

## MEMORANDUM

DATE 31 August 2007

REF NLA07/1243

TO NSLA

FROM Rose Holley – Manager Newspaper Digitisation Program

SUBJECT Progress Report for NSLA on the Australian Newspapers Digitisation Program August 2007

---

### **Purpose**

This report outlines the background to the Australian Newspapers Digitisation Program which commenced in February 2007. A summary of progress over the last 6 months and work to be completed over the next 8 months is outlined (February 2007 – April 2008).

---

### **1. Australian Newspapers Digitisation Program Background (NDP)**

On 29 November 2006 the National Library received approval from the Minister for the Arts and Sport to enter into a major newspaper digitisation contract.

The National Library is aiming to build a database containing content from the first Australian newspaper in 1803 through to the copyright cut-off date of 1954. We aim to digitise, during the four years of this contract, one major newspaper from each state and territory, amounting to about three million pages of content. We will extend this database in the future, and if possible during the first four years, through the addition of other content such as regional newspapers.

The Library will offer the Australian Newspapers Online service free of charge to researchers and the general public. The service will support 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”.

The National Library sees this Project as a key component of an ongoing digitisation program which will make Australian collections more visible and accessible. The Project will also give the Library valuable experience with “industrial scale” digitisation and the contract may provide a process for the future digitisation of other text-based content, such as out-of-copyright journals and books.

State libraries are supporting the National Library by ensuring the availability of quality microfilm versions of the relevant newspapers as the source material for the digitisation process. Some state libraries have secured financial support for the digitisation of additional newspapers, such as regional newspapers from their state.

The Library has engaged a Sydney-based company, W. & F. Pascoe Pty Ltd, which has commenced the conversion of the microfilm editions into digital page images.

The key part of this Project, the one covered by the Australian Newspapers Online Services contract, will convert the digital page images into text-searchable files through the use of Optical Character Recognition (OCR) technology and other processes including the “zoning” of the newspaper articles. The Library will use the services of Apex Publishing to undertake these processes.

The completion of contract negotiations with Apex Publishing in February 2007 marked the end of the 2 year scoping stage of the Newspaper Digitisation Project. The contract with Apex came into effect on 1 March 2007 and this marked the beginning of the Australian Newspaper Digitisation Program. From this date onwards practical development of the service would take place and a project plan would be developed for the initial year. It was anticipated that it would take approximately a year to develop systems, infrastructure, workflow processes and a data corpus so that a public service could be launched in early 2008. In April 2007 Rose Holley was appointed to manage the Newspaper Digitisation Program.

## **2. Key Milestones**

March 2007 – Contract with Apex for OCR and content analysis services commences.

April 2007 – Newspaper Digitisation Program Manager commences at NLA.

May 2007 – NLA complete development of workflow system to handle quality assurance of scanned images.

June 2007 – NLA complete scanning, quality assurance and metadata entry on 50,000 newspaper pages for the pilot phase of project.

July 2007 – NLA upgrade infrastructure to meet storage needs of Newspaper Digitisation Program (62 TB required in first year).

August 2007 – Apex complete development of production software to NLA specification for processing and quality assurance of newspaper digital images, and begin to process 50,000 page pilot.

September 2007 – 50,000 page image pilot completed and evaluated.

October 2007 – Phase 1 (500,000 page images) commences.

October 2007 – Search and delivery system prototype completed and 50,000 pilot data ingested to system.

November 2007 – Search and Delivery v1 released to stakeholders for feedback.

April 2008 - Phase 1 (500,000 page images) completed.

April 2008 – Public Launch of service (anticipated).

In addition the National Library will be developing models for contribution and a digitisation schedule of titles for the next 3 years. Feedback will be sought from State Libraries. Order of digitisation of titles depends largely on quality of microfilm, accessibility of microfilm and timeframe of new and re-filming work being carried out as part of ANPLAN for some of the titles which have been identified as significant by State Libraries.

### 3. Deliverables and Outcomes

In order to meet the objectives of the program, the following project deliverables will be developed internally by the National Library of Australia:

- Documented workflows and processes which enable the Library to undertake large scale digitisation in an efficient and effective way, and which can be applied to other collections formats such as books and/or journals.
- Workflow support system to manage the workflows for producing digitised newspaper content. This includes an internal quality assurance tool with appropriate development, test and training environments.
- Content management system to manage and store the digitised newspaper content produced.
- Search and delivery system which will support users in searching and accessing the digitised newspaper content produced.
- System infrastructure required to achieve the above.

The developed applications will conform to the Library's application framework and service-oriented architecture.

The developed applications will have interfaces to the Library's persistent identifier standards and resolver service.

The system infrastructure will conform to the Library's IT security, data backup and other existing infrastructure procedures.

In order to meet the objectives of the program, the following project deliverables will be developed externally by companies contracted by the Library:

- 17,500 newspaper page images digitised per week.
- Conversion of newspaper page image files using OCR and content analysis processes to identify unique page segments, and generation of required file outputs to the Library based on 17,500 page image files per week.
- Pro-Qual quality acceptance software developed and maintained by Apex.
- Tracking and issue management systems developed and maintained by Apex.

#### **4. External dependencies**

The success of the Australian Newspapers Digitisation Program and keeping within preferred timeframes is largely dependant upon the following:

- Owners of master microfilm being able to supply high quality microfilm masters of historical Australian newspapers as required.
- W. & F. Pascoe's PTY Ltd (or any alternative agencies) being able to create digital newspaper content to the quality and volume required.
- Apex Publishing (or any alternative company) being able to undertake OCR and other processing of the digital newspaper content to the quality and volume required.

#### **5. Evaluation of Success**

The National Library will be able to determine the success of the Program through the following achievements:

- User satisfaction with the search and delivery service;
- Stakeholder satisfaction with the model developed for newspaper content contribution by other institutions;
- Library processes, workflows and IT infrastructure developed to support mass digitisation activities.

#### **6. Work Progress February 2007 – August 2007**

- Project teams have been established and both permanent, fixed term and casual staff appointed to various roles. Project teams are: NDP Board, NDP IT team, NDP Search and Delivery Team, NDP Digitisation Team (including NDP Quality Assurance Team).
- Project management tools and processes have been implemented for the initial year of the program and will be maintained until the service is publicly launched and enters operational phase.
- From November 2006 – August 2007 350,000 newspaper pages have been scanned from microfilm by W. & F. Pascoe Pty Ltd. Of these 200,000 have had metadata added (newspaper title, volume, issue and page information), and been quality assured (sequencing of pages, identifying duplicate and missing pages) by the NDP Quality Assurance Team.
- The first parts of the workflow management system have been developed in house by the National Library to support the work of the NDP Quality Assurance Team. The main parts of this system are the NLA Quality assurance tool and the Ingest system.

- The National Library infrastructure and storage has been upgraded so that enough capacity and working space is available for the Newspaper Digitisation Program on an ongoing basis. To date 62 TB of storage space has been purchased and implemented.
- Apex have developed the OCR, zoning and categorisation production software and quality assurance software to meet the National Library's work specification and requirements. The Apex quality assurance software is called Pro-Qual. The systems and specification have been documented in the Conversion System Design Document.
- The National Library has developed guidelines for metadata requirements (METS and ALTO files) and defined categories for articles.
- Apex production software has been successfully tested and Apex have been given the go ahead to proceed with production of a 50,000 pilot. The original pilot selection identified in 2006 fell from 50,000 to 39,758 images after duplicates were removed and quality assurance took place. To bring the pilot size back up to 50,000 images the Maitland Mercury 1843-1855 and Brisbane Courier 1879-1881 were included, as these were microfilms of reasonable quality that we already had the digital images for. The contents of the 50,000 pilot are outlined below:

Newspaper Title	Date Range
Canberra Times	03/09/1926 to 08/12/1929
Sydney Gazette	05/03/1803 to 30/12/1815
Northern Territory Times and Gazette	07/01/1898 to 31/12/1914
Maitland Mercury	07/01/1843 to 31/12/1855
Maitland Mercury	03/07/1880 to 01/11/1883
Brisbane Courier	01/01/1879 to 31/12/1881
Courier Mail	28/08/1933 to 30/04/1934
South Australian Advertiser	01/07/1858 to 31/12/1861
Hobart Town Gazette and Southern Reporter	11/05/1816 to 13/01/1821
Hobart Town Gazette and Van Diemen's Land Advertiser	20/01/1821 to 12/08/1825
The Mercury	01/01/1916 to 31/12/1917
The Argus	01/01/1945 to 29/09/1945
Perth Gazette and Western Australian Journal	05/01/1833 to 25/12/1847

- Wireframes have been developed for the search and delivery prototype, and the range of functional requirements is under discussion. (Wireframes are a basic visual guide used to suggest the layout and placement of fundamental design elements in the web interface design. They provide a visual reference upon which to structure each page.)

- A new website has been created to enable the public and stakeholders to gain information about the program <http://www.nla.gov.au/ndp/>

## **7. Work to be completed September 2007 – April 2008**

- Apex will process the 50,000 pilot data and this will be quality assured by both Apex and National Library. After review amendments and changes may be made to production software and work specifications, before production begins on a further 500,000 images.
- National Library will continue to develop the software required to support the workflow process, content management, and search and delivery systems.
- National Library will create specifications for derivative images for use in the search and delivery system, and then generate derivatives across the data corpus.
- Search and Delivery Prototype is to be developed by the National Library. This will be released to stakeholders with the 50,000 pilot data in, for feedback.
- Microfilm will continue to be scanned at a rate of 17,500 images per week until the 500,000 page milestone is reached, when outputs will be reviewed.
- A long term plan for selection and prioritisation of newspaper titles will be developed in consultation with stakeholders, taking into account availability and quality of microfilm, as well as significance of titles. The plan needs to cover 3 million pages for the first 3 years. Previous scoping plans from 2005 and 2006 will be reviewed taking into account new information and the current co-operative re-filming program of ANPLAN.
- A model for national collaboration will be developed in consultation with stakeholders. This will include standards to work to and digital storage scenarios.
- The NDP service will be launched to the public (no date set but hope to be early 2008. The date cannot be set at this stage since it will be influenced by dependencies outlined in section 5. Service will not be officially launched until there is suitable body of content to enable successful searching).
- An operational model and staffing will be established for the ongoing program after public launch.
- Communication and marketing plan to be developed.
- Service to be evaluated (this will be some time after launch).

## **8. Summary**

Implementing the NDP program is proving to be a complex and challenging experience for the National Library. To date the progress is within agreed timeframes. The entire workflow process is made up of many parts and there are significant dependencies on external vendors and software. This makes the program high risk. The development process has been very agile since technology in this area is changing rapidly and some aspects have changed since

the initial project scoping was undertaken 2 years ago. Moving a theoretical workflow and service model into a practical application has raised both expected and unexpected issues. There have been more issues than were expected, not just in the technology area. The quality assurance processes, software and staffing both internal and external are proving to be a major and important aspect of the overall program. Working across 3 time zones (Australia, USA and India) is also challenging.

Despite all this the National Library are pleased with the progress to date of the program and are keen to be able to provide stakeholders with a prototype to view as early as they can. There has been enormous interest in the program from the government, media, stakeholders and public and the National Library are well aware that the potential users would like the service delivered as soon as possible.