Annual report to partners 2006-2007

Contents

PANDORA, Australia's Web Archive < http://pandora.nla.gov.au/index.html>, is a selective archive of Australian online publications and web sites which is built collaboratively by the National Library, all of the mainland State libraries, the Northern Territory Library, the National Film and Sound Archive, the Australian War Memorial, and the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). This is a report to contributing partners on activities and developments in the 2006-2007 financial year.

1. PANDORA participants working together

1.1 Consultation mechanisms

The National Library continued to inform and consult with other PANDORA participants about the operation of PANDORA through the two email discussion lists – pandoraconsult-l and pandora-l respectively.

1.2 Reports

Each month, a report on the growth of the Archive, usage statistics, and a summary of responses to the online PANDORA user survey forms are sent to both email discussion lists. This report includes information about the ten most popular sites for the month and which agency has archived them.

On a bi-monthly basis, the Library compiles two lists of instances[1] archived by each partner agency – one list contains all instances archived during the period and the other details government publications only. These lists are put up on the PANDORA web site at http://pandora.nla.gov.au/newtitles/new_titles_reports.html and partners are advised of their availability via a message to the two email discussion lists.

An annual report of progress and activities to the Chief Executive Officers of partner agencies is also provided.

2. Growth of the Archive

	30 June 2006	30 June 2007	Growth 2006-07
Titles	12,297	15,236	2,939 (23.9%)
Instances	24,536	30,285	5,749 (23.4%)
Gigabytes[2]	1180	1449	269 (22.8%)
Usage (page views)	8,794,547	5,708,690	-3,085,853 (-35%)

The Archive continued to show good content growth during 2006-2007, with percentage increases for the number of titles and number of instances at just over 23%. The data size of the Archive approached 1.5 Terabytes. There has, however, been a noticeable downward trend in usage (page views) of the Archive since October 2006.

Government publications comprise approximately 50 per cent of the Archive. In July 2006 the Commonwealth Copyright Reproduction Licence between the National Library and the Commonwealth Copyright Administration (CCA) was renewed for a second four

1 An 'instance' is a single gathering of a title. It includes the gathering of a monograph that has been archived once only, the first gathering of a serial title or integrating title (for example, a web site that changes over time), and all subsequent gatherings.
2 This figure does not include the preservation and other master and back up copies.

year period. This licence forms the basis by which the CCA seek to secure permission to archive content from Commonwealth government domains on behalf of the Library.

3. Development of the Archive

To keep pace with a rapidly changing web archiving environment the National Library is committed to ongoing development of the policy, procedures and technical infrastructure which support the Archive.

3.1 Development of PANDAS

PANDAS (the PANDORA Digital Archiving System) is web-based software developed by the Library to enable PANDORA staff in participating agencies to carry out all of the tasks involved in contributing selected online publications and web sites to PANDORA. (This does not include cataloguing, which is carried out in separate systems.)

During 2006-2007 the Library completed the re-development of the web archiving management system and PANDAS version 3 was deployed and released to partners on 27 June 2007.

The original aim in developing PANDAS 3 was to tidy up some code and improve stability and maintainability. However, we determined that it would be more effective to do a full re-engineering of the software. This proved to be a major undertaking but has provided the opportunity to include some enhanced functionality and to redesign the user interface to make it better able to support the web archiving workflows. On one level the re-engineering of the code will provide better system performance. In addition to this, the improvements to the interface, including better navigation and the introduction of personal 'worktrays' to manage titles at all stages of the web archiving workflow will improve the efficiency of the system.

3.2 Whole domain harvest

In August and September 2006 the Library conducted the second large scale harvest of the Australian web domain, following on from the first harvest conducted in 2005. Despite the advantages of the selective approach to archiving, its disadvantages have long been recognised by the Library. Resources are taken out of context, and their links to other web documents are broken. In addition, important resources are missed. Government publications comprise just one category of material that we know none of the PANDORA participants can address adequately via the selective approach. There are just too many titles to be captured.

As with the 2005 harvest, in 2006 the National Library contracted the Internet Archive to undertake the whole domain harvest crawl on our behalf. The Internet Archive has extensive experience in this form of web archiving. This second harvest had the objective of harvesting 500 million unique URLs from the .au web domain and other resources on hosts located in Australia (where these could be automatically identified as such). The crawl was seeded with a large number of URLs from the previous domain crawl and the National Library also provided smaller seed lists which included the URLs of priority web sites, particularly in the government and educational sectors. The intention was to bias the harvest towards these sectors to capture as many of these resources as possible.

The harvest was run for four to five weeks during August and September 2006 and around 596 million unique documents were captured, amounting to 19.04 terabytes of

data. The Internet Archive indexed the contents of the 2006 and 2005 harvests and shipped it to the National Library in late 2006, where it was installed for local access.

In the absence of legal deposit provisions for online publications and web sites at the Commonwealth level, the access that the Library can provide to the whole domain harvest remains limited. It is not currently available to the general public. Unlike the selective Archive, we have not negotiated permission from publishers to archive and to provide access to the contents of the whole domain harvest. An issues paper on access to the domain harvest content was prepared in May 2006 and the Library will consider options for providing wider access to content from the domain harvest.

The Library is preparing for a third whole domain harvest commencing in September 2007. This third harvest will follow the 2006 harvest with an objective of at least 500 million unique URLs.

4. Focus on users

Once again this year an analysis of usage of the Archive during the previous financial years, 2005-2006 and 2006-2007, was undertaken.

The analysis showed a continued growth in usage during the 2005-2006 financial year over the previous year (22.5 %), but a corresponding decrease in usage in the 2006-2007 financial year (down 23 %). The reason for this is not clear. It is not clear if this is due to an abnormal increase in usage in 2005-2006 and the 2006-2007 figures are normalisation of a small usage growth; or whether there are other reasons for the decrease in 2006-2007.

Usage in 2006 - 2007

Total pageviews	Average per month	Month of highest use	Month of lowest use
5,708,690	475,724	September – 650,553	February – 375,962

Usage in 2005 - 2006

Total pageviews	Average per month	Month of highest use	Month of lowest use
7,422,601	618,550	October - 764,662	July – 435,925

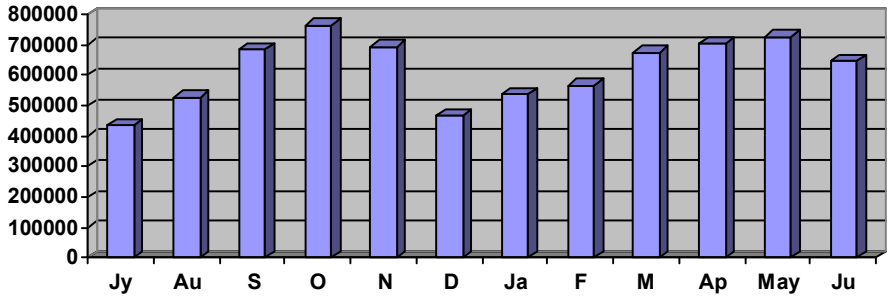
Usage in 2004 - 2005

Total pageviews	Average per month	Month of highest use	Month of lowest use
5,390,459	449,204	May - 649,786	December – 276,983

Usage in 2003 - 2004

Total pageviews	Average per month	Month of highest use	Month of lowest use
4,398,148	366,512	May – 560,168	December – 174,070

Month by month usage (pageviews) for July 2006 – June 2007



As another measure of Archive usage, according to Google, as at the end of the 2006-2007 financial year there are around 9,000 external links to the PANDORA website. Of these, 3,500 point to the PANDORA home page while a significant number, 1,767, point to the important election campaigns subject listing page. This points to the value of the archiving of election campaign web content especially as we approach a federal election later in 2007.

While only 4 per cent of the titles archived in PANDORA have completely disappeared from the live Web, 53 per cent (between five and six) of the ten most heavily-used titles in the period July 2006 to June 2007 are no longer available from the publishers' web sites. This suggests that a reasonable proportion of users are coming to the Archive for its primary purpose, which is to provide access to online publications and web sites that are no longer available elsewhere.

There were 74 responses to the PANDORA User Survey questionnaire between July 2006 and June 2007. These responses indicated that usage of PANDORA was spread fairly evenly across all age groups, with those in the 36 – 49 (22%) and 50 – 64 (22%) age groups being the most frequent users. The primary purpose for visiting the Archive was academic or professional research (38%), but family history was also popular (16% of respondents most of whom were interested in the *First Families 2001* site which is consistently the most heavily-used title).

5. Preservation

The National Library continued to monitor the range of file formats entering the PANDORA archive, maintaining an ongoing profile of its technical makeup.

Working as part of the International Internet Preservation Consortium (IIPC), the Library provided input to the submission of the WARC Web archiving format as an ISO standard, and is also leading the IIPC Preservation Working Group, which is charged with identifying approaches, standards and practices for digital preservation that are applicable to Web archives.

As a partner in the Australian Partnership for Sustainable Repositories (APSR), the National Library also led a project to further develop an Automated Obsolescence Notification System (AONS), which will aid repository managers in assessing and monitoring the risk of obsolescence for the range of formats held in their repositories.

An adapter to allow monitoring of the PANDORA archive is expected as part of development, and a sample of the PANDORA archive may be used to test the efficacy of the application for Web archives.

6. Organisational Restructure

In December 2006 the Digital Archiving Branch which was responsible for the business management of the PANDORA Archive was transformed into a new branch within the Library's Collections Management Division. This new branch brings together in a closer working relationship the Library's web archiving activities and digital preservation activities under the directorship of Colin Webb. The new branch called the Web Archiving and Digital Preservation Branch has responsibility for the PANDORA Archive and whole domain web archiving, digital preservation, preservation standards and associated collaborative involvements including that with APSR (Australian Partnership for Sustainable Repositories), MAGDIR (Managing Australian Government Digital Information Resources) and the IIPC (International Internet Preservation Consortium).

7. International relations

During 2006-2007 the Library continued its active participation in the International Internet Preservation Consortium[3]. The Library decided to continue its membership of the IIPC Steering Committee and at the January 2007 meeting of the Committee a new Working Group on Preservation was established which will be lead by National Library of Australia. The initial task of the this working group – chaired by Colin Webb – will be to characterise large scale web archives so as to identify current approaches to preservation including standards and practices and to make recommendations for enhancements or additions to tools, standards, practice guidelines, testing and possible further research.

8. Promoting the Archive

As the focus of activity over the last year has been on the development of PANDAS 3 we have not undertaken significant promotion of the Archive in that period. This may have had some bearing on the figures. The Library recognises the need to resume more active promotion of the Archive and our web archiving activities in the coming year.

8.1 PANDORA Fact Sheet

The Library has continued to update the PANDORA Fact Sheet on a monthly basis and to distribute it to participants for publicity purposes. It summarises key information about the Archive and supplements the printed PANDORA Brochure.

8.2 Papers and articles

A number of papers and articles were published and presented during the year for the dual purpose of promoting our work and sharing what we have learned. These include:

- Phillips, M. and Koerbin, P. *PANDORA, Australia's Web Archive: how much metadata is enough?* An article published in the Journal of Internet Cataloging, volume 7 issue 2 (cover date 2004, published in 2007).

3 Information about the IIPC is available from its web site at http://netpreserve.org/about/index.php

- Koerbin, P. *Web archiving at the National Library of Australia. PANDORA: Australia's Web Archive*. A short article published in the March 2007 issue (no. 58) of the CDNLAO Newsletter, available online at http://www.ndl.go.jp/en/publication/cdnlao/058/581.html

- Crook, E. *For the Record: Assessing the Impact of Archiving on the Archived*. A paper by Edgar Crook based on a survey of publishers of titles archived in PANDORA. Published in RLG DigiNews, August 15 2006. Available online at http://www.rlg.org/en/page.php?Page_ID=20962#article0 '

- Crook, E. *The work of PANDORA*. A brief history of the first 10 years of the PANDORA Archive. Published in Gateways, August 2006. Available online at http://www.nla.gov.au/pub/gateways/issues/82/story01.html

- Koerbin, P. "Purpose, Pragmatism and Perspective: Preserving Australian Web Resources at the National Library of Australia". A paper presentation by Paul Koerbin at the *Internet Convergences 7.0* the 2006 international conference of the Association of Internet Researchers held in Brisbane in September 2006.

- Paul Koerbin also presented papers on PANDORA and web archiving at the *DCM Evolution 2006* conference hosted by The Government Practice in Canberra in November 2006; and at the *Digital Preservation Seminar* organised by Charles Sturt University and hosted at the National Library in November 2006.

9. Concluding summary

Some of the highlights of 2006-2007 include:

- Content of the Archive grew by 23.4 per cent[4] (section 2);

- The completely re-engineered and enhanced version of the Library's web archiving management system was completed and deployed in June 2007. PANDAS version 3 will provide improved system performance and includes a number of workflow and interface enhancements (section 3.1);

- The second large scale harvest of the Australian web domain was undertaken in August – September 2006. This resulted in an archival collection of 596 million files amounting to 19 terabytes of data. This is three times the size of the first domain harvest conducted in 2005. A third large scale domain harvest is planned for August-September 2007 (section 3.2);

- The National Library continued to monitor the range of file formats entering the PANDORA archive, maintaining an ongoing profile of its technical makeup (section 5); and,

- Active participation in the International Internet Preservation Consortium continued with the National Library leading a newly established IIPC Preservation Working Group (section 7).

4 Calculated on number of instances added.

Appendix 1

PANDORA Consultative Committee – list of representatives

Australian Institute of Aboriginal and Torres Strait Islander Studies
Pat Brady, Collection Manager, AIATSIS Library

Australian War Memorial
Mal Booth, Head, Research Centre

National Film and Sound Archive
Matthew Davies, Manager, Collection Development

National Library of Australia
Colin Webb, Director, Web Archiving and Digital Preservation (Chair of Committee)

Northern Territory Library and Information Service
Ann Ritchie, Assistant Director

State Library of New South Wales
Jim Tindall, Senior Librarian, Collection Services

State Library of Queensland
Sharon Nolan, Manager, Published Material, Heritage Collections

State Library of South Australia
Tony Leschen, Manager, Collection Development

State Library of Victoria
Liz Jesty, Manager, Collections Management

State Library of Western Australia
Monika Szunejko, Manager, Access