fig 2. Front page of the Melbourne Argus August 22 1945, one of the later papers included in the program.

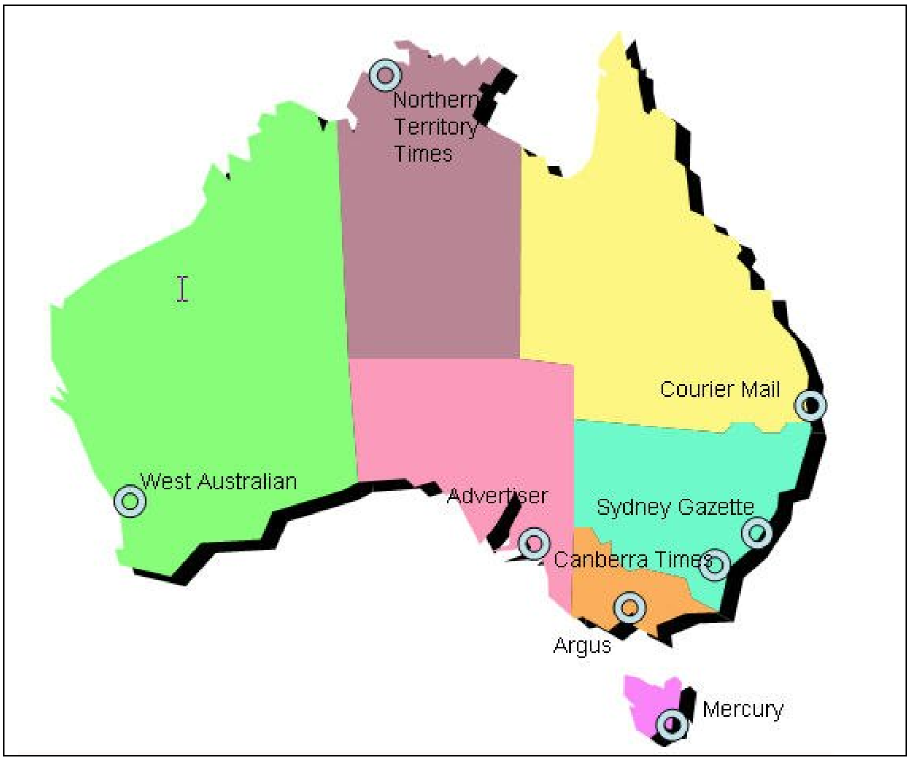
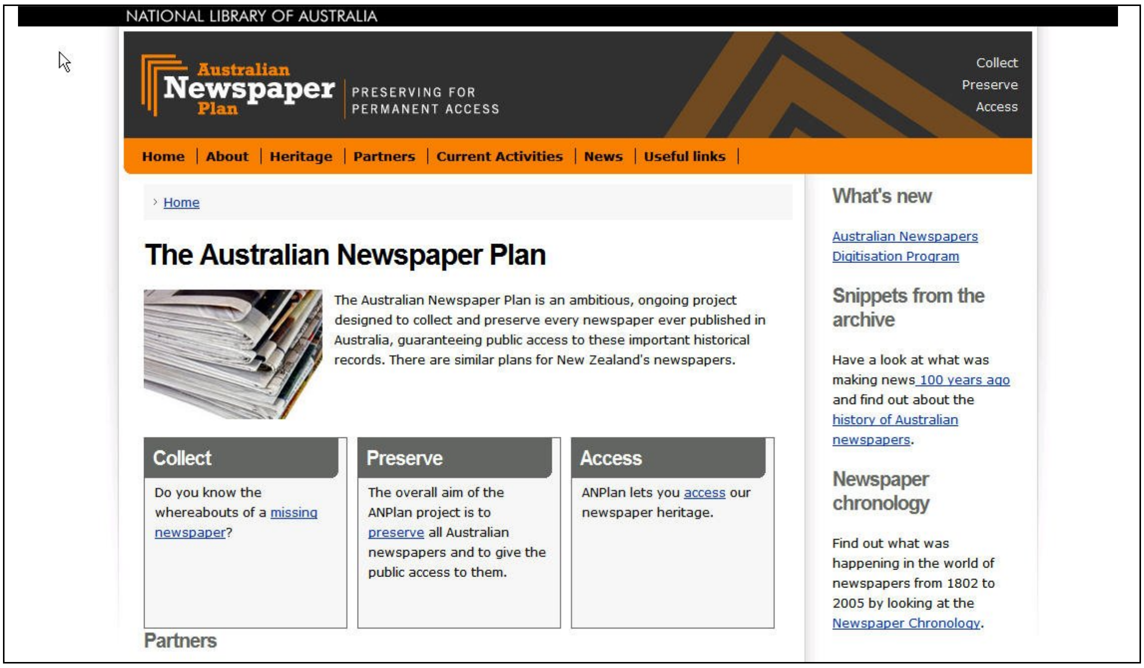


Fig 3. Initial scope and coverage of papers from Australian States and Territories.

The ANDP is being undertaken in conjunction with ANPLAN (The Australian Newspapers Plan). There is a close working relationship between ANPLAN and ANDP. ANPLAN has been in existence for a number of years and its members comprise librarians from State and Territory Libraries. ANPLAN has recently released a new website outlining their objectives. http://www.nla.gov.au/anplan The 3 objectives of ANPLAN are – to collect copies of all Australian newspapers (in conjunction with public help); to preserve copies of all Australian newspapers (by microfilming); and to provide access to Australian newspapers so that all Australians can explore their rich newspaper heritage. It is the third objective which the ANDP seeks to achieve. Rather than Australians being limited to accessing newspapers in Libraries on microfilm readers the ANDP will also enable online access from any location. The microfilm will be digitised and made available via the internet for free.

The website for the ANDP has been built to complement the ANPLAN website, but not duplicate information. It provides information to the public and ANPLAN members about the ANDP. http://www.nla.gov.au/ndp/ The website is being updated regularly to reflect work on the program and progress of digitisation. Under the project detail tab there are also photographs of the entire process.


NATIONAL LIBRARY OF AUSTRALIA
Australian Newspaper Plan
PRESERVING FOR PERMANENT ACCESS
Collect
Preserve
Access
Home | About | Heritage | Partners | Current Activities | News | Useful links |
› Home

The Australian Newspaper Plan

The Australian Newspaper Plan is an ambitious, ongoing project designed to collect and preserve every newspaper ever published in Australia, guaranteeing public access to these important historical records. There are similar plans for New Zealand's newspapers.

Collect

Do you know the whereabouts of a missing newspaper?

Preserve

The overall aim of the ANPlan project is to preserve all Australian newspapers and to give the public access to them.

Access

ANPlan lets you access our newspaper heritage.

Partners

What's new

Australian Newspapers Digitisation Program

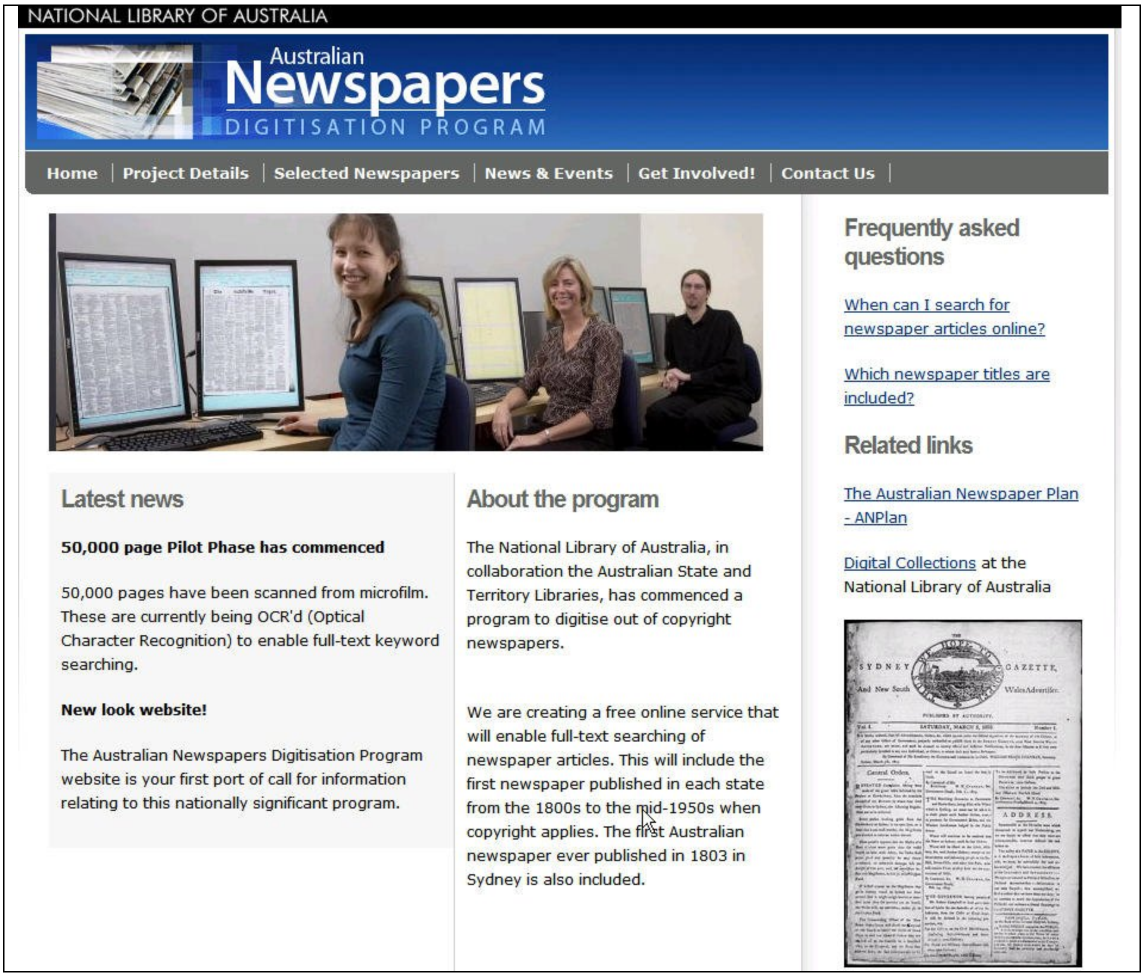
Snippets from the archive

Have a look at what was making news 100 years ago and find out about the history of Australian newspapers.

Newspaper chronology

Find out what was happening in the world of newspapers from 1802 to 2005 by looking at the Newspaper Chronology.

Fig 4. ANPLAN website http://nla.gov.au/anplan


NATIONAL LIBRARY OF AUSTRALIA
Australian Newspapers DIGITISATION PROGRAM
Home | Project Details | Selected Newspapers | News & Events | Get Involved! | Contact Us |

Latest news

50,000 page Pilot Phase has commenced

50,000 pages have been scanned from microfilm. These are currently being OCR'd (Optical Character Recognition) to enable full-text keyword searching.

New look website!

The Australian Newspapers Digitisation Program website is your first port of call for information relating to this nationally significant program.

About the program

The National Library of Australia, in collaboration the Australian State and Territory Libraries, has commenced a program to digitise out of copyright newspapers.

We are creating a free online service that will enable full-text searching of newspaper articles. This will include the first newspaper published in each state from the 1800s to the mid-1950s when copyright applies. The first Australian newspaper ever published in 1803 in Sydney is also included.

Frequently asked questions

When can I search for newspaper articles online?

Which newspaper titles are included?

Related links

The Australian Newspaper Plan - ANPlan

Digital Collections at the National Library of Australia

Fig 5. ANDP website http://nla.gov.au/ndp

From the initial establishment of the program it was intended that this would be a large scale ongoing digitisation program, not a short term project. The scope of the objective is so large that final completion of digitisation has no date. For the initial setting up of workflows, processes, software and contractors the program is in 'project phase', but once all systems are bedded down and a service is launched to the public it is intended that the program will be ongoing long-term. For this reason the ANDP is not being referred to as a 'digitisation project'. The first phase of the program is to digitise 3 million pages over 4 years.

SUMMARY OF WORKFLOW

The National Library of Australia is leading this national program, which is a collaborative effort. This is truly a national program since the Library does not own most of the microfilm needed for digitisation. This is distributed throughout the country at State and Territory Libraries and in addition a large amount is owned and housed at Sydney by W & F Pascoe's Ltd - a microfilm bureau and supplier. State and Territory Libraries, newspaper historians and a microfilm supplier have made recommendations for which titles should be included in the first 3 million pages. We are sourcing the microfilm and sending it to the scanning contractor in Sydney. This same group of stakeholders will be providing us with feedback on the search and delivery prototype which is to be released shortly.

The Library is also developing a model for national collaboration. At present the Library is funding and organizing the digitisation process from start to finish, but the Library is keen to enable participation by the wider community. For example if a public library or local historical society has digitised regional newspapers themselves, or has obtained a grant for a specific title the Library is exploring ways that these titles may be integrated into the program and stored at the Library.

The entire workflow process is very complex. The steps below illustrate a simplified version of the process.

1. Source selected newspaper microfilm masters from around Australia.
2. Outsource the scanning of microfilm to a contractor in Sydney who converts the microfilm into digital images.
3. Take receipt of the image files at the Library and quality assure them (check all the images are there and in the right order) and add page level metadata to each image.
4. Create batches of 2000 images and outsource these to a contractor in USA/India who will perform content analysis (zoning and categorising articles on a page) and OCR on the files (OCR is an automated means of converting an image into a full text searchable resource. This is done by use of specialist software and the files are provided as xml).
5. Take receipt of the files at the Library and quality assure them.
6. Ingest the files into the storage system; create derivative files for loading into the search and delivery system at the Library so that digital newspapers may be made available for public delivery.

The logistics of managing the process across different continents, business cultures and time zones has been challenging. The Library is co-ordinating newspaper sourcing around Australia, the contractor for digitisation of microfilm is based in Sydney, the contractor for OCR has its headquarters, project management and software development team in the USA, but the actual OCR processing work is being carried out at various facilities locations throughout India.

PROGRESS

The program has been going for 6 months now and in this time the following major achievements have been made:

- IT infrastructure and storage implemented at the Library for ANDP
- In-house software development – ingest system, content management system and associated workflow processes.
- Development of in-house quality assurance software and working with contractor on development of external quality assurance software.
- Digitisation of 500,000 microfilm images.
- Identification of 'pilot' data. The pilot data sent to contractors to test workflows, systems and software against agreed project specifications.

Although a considerable amount of scoping had been carried out in the 2 years before the program commenced once we moved from theoretical stage into practical stage a number of previous decisions and assumptions were reviewed and changed in light of changing technology and new information gained about newspapers. Major changes and new developments in digitisation software and hardware had taken place over the 2 years and there were now more efficient and different ways to deal with some aspects of the program. Two of the key assumptions that were incorrect were that all key newspapers would be available on microfilm, and that quality assurance processes would be minimal and not very time consuming.

Tasks ahead for the next 6 months are to:

- Evaluate the pilot data when it is returned from the contractor.
- Review specifications.
- Continue to digitise microfilm into digital images
- Proceed with OCR processing the first 500,000 pages of the 3 million.
- Develop the search and delivery prototype and subsequent public system.
- Public launch of service in 2008 with a good body of content.
- Progressive addition of content – national ongoing.

INFRASTRUCTURE AND TECHNOLOGY

Development of the search and delivery system is being carried out in-house. Although it is the part of the program that users are most interested in the software development has been left until last since the Library needed to establish the mass storage and infrastructure and ingest system first. All the workflow processes needed to be established and sample files obtained from contractors before work could seriously commence on the search and delivery process. A team of programmers at the Library have been working on the software development process.

It is the intent of the Library that a search and delivery prototype will be developed and external feedback gained on this. Then a search and delivery version 1 and 2 will be released to a wider audience including the public before the final service is publicly launched. By using this method it is hoped that when the service is launched it will be good workable product with a good body of data.

The latest technology is being utilized for the program. The Library has examined how other newspaper projects are being achieved and where applicable applied similar methods. The Library has taken a close interest in Google technologies particularly Google maps zoom and navigation technology to see how these may be applied to the service. The Library does not believe that the ANDP will result in a traditional library style database but rather is exploring innovative ways for delivery and user interaction with a full text resource and integration with other relevant collections both internal and external to the Library.

The Library is providing the infrastructure for the nation for the ANDP. Software development is taking place in-house. Key parts are storage system; ingest system, content management and quality assurance system, and search and delivery system. Some parts of the software have been developed in-house because:

- The current in-house content management system is not able to cope with the scalability and nature of newspaper pages.
- It is intended that some parts of the system e.g. Quality assurance module may be beneficial to be open source and provided to contributors to complete parts of the workflow process later.
- No suitable search and delivery system was identified on the market and the Library wanted to take this opportunity to deliver integrated resources using open source software (Lucene).
- The Library wanted to ensure that the newspaper program adhered to digital preservation standards.

The software development is using open-source tools where possible and the Library has learnt from its previous software development experiences. Therefore for newspapers the software will not be irrevocably bolted into in-house systems, but has been developed in modules that could all be supplied as open source product to contributing partners and will operate on a standalone basis.

It is not expected that contributing libraries will need to have their own capacity for storage of files. Indeed most have indicated they definitely do not have the capacity to store any of their own files. The Library is offering a central storage service for newspaper files (preservation masters and derivatives for service) to contributors. The storage required for a program of this scale is considerable. It has taken the Library some time to estimate storage space required and then work through the RFT process and subsequent selection, ordering and implementation of servers and storage arrays to cope with the capacity needed for the ANDP.

Estimates for storage to date are 70 TB for the first 3 million pages. The storage is partitioned. The working space online is 40 TB. The working space contains master files (500,000) that have not yet reached the end of the workflow process. Because the workflow process is a complex cycle files may remain in the working space for several months until all the processing stages are completed. Once processing is completed master TIFF files are moved to offline storage (LTO2 tapes) and derivatives created for delivery. Several different file types are created for delivery and also different file sizes and resolution for use in the zoom feature in the search and delivery system. It was difficult to estimate storage required since until recently we had not made decisions about the number and type of delivery derivatives required and whether there would be 1, 2 or 3 master TIFF images. We were uncertain which version of the TIFF to keep as a master. The original TIFF image had cropping and image manipulation applied to it at various parts of the process, in addition we also had both bi-tonal and greyscale TIFF images since we were unsure which would be the optimal files for OCR processing and also delivery. The size of derivative storage space required for 3 million images is 30 TB. The database itself, indexes and XML files will take up the smallest part of storage at only 1TB. Files are stored in line with the Library’s preservation policy and master files offline are stored in triplicate on and offsite, with online storage having appropriate back-up and disaster recovery plans in place.

The diagram below gives an indication of the complexity of the storage workflow which has been through several iterations to date, and may change in the future.

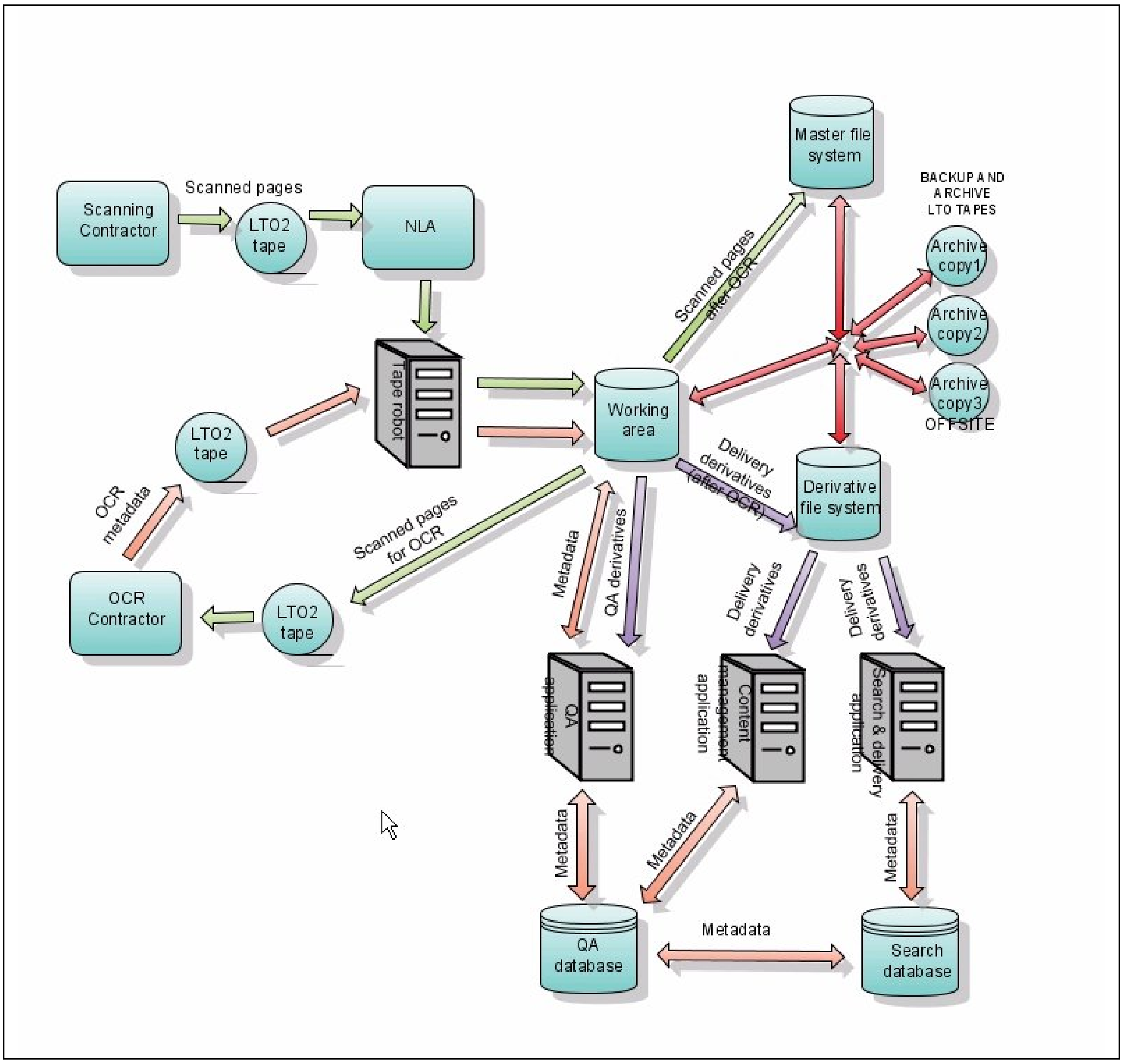


Fig 6. Storage workflow November 2007.

The Library has required that its scanning contractor make recommendations for the best scanning hardware and software on the market that will work efficiently and effectively to meet the Library's needs. The technology selected is the latest scanning software and hardware from NextScan Ltd. This equipment has been developed in the USA in close consultation with the Mormons who currently have about 70 scanners in use to digitise genealogical resources. The Eclipse microfilm scanner with the latest Nextstar software scans and captures a newspaper microfilm reel in about 12 minutes (approx 8000 images). The images are then reviewed and manipulated in the Nextstar audit software. A key feature of this is the page splitter which separates multiple pages on a single microfilm frame into single pages. It is very important for the ANDP that all newspaper pages are separate digital images. Our current scanning outputs are 20,000 images a week but we are hoping to significantly increase this in the immediate future. Many of the microfilm masters are owned by W & F Pascoe's based in Sydney and are not in Library premises.

Fig 7. Master microfilm reels awaiting scanning.

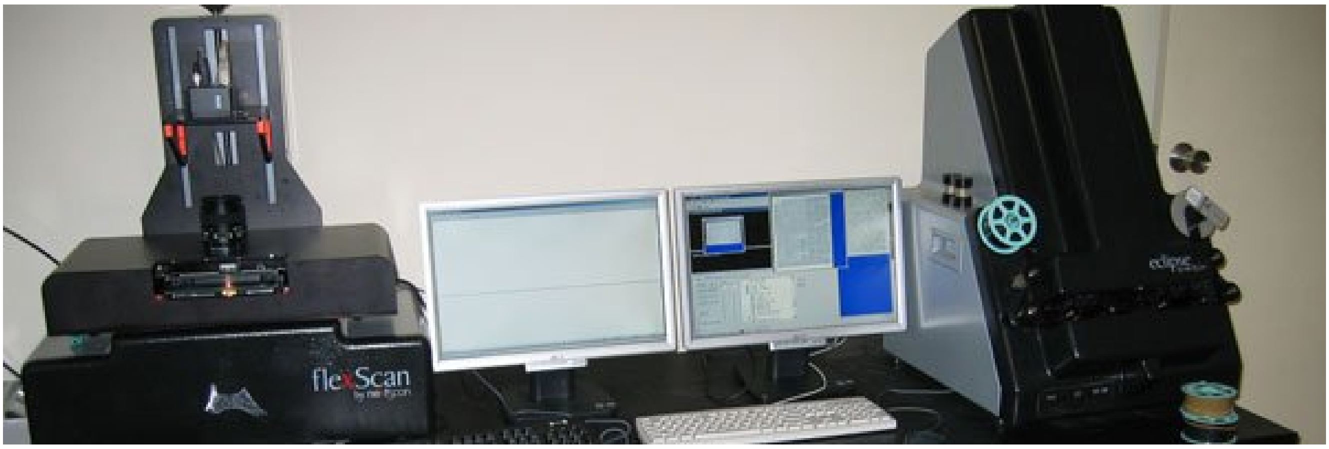


Fig 8. NextScan scanning equipment.

When digital images are returned to the Library on LT02 tape they are loaded into the Library's in-house Quality Assurance system. This is accessed by Library staff using a double monitor setup. We are using 2 widescreen monitors at each workstation and these have been turned vertically on their stands. Widescreen monitors turned the other way up are perfect for viewing newspaper pages and give staff an excellent view of the papers at a large size. It is a particularly useful setup for comparing duplicate pages, or viewing thumbnails of the paper and one screen and individual selected pages on the other screen. We have also trialled the widescreen monitors turned vertically in the newspaper reading room at the Library and received very positive feedback from users.

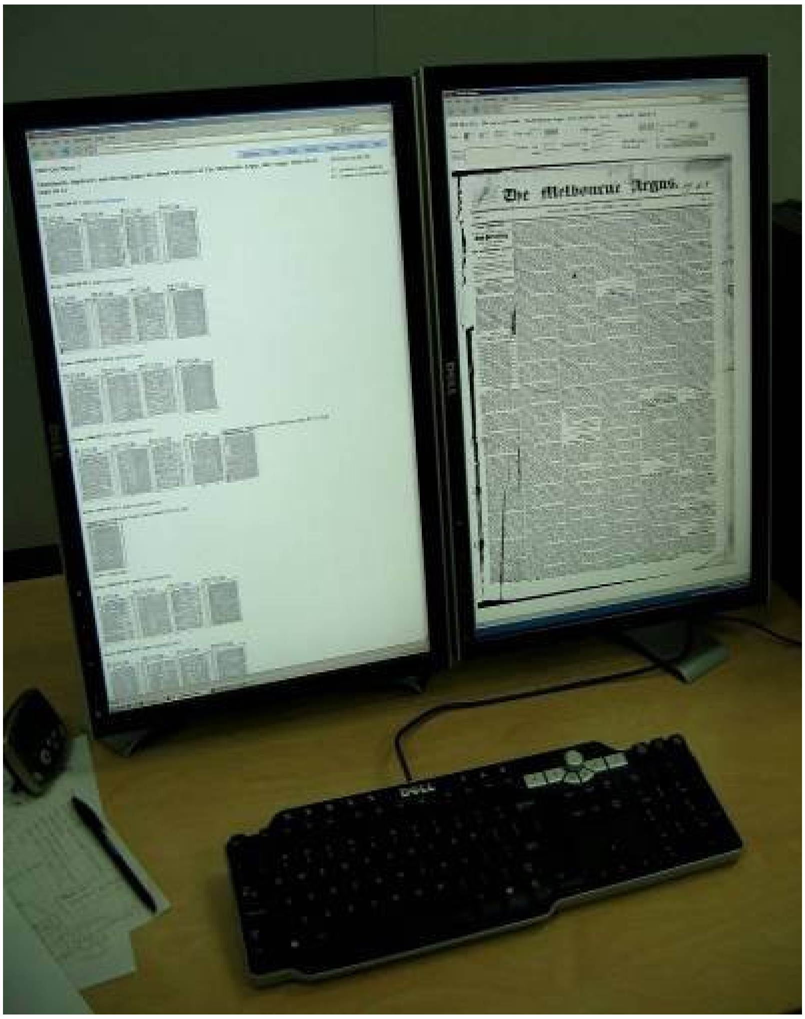
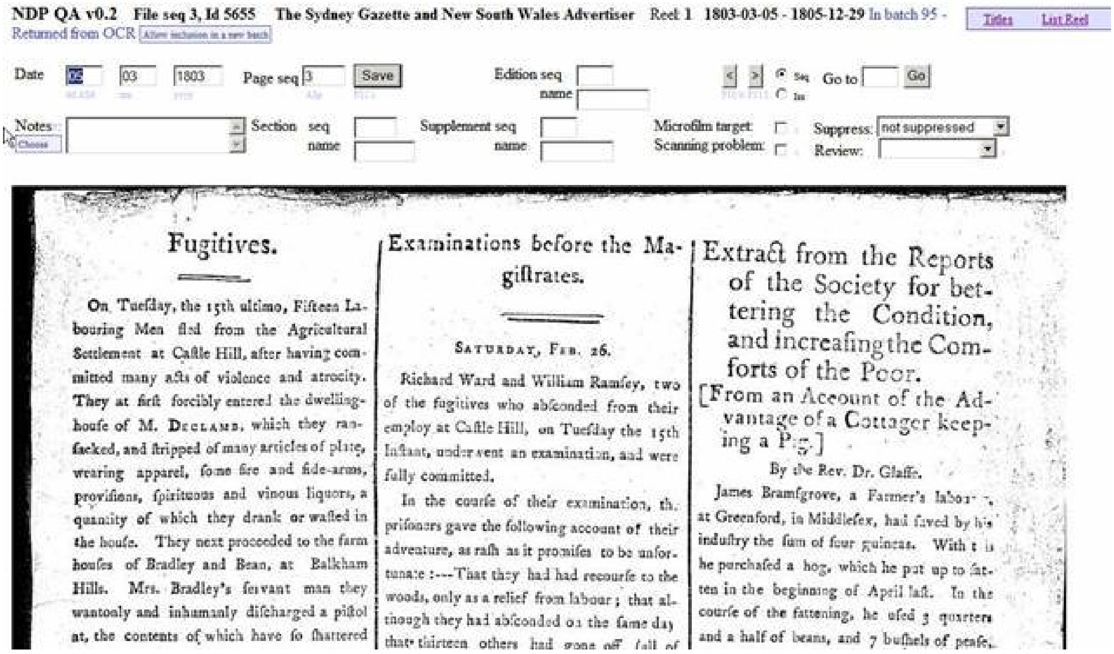


Fig 9. Two widescreen monitors set up vertically for Quality Assurance.

The quality assurance process involves adding metadata at page level – title, volume, issue, page sequence as a minimum. There is also the facility to add supplement information and notes. The most important thing at this stage is page sequencing. Most newspapers don't have a page number showing, or often have an incorrect page number so we give them a virtual page number which is really a sequence number. This is fairly time consuming since often the order of the pages on the microfilm is out of sequence, or many pages are missing which makes it difficult to identify the page sequence of remaining or fragmented pages.



NDP QA v0.2 File seq 3, Id 5655 The Sydney Gazette and New South Wales Advertiser Reel: 1 1803-03-05 - 1805-12-29 In batch 95 - Returned from OCR

Date 05 03 1803 Page seq 3

Edition seq
name

○ Seq ○ Iss Go to

Notes:

Section seq
name

Supplement seq
name

Microfilm target: □
Scanning problem: □

Suppress: not suppressed
Review:

Fugitives.

On Tueſday, the 15th ultimo, Fifteen Labouring Men fled from the Agricultural Settlement at Caſtle Hill, after having committed many acts of violence and atrocity. They at firſt forcibly entered the dwelling-houſe of M. DECLAMB, which they ranſacked, and ſtripped of many articles of plate, wearing apparel, ſome fire and ſide-arms, proviſions, ſpirituous and vinous liquors, a quantity of which they drank or waſted in the houſe. They next proceeded to the farm houſes of Bradley and Bean, at Balkham Hills. Mrs. Bradley's ſervant man they wantonly and inhumanly diſcharged a piſtol at, the contents of which have ſo ſhattered

Examinations before the Magiſtrates.

SATURDAY, FEB. 26.

Richard Ward and William Ramſey, two of the fugitives who abſconded from their employ at Caſtle Hill, on Tueſday the 15th Inſtant, underwent an examination, and were fully committed.

In the courſe of their examination, the priſoners gave the following account of their adventure, as raſh as it promiſes to be unfortunate :---That they had had recourſe to the woods, only as a relief from labour; that although they had abſconded on the ſame day that thirteen others had gone off (all of

Extract from the Reports of the Society for bettering the Condition, and increaſing the Comforts of the Poor.

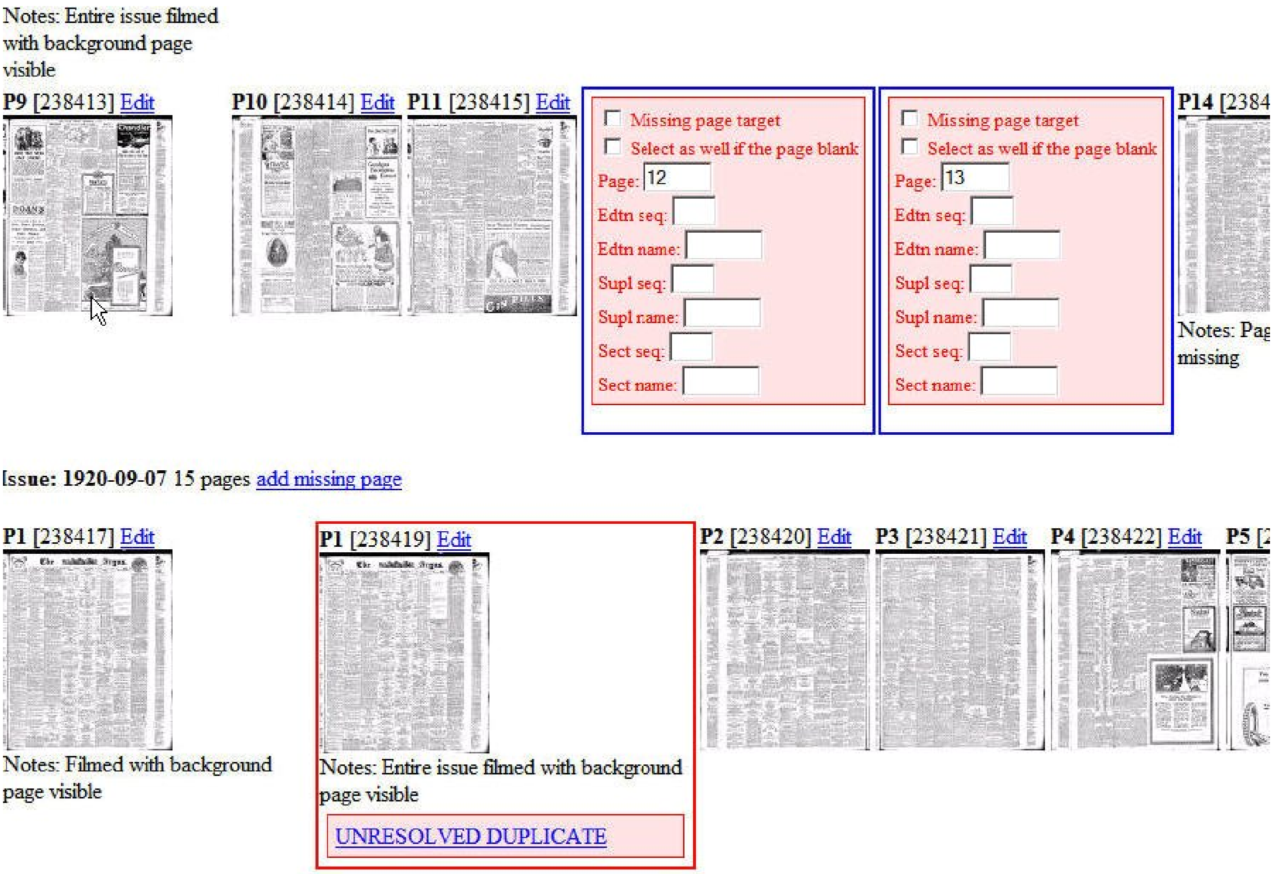
[From an Account of the Advantage of a Cottager keeping a Pig.]

By the Rev. Dr. Glaſſe.

James Bramſgrove, a Farmer's labourer, at Greenford, in Middleſex, had ſaved by his induſtry the ſum of four guineas. With this he purchaſed a hog, which he put up to fatten in the beginning of April laſt. In the courſe of the fattening, he uſed 3 quarters and a half of beans, and 7 buſhels of peaſe,

Fig 10. Addition of metadata and page sequencing.

Right from the start of the program it had been agreed that adding missing pages into the sequence at a later date and flagging known missing pages to users was of critical importance. Most other newspaper projects had come to the 'missing page issue' at a much later stage of the project when it was impossible to change workflows and therefore add in missing pages when they were found. However since we had identified the missing page issue at the start of the program as something that we would need to address we set up workflows so that if and when any missing pages are later sourced the virtual page sequence number can be shuffled so they can be added in. Currently we perform page verification which is part automated and part manual. If we sequence a page number 1 and then a page number 3 the system assumes that page 2 is missing and automatically generates a digital missing page target. It also identifies 2 pages that are the same and automatically generates digital duplicate page targets. These targets are manually checked by staff and verified or deleted. Staff also have the opportunity to manually add digital missing page/issue targets, or duplicate targets. In theory missing issues and pages should have a microfilm target but these are often incorrect or missing. To achieve consistency and flexibility all microfilm targets are being verified and replaced with uniform digital targets. The digital target has a virtual sequence number so that if the missing item is found it can be easily switched with the target. The digital target is actually acting as a place holder. Duplicate pages in the system are not deleted they are suppressed so that if they need to be switched later they can be.



Notes: Entire issue filmed with background page visible

P9 [238413] Edit

P10 [238414] Edit P11 [238415] Edit

□ Missing page target
□ Select as well if the page blank
Page: 12
Edtn seq:
Edtn name:
Supl seq:
Supl name:
Sect seq:
Sect name:

□ Missing page target
□ Select as well if the page blank
Page: 13
Edtn seq:
Edtn name:
Supl seq:
Supl name:
Sect seq:
Sect name:

P14 [2384

Notes: Pag missing

Issue: 1920-09-07 15 pages add missing page

P1 [238417] Edit

Notes: Filmed with background page visible

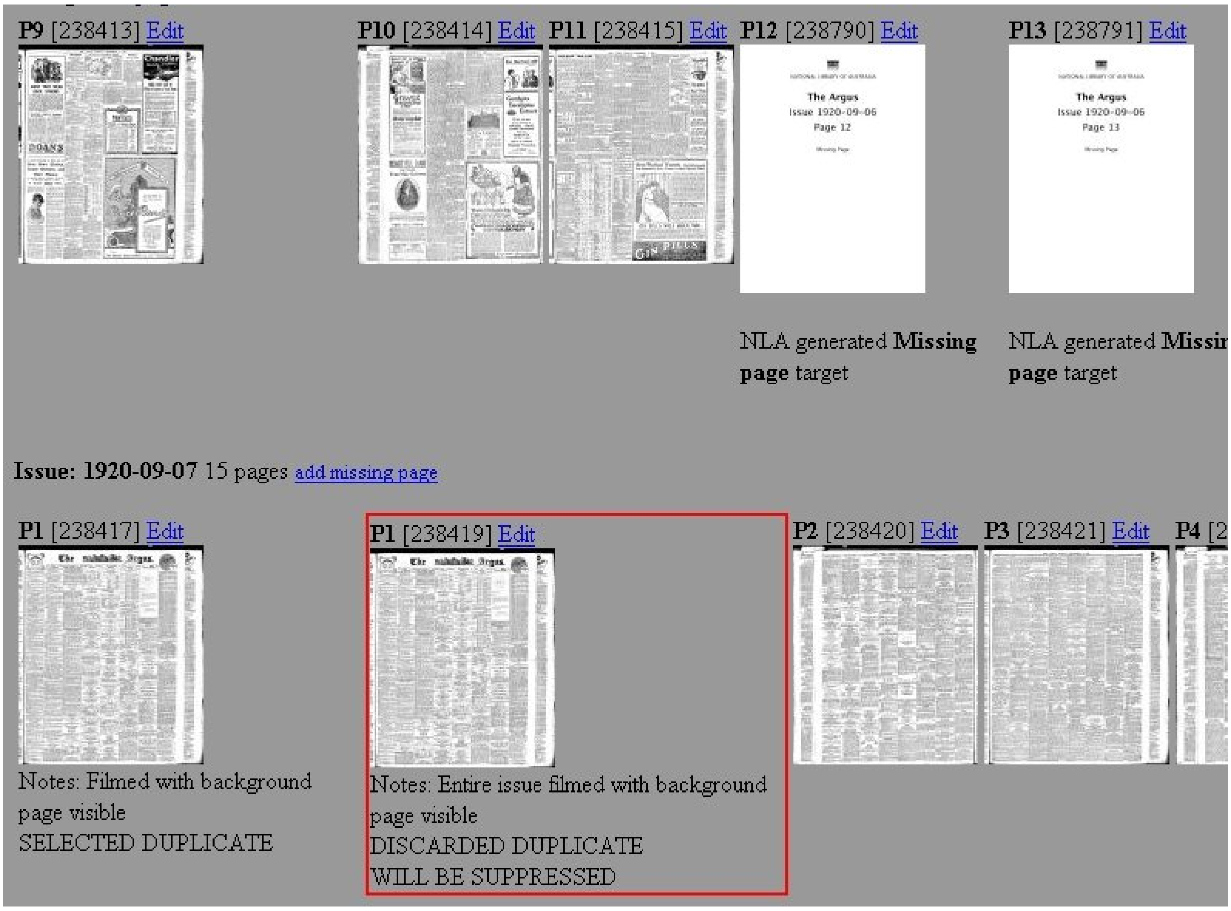
P1 [238419] Edit

Notes: Entire issue filmed with background page visible

UNRESOLVED DUPLICATE

P2 [238420] Edit P3 [238421] Edit P4 [238422] Edit P5 [2

Fig 11. Identification of missing and duplicate pages.



P9 [238413] Edit

P10 [238414] Edit P11 [238415] Edit P12 [238790] Edit

NATIONAL LIBRARY OF AUSTRALIA
The Argus
Issue 1920-09-06
Page 12
Missing Page

NLA generated **Missing page** target

P13 [238791] Edit

NATIONAL LIBRARY OF AUSTRALIA
The Argus
Issue 1920-09-06
Page 13
Missing Page

NLA generated **Missir page** target

Issue: 1920-09-07 15 pages add missing page

P1 [238417] Edit

Notes: Filmed with background page visible
SELECTED DUPLICATE

P1 [238419] Edit

Notes: Entire issue filmed with background page visible
DISCARDED DUPLICATE
WILL BE SUPPRESSED

P2 [238420] Edit P3 [238421] Edit P4 [2

Fig 12. Verification of missing pages and duplicate pages and target generation.

Once the initial quality assurance processes are completed the pages are placed into batches of 2000 pages and send to the OCR contractor. The OCR contractor is performing several processes which we refer to as content analysis and OCR:

- Zoning the page into areas (e.g. article, masthead).
- Linking multiple page articles and illustrations to articles.
- Categorising articles e.g. news, advertising.
- OCR'ing article text using Abby FineReader.

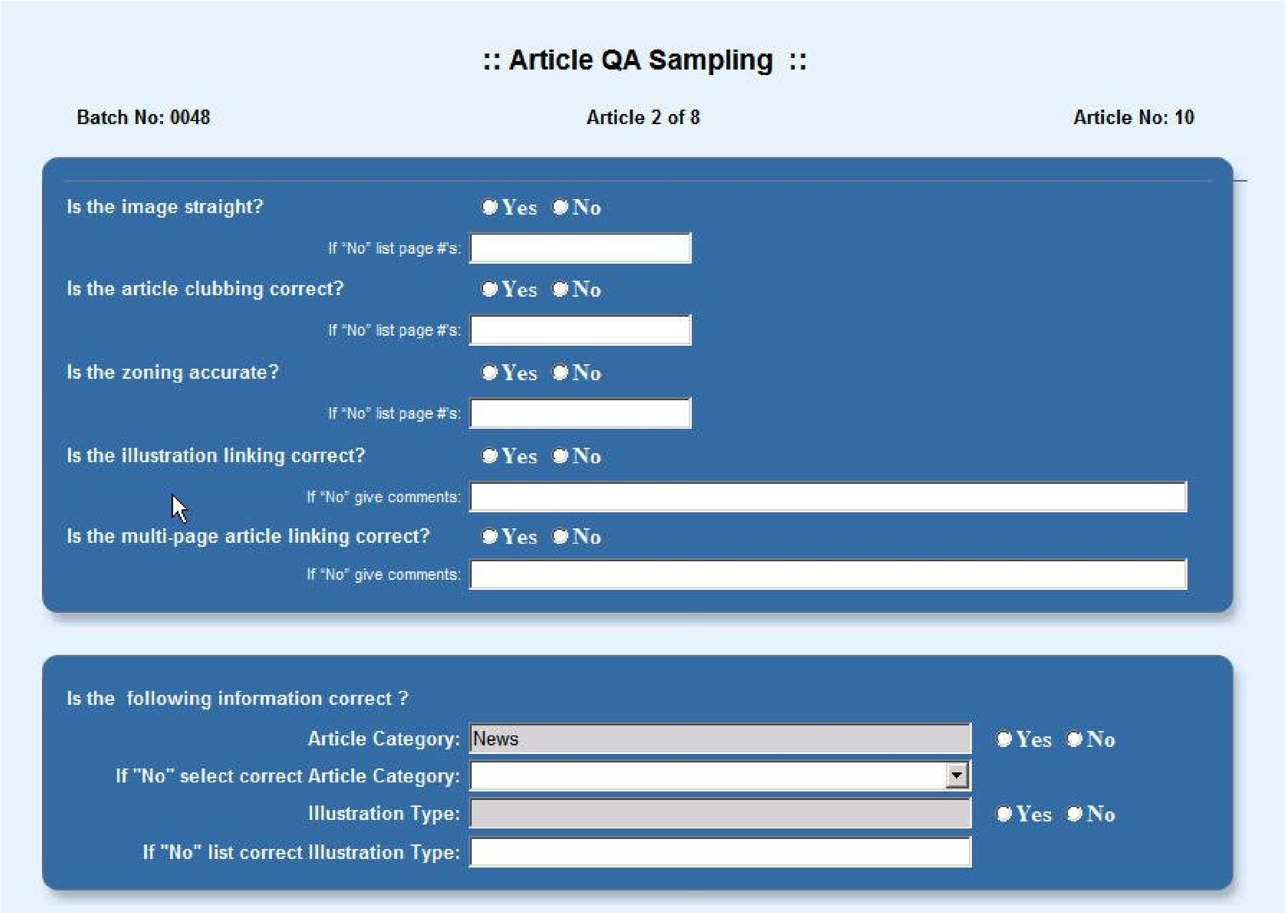
The quality of OCR text results can vary greatly. Very good OCR results can be achieved if the image text is clean and clear black on a white background. However most newspapers are not like this. For this reason the Library had decided to pay an additional fee to have titles and subtitles of articles and the first four lines of articles text re-keyed. This crucial text will be therefore have a high accuracy rate and this will greatly improve the user experience in searching and discovery capability. Searching will be more accurate and the initial results list will be more meaningful to users.

The OCR text is delivered back as XML files using the ALTO format, one ALTO file for each page. XML files are also supplied at the newspaper issue level. These files provide structural and descriptive information about the pages and articles in an issue and about the ALTO and image files which relate to them. The files are in a standard metadata format called METS (Metadata Encoding Transmission Standard) which is applicable to complex digital objects. Newspapers are considered complex digital objects since each newspaper issue comprises of multiple digital pages which are made up of multiple articles, some of which are linked across pages. The METS files include another metadata standard called MODS for the descriptive bibliographic information.

In September representatives from the Library visited our contractors based in India to view the processes being undertaken. The volume and nature of the work is very labour intensive. Most of the major OCR contractors are based in India for this reason. The picture below shows contractor staff working on the Library's program.

Fig 12. Contractor staff in India performing OCR and content analysis.

The first step of the content analysis process is zoning. Each newspaper page is divided into zoned areas e.g. articles, masthead area. Different types of zone are colour coded e.g. advert zones have a different colour to news articles. This speeds up the next stage of the process which is to categorise articles into different types e.g. news articles, births deaths and marriages. This is done manually. Additional steps such as linking articles which continue over multiple pages, and linking illustrations to relevant articles are undertaken. All zoned text is OCR'd using Abby Finereader software and in addition title headings, subheadings and the first 4 lines of text are manually re-keyed. Both our own staff and the contractor's staff perform quality assurance on the work undertaken using the same software module. This works on a batch sampling basis and staff check that the work specification has been carried out correctly and to the % accuracy level agreed. There are 15 criteria to check. The software enables tracking on the status of batches and reports of batch acceptance/rejection are generated on demand.



:: Article QA Sampling ::

Batch No: 0048 Article 2 of 8 Article No: 10

Is the image straight? ○ Yes ○ No

If "No" list page #'s:

Is the article clubbing correct? ○ Yes ○ No

If "No" list page #'s:

Is the zoning accurate? ○ Yes ○ No

If "No" list page #'s:

Is the illustration linking correct? ○ Yes ○ No

If "No" give comments:

Is the multi-page article linking correct? ○ Yes ○ No

If "No" give comments:

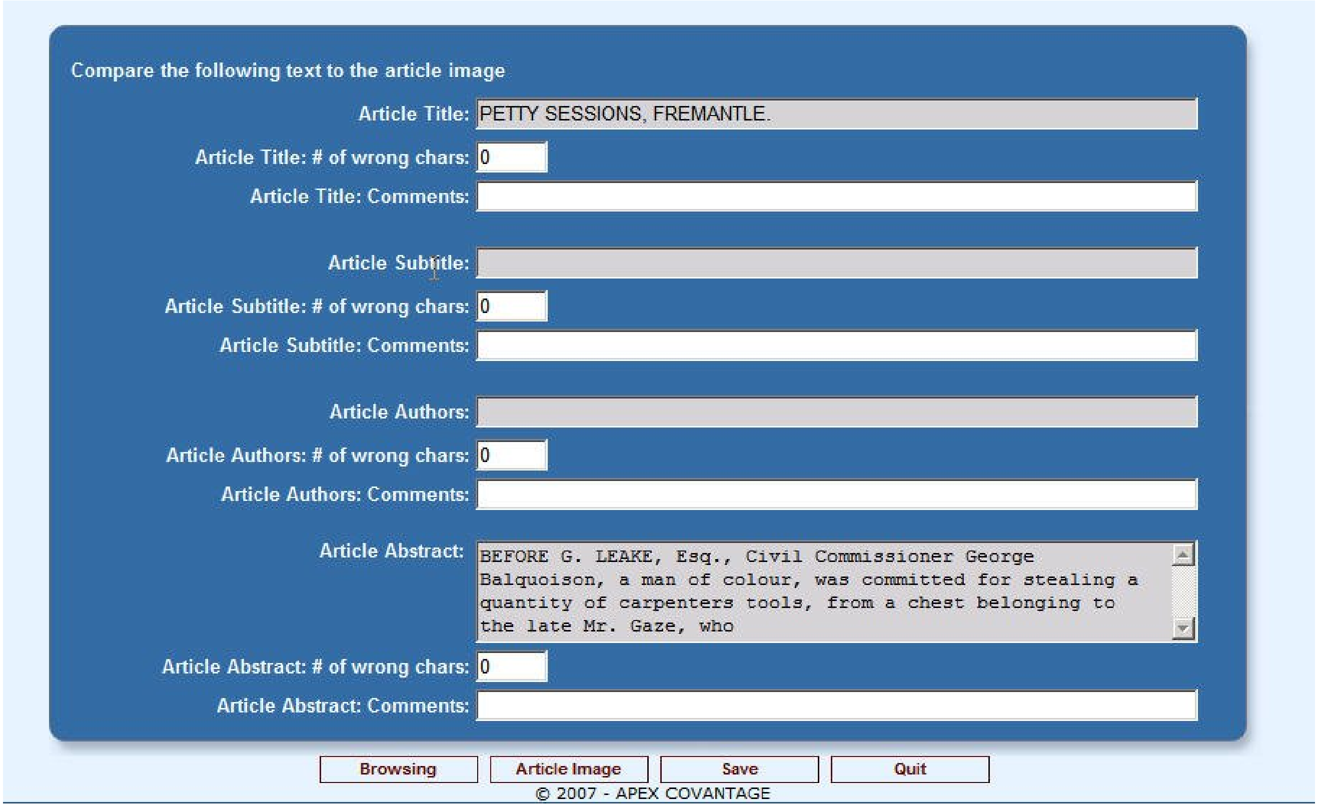
Is the following information correct ?

Article Category: News ○ Yes ○ No

If "No" select correct Article Category:

Illustration Type: ○ Yes ○ No

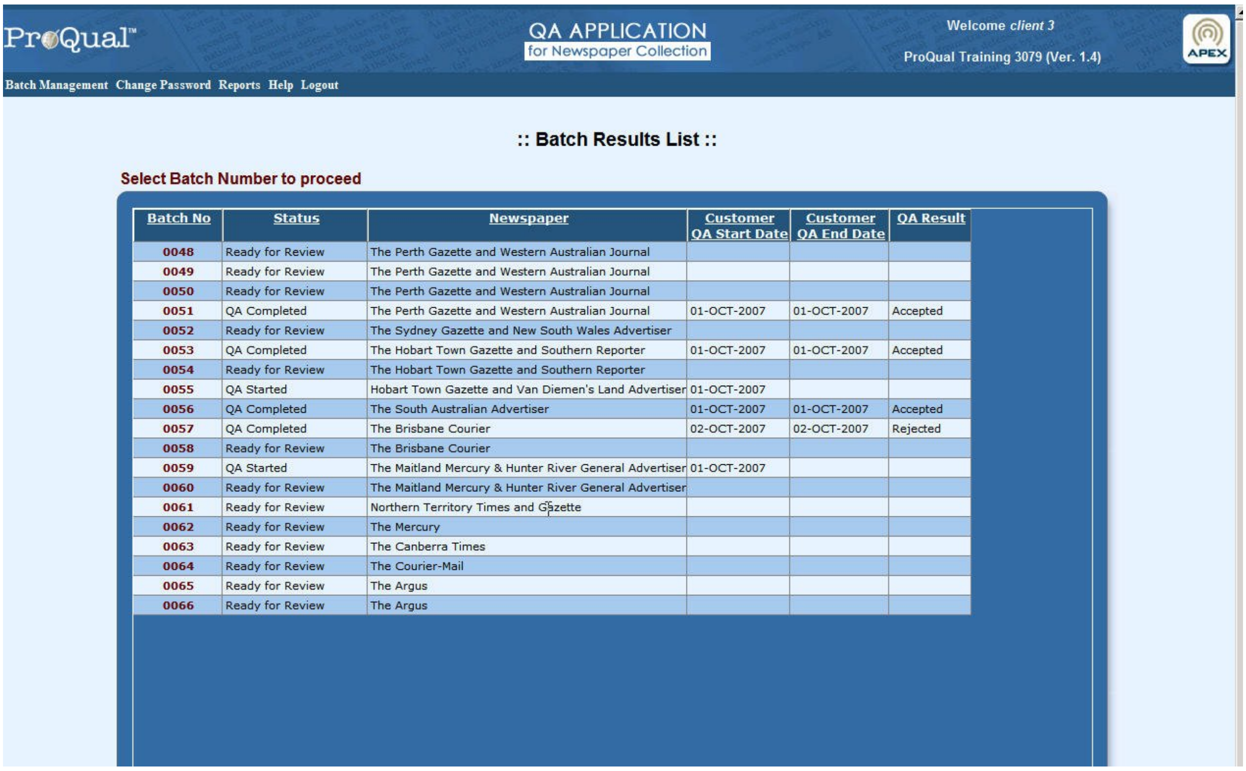
If "No" list correct Illustration Type:



Compare the following text to the article image

Article Title:	PETTY SESSIONS, FREMANTLE.
Article Title: # of wrong chars:	0
Article Title: Comments:	
Article Subtitle:	
Article Subtitle: # of wrong chars:	0
Article Subtitle: Comments:	
Article Authors:	
Article Authors: # of wrong chars:	0
Article Authors: Comments:	
Article Abstract:	BEFORE G. LEAKE, Esq., Civil Commissioner George Balquoison, a man of colour, was committed for stealing a quantity of carpenters tools, from a chest belonging to the late Mr. Gaze, who
Article Abstract: # of wrong chars:	0
Article Abstract: Comments:	

Fig 13. Quality Assurance processes.



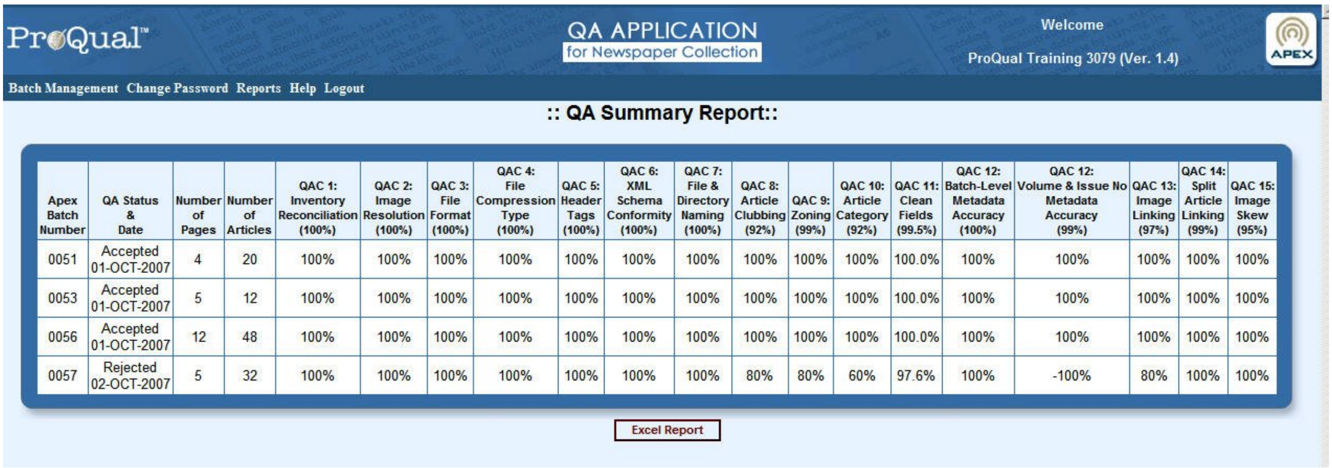
ProQual™
QA APPLICATION
for Newspaper Collection
Welcome client 3
ProQual Training 3079 (Ver. 1.4)
APEX
Batch Management Change Password Reports Help Logout

:: Batch Results List ::

Select Batch Number to proceed

Batch No	Status	Newspaper	Customer QA Start Date	Customer QA End Date	QA Result
0048	Ready for Review	The Perth Gazette and Western Australian Journal			
0049	Ready for Review	The Perth Gazette and Western Australian Journal			
0050	Ready for Review	The Perth Gazette and Western Australian Journal			
0051	QA Completed	The Perth Gazette and Western Australian Journal	01-OCT-2007	01-OCT-2007	Accepted
0052	Ready for Review	The Sydney Gazette and New South Wales Advertiser			
0053	QA Completed	The Hobart Town Gazette and Southern Reporter	01-OCT-2007	01-OCT-2007	Accepted
0054	Ready for Review	The Hobart Town Gazette and Southern Reporter			
0055	QA Started	Hobart Town Gazette and Van Diemen's Land Advertiser	01-OCT-2007		
0056	QA Completed	The South Australian Advertiser	01-OCT-2007	01-OCT-2007	Accepted
0057	QA Completed	The Brisbane Courier	02-OCT-2007	02-OCT-2007	Rejected
0058	Ready for Review	The Brisbane Courier			
0059	QA Started	The Maitland Mercury & Hunter River General Advertiser	01-OCT-2007		
0060	Ready for Review	The Maitland Mercury & Hunter River General Advertiser			
0061	Ready for Review	Northern Territory Times and Gazette			
0062	Ready for Review	The Mercury			
0063	Ready for Review	The Canberra Times			
0064	Ready for Review	The Courier-Mail			
0065	Ready for Review	The Argus			
0066	Ready for Review	The Argus			

Fig 14. Batch tracking and status



:: QA Summary Report::

Apex Batch Number	QA Status & Date	Number of Pages	Number of Articles	QAC 1: Inventory Reconciliation (100%)	QAC 2: Image Resolution (100%)	QAC 3: File Format (100%)	QAC 4: File Compression Type (100%)	QAC 5: Header Tags (100%)	QAC 6: XML Schema Conformity (100%)	QAC 7: File & Directory Naming (100%)	QAC 8: Article Clubbing (92%)	QAC 9: Zoning (99%)	QAC 10: Article Category (92%)	QAC 11: Clean Fields (99.5%)	QAC 12: Batch-Level Metadata Accuracy (100%)	QAC 12: Volume & Issue No Metadata Accuracy (99%)	QAC 13: Image Linking (97%)	QAC 14: Split Article Linking (99%)	QAC 15: Image Skew (95%)
0051	Accepted 01-OCT-2007	4	20	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100.0%	100%	100%	100%	100%	100%
0053	Accepted 01-OCT-2007	5	12	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100.0%	100%	100%	100%	100%	100%
0056	Accepted 01-OCT-2007	12	48	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100.0%	100%	100%	100%	100%	100%
0057	Rejected 02-OCT-2007	5	32	100%	100%	100%	100%	100%	100%	100%	80%	80%	60%	97.6%	100%	-100%	80%	100%	100%

Fig 15. Batch reports measuring against quality assurance acceptance criteria

SEARCH AND DELIVERY SYSTEM DESIGN

Two weeks ago work began on the search and delivery prototype development. Initial items that have been discussed for this are derivative types and sizes, zoom technology for images, search and browse features, results and refinement of results, possible user interactions with the resource and interface design. It is intended that the delivery service will enable:

- Full text searching of newspapers
- Searching of image captions
- Searching across multiple papers
- Refining searching by date, title and state published
- Browsing papers page by page
- Printing articles, pages and issues
- Zooming in and out of images (to read small text, to view context of an article within the page).

During the pilot the usefulness of categories is being explored. The pilot categories are:

- News
- Advertising
- Birth Death Marriage Notices
- Obituaries
- Editorial Commentary and Letters
- Shipping News
- Arts and Leisure
- Detailed lists, results and guides.

Categories may assist searchers in the delivery service, but it is not clear at this stage whether categories can be assigned correctly and whether the additional cost involved in manually categorizing articles actually does help the average user. The implementation and use of categories is therefore being monitored closely and feedback will be sought.

Illustrations are being sub categorized as:

- Photo
- Cartoon
- Map
- Graph

- Illustration

There is a flag in the metadata file to indicate if an article contains or consists of an illustration so that if it is desirable to search across these separately it can be achieved.

The size and number of derivatives and the method by which users may navigate around a page and zoom in and out has been a difficult issue to address. Although users will have the ability to zoom in and out of a page image when they first come to the image a default size must be selected by us. Feedback from testers indicated that the default page size should be at readable size rather than a thumbnail. However obtaining a standard readable size when the actual size and quality of the papers varies so much has been difficult. There are both tabloid size pages and broadsheet pages within the corpus and the font size and type varies greatly.

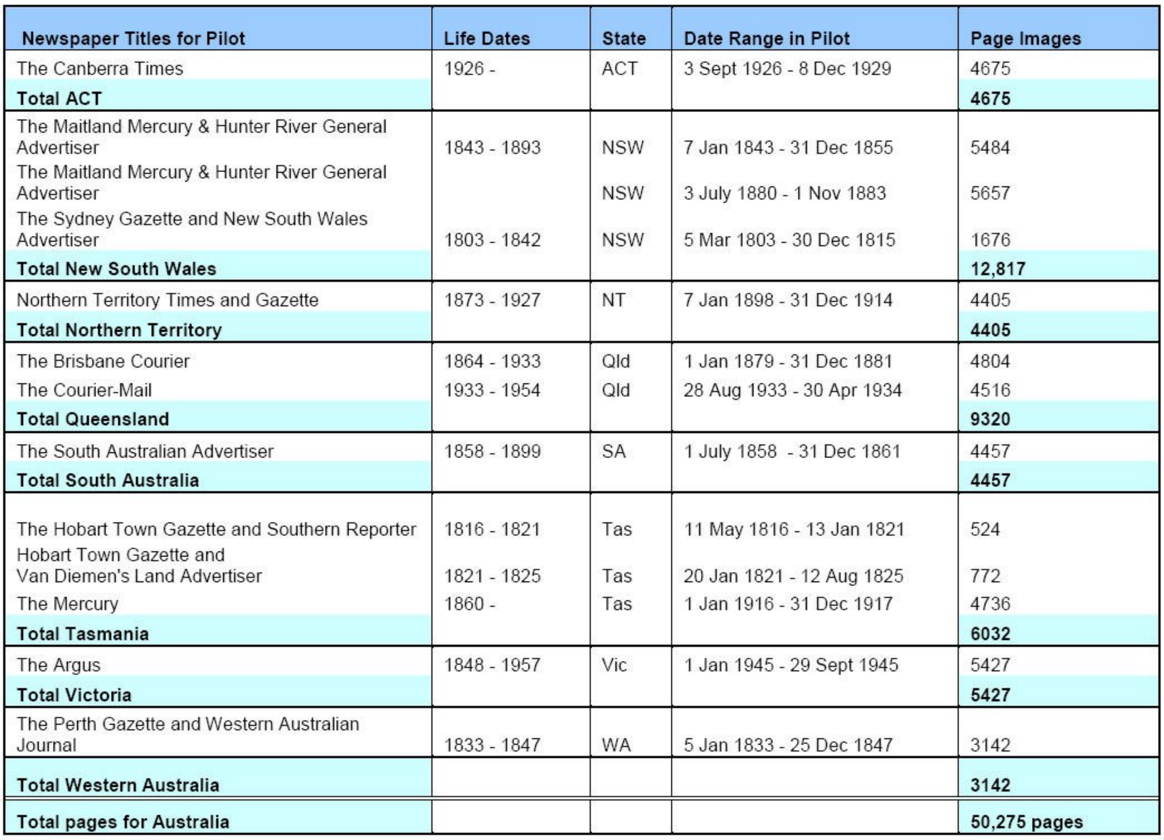
Whether standard zoom sizes e.g. 10. 20. 30. 40 % etc should be fixed and images pre-generated or whether image zooming should be done on the fly will significantly effect both speed of access and also image storage, so a compromise will need to be made. Variations in resolution and file type also effect storage and the user experience of quality. A happy medium between quick access and acceptable quality will need to be reached. The development team have spent time exploring the Google maps zoom technology which pre-generates a standard number of derivatives at fixed sizes (5%, 15%, 25%, 33%, 66%, 100%) and uses a navigation bar as the larger/smaller prompt. The Library is keeping in mind the requirements of visually impaired users and the abilities of some of the new zoom and image tools available to work with software products for the vision impaired. These explorations are at an early stage.

Other features currently under discussion are:

- The ability for users to correct OCR (the human eye is much better at reading text correctly than the OCR software eye, and correcting OCR will add value to the resource and improve searching for subsequent users. How and if OCR corrections will be moderated requires a lot more thought).
- Personal annotation (adding notes) to articles by users that others can also see.
- Tagging results and viewing tag clouds.
- Users creating public search sets of articles e.g. on historical or family events. This may be especially useful in the education sector.
- Clustering results e.g. by date, time to assist with discovery of articles
- Searching across other relevant resources, or integrating newspapers with other existing resources including the National Bibliographic Database, Picture Australia, newspaper paid subscription services etc.

The prototype will initially be released to stakeholders that have contributed newspapers and they will be asked for feedback on their own data and the proposed search and delivery features. It is hoped that the prototype will later be released to a wider audience including members of the public. In 2008 the prototype will be further developed taking into consideration the feedback obtained from stakeholders and users and will evolve into the search and delivery system version 1. It is anticipated that the digital newspaper service will be publicly launched in 2008 when a good body of digital newspaper data has been processed.

The table below summarises the content and scope of the pilot data that will be loaded into the prototype. It comprises 12 different titles, which equates to 8000 issues/50,000 pages/500,000 articles. These pages have been carefully selected to be a representative sample of the first 3 million pages and encompass a variety of formats, dates, and levels of quality. The pilot data needs to be a fair representation of the corpus so that all potential processing issues with different titles, dates, fonts and formats can be identified and resolved at the start of the program. This will enable efficient mass scale processing of titles once production begins without further unexpected issues cropping up along the way.



Newspaper Titles for Pilot	Life Dates	State	Date Range in Pilot	Page Images
The Canberra Times	1926 -	ACT	3 Sept 1926 - 8 Dec 1929	4675
Total ACT				**4675**
The Maitland Mercury & Hunter River General Advertiser	1843 - 1893	NSW	7 Jan 1843 - 31 Dec 1855	5484
The Maitland Mercury & Hunter River General Advertiser		NSW	3 July 1880 - 1 Nov 1883	5657
The Sydney Gazette and New South Wales Advertiser	1803 - 1842	NSW	5 Mar 1803 - 30 Dec 1815	1676
Total New South Wales				**12,817**
Northern Territory Times and Gazette	1873 - 1927	NT	7 Jan 1898 - 31 Dec 1914	4405
Total Northern Territory				**4405**
The Brisbane Courier	1864 - 1933	Qld	1 Jan 1879 - 31 Dec 1881	4804
The Courier-Mail	1933 - 1954	Qld	28 Aug 1933 - 30 Apr 1934	4516
Total Queensland				**9320**
The South Australian Advertiser	1858 - 1899	SA	1 July 1858 - 31 Dec 1861	4457
Total South Australia				**4457**
The Hobart Town Gazette and Southern Reporter	1816 - 1821	Tas	11 May 1816 - 13 Jan 1821	524
Hobart Town Gazette and Van Diemen's Land Advertiser	1821 - 1825	Tas	20 Jan 1821 - 12 Aug 1825	772
The Mercury	1860 -	Tas	1 Jan 1916 - 31 Dec 1917	4736
Total Tasmania				**6032**
The Argus	1848 - 1957	Vic	1 Jan 1945 - 29 Sept 1945	5427
Total Victoria				**5427**
The Perth Gazette and Western Australian Journal	1833 - 1847	WA	5 Jan 1833 - 25 Dec 1847	3142
Total Western Australia				**3142**
Total pages for Australia				**50,275 pages**

Fig 16. Table of newspapers included in the pilot data sample.

THE SYDNEY MORNING HERALD

The Library is pleased that very recently the Vincent Fairfax Family Foundation granted the Library $1 million to digitise the Sydney Morning Herald and include this in the ANDP. The Sydney Morning Herald is Australia’s longest running newspaper and of significance to the nation. It had been unable to be included in the program since the existing microfilm was incomplete and digitisation from hard copy was required. There is not a complete hard copy set of papers in single location so the digitisation of the Sydney Morning Herald would be a challenging and costly process. This issue has now been resolved with the grant of money which the National Library is matching.

SUMMARY

A tremendous amount of work has taken place in the Library over the last 6 months on the ANDP. I hope I have given you a useful insight into some aspects of the program and our plans to date. It is hard to cover the technical detail of such a large scale program in such a limited time. I would urge you to refer to the ANDP website for progress reports and more detailed information. The website is being regularly updated with documentation as we progress. Any specific questions that I do not have time to answer today can be e-mailed to me or sent via the website contact e-mail address which is received by members of the ANDP team. All feedback is useful to us and all questions will be answered.

I would also like to thank and acknowledge the work of the ANDP Team at the National Library of Australia and our State and Territory Library collaborators for their assistance in working towards our objective of helping communities access and explore their newspaper heritage.