Annual report to partners 2007-2008

Contents

PANDORA, Australia's Web Archive http://pandora.nla.gov.au/, 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 2007-2008 financial year.

1. PANDORA participants working together

1.1 Consultation mechanisms

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

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 is sent to both email discussion lists. This report includes information about the ten most popular (most viewed) 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 published 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. These reports are also available on the PANDORA website Partners page http://pandora.nla.gov.au/partners.html.

2. Growth of the Archive

	30 June 2007	30 June 2008	Growth 20076-08
Titles	15,236	19,308	4,072 (26.7%)
Instances	30,285	38,183	7,898 (26.1%)
Terabytes[2]	1.42	1.86	269 (30.1%)
Usage (page views)	5,708,690	7,295,996	1,587,306 (27.8%)

The Archive continued to show steady growth and a small increase in the growth rate during 2007-2008 over the previous year, with percentage increases for the number of titles and number of instances at just over 26 %. The data size of the Archive grew to more than 1.8 terabytes. After a downturn in the previous year in the number of user page views, 2007-2008 showed a strong increase in usage (up by just over 27 %).

Government publications comprise approximately 50 % of the Archive. Serial (16 %) and integrating (34 %) titles comprise 50 % of the archived titles. These formats are

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.

usually scheduled to be re-harvested at regular periods as content changes. The remaining 50% are treated as monographic titles with no change expected and are generally harvested only once.

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.)

The third version of PANDAS was deployed and released to partners on 27 June 2007 and the past financial year has involved bedding down the completely re-engineered and enhanced system and providing support to PANDAS users. The new version of PANDAS provides enhanced workflows and a more stable system. User documentation was written and made available on the PANDORA website and training was provided in Canberra. Training was also provided to PANDORA participant agencies outside Canberra on the basis of the participant organisation funding the travel costs for their staff or the NLA officer. Two partner agencies took up this opportunity.

The implementation of PANDAS version 3 has provided the new functionality to manage the PANDORA subject listings through the user interface. The existing subject listings were reviewed and a new more substantial and useful list of subject headings was created in consultation with staff in participant agencies. Adding titles to the new subject headings remains an ongoing task to be completed by all participants.

3.2 Whole domain harvest

In August and September 2007 the Library conducted the third large scale harvest of the Australian web domain, following on from previous harvests in 2005 and 2006. 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 and 2006 harvests, in 2007 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 third 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 harvest was run for four weeks during August and September 2007 and around 516 million unique documents were captured, amounting to 18.47 terabytes of data. The combined total for all three Australian domain harvests has now reached 1.3 billion files amounting to 44.2 terabytes of data.

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. Programmatic access has been provided to some university based researchers.

The Library is undertaking a fourth Australian domain harvest to commence in July 2008. This harvest has the objective to collect 1 billion unique documents, making it the largest single crawl of the Australian domain yet undertaken. It is expected to take around 12 weeks to complete.

3.3 Review of web archiving

The National Library commenced a review of its web archiving activities with the purpose of 're-visioning' its approach to web archiving based on 12 years experience leading the PANDORA Archive and its scoping and commissioning of large scale domain harvests since 2005. The wide ranging review will report, with recommendations, in late 2008.

4. Focus on users

Once again this year an analysis of usage of the Archive over the last three financial years was undertaken.

The analysis showed a continued growth in usage during the 2007-2008 financial year over the previous year (27.8 %), bringing usage back to the level of 2005-2006. The reason for the drop in usage in the 2006- 2007 financial year remains unclear, however the trend is again towards an increase in usage.

Usage in 2007 – 2008

Total pageviews	Average per month	Month of highest use	Month of lowest use
7,295,996	607,999	January 2008 1,084,499	July 2007 303,855

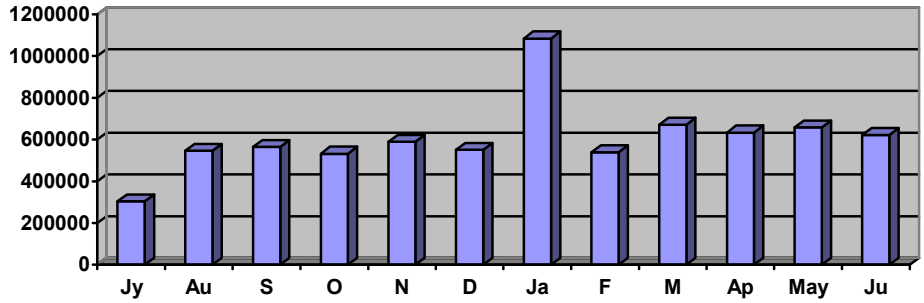
Usage in 2006 - 2007

Total pageviews	Average per month	Month of highest use	Month of lowest use
5,708,690	475,724	September 2006 650,553	February 2007 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 2005 764,662	July 2005 435,925

Month by month usage (pageviews) for July 2007 – June 2008



As another measure of Archive usage, according to Google, as at the end of the 2007-2008 financial year there are around 50,000 external links to the PANDORA website (a considerable increase over the 9,000 reported last year). Of these, around 33,000 point to the PANDORA home page.

While only 4 % of the titles archived in PANDORA have completely disappeared from the live Web, 65 % of the ten most heavily-used titles in the period July 2007 to June 2008 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 53 responses to the PANDORA User Survey questionnaire between July 2007 and June 2008. No respondents indicated this was their first visit but 60 % indicated they had only previously visited once while 26 % had visited three or more times in the past 12 months. The primary purpose for visiting the Archive remained academic or professional research (52 %), but family history was also popular (15 % of respondents most of whom were interested in the *First Families 2001* site which continued to be 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.

The National Library's collaboration with the Australian Partnership for Sustainable Repositories (APSR) at the Australian National University produced two outcomes: a prototype software product for preservation support called AONS II (Automatic Obsolescence Notification System, version 2); and the development of the Australian METS (Metadata Exchange and Transmission Standard) Profile.

AONS II is a software tool that allows users to automatically monitor the status of file formats in their repositories, make risk assessments based on a core set of obsolescence risk questions, and receive notifications when file format risks change or other related events occur. The source code was placed on SourceForge.net as open source.

The Australian METS Profile describes the rules and requirements for using METS to support the collection of, and access to, content in Australian digital repositories. This

profile has been used by the Library as well as being adopted by the National Library of Norway for .pdf newspaper collections. In December 2007 this profile was registered with the Library of Congress.

During the 2007-2008 Financial year the Digital Preservation Section has primarily been involved in designing and building a workflow solution for the most pressing preservation risks confronting its collection: digital materials on physical carriers such as handheld optical and magnetic disks. The Digital Preservation Workflow Project produced a semi-automated, scalable process for transferring data from physical carriers to the digital repository, helping to mitigate risk associated with the deterioration of media and obsolescence of the technology required to access content. This workflow system, called 'Prometheus', allows staff to process relatively standard physical media, while remaining flexible enough to accommodate special cases as required. At the end of the project, the source code was placed on SourceForge.net as open source.

6. International relations

During 2007-2008 the National Library continued its active participation in the International Internet Preservation Consortium[3] and continued as lead participant in the IIPC Working Group on Preservation. In April 2008 the National Library hosted the IIPC General Assembly, a week-long meeting of IIPC members from around the world.

7. Promoting the Archive

7.1 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.

7.2 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:

- Curtis, J., Koerbin, P., Raftos, P., Berriman, D., Hunter, J. *AONS – An obsolescence detection and notification service for Web archives and digital repositories.* A joint article by J. Curtis and P. Koerbin of the National Library, P. Raftos and D. Berriman of the ANU, and J. Hunter of the University of Queensland. Published in a special web archiving issue (vol. 13, issue 1, January 2007, pages 39-53) of New Review of Hypermedia and Multimedia. Available online at http://www.informaworld.com/smpp/content~content=a780448483~db=all~order=page

- Koerbin, P. *PANDORA : collecting in a digital world – where is the artefact?* A paper by Paul Koerbin presented at the ALIA symposium The Acquisition of Cultural Artefacts, University of South Australia, October 2007. Available online at http://pandora.nla.gov.au/documents/artifacts_symposium_pandora.pdf

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

- Koerbin, P. *A new version of the PANDORA Digital Archiving System*. A brief notice on the release of PANDAS version 3 published in Gateways, October 2007. Available online at http://www.nla.gov.au/pub/gateways/issues/89/story01.html

- Crook, E. *The 2007 Australian Federal Election on the Internet.* A paper by Edgar Crook discussing archiving issues and the use of the Internet during the 2007 federal election campaign. Available online at http://www.nla.gov.au/nla/staffpaper/2007/documents/Election2007.pdf

- Koerbin, P. *The Australian web domain harvests: a preliminary quantitative analysis of the archive data*. A report by Paul Koerbin on the size and nature of the content of the 2005, 2006 and 2007 Australian web domain harvests. Available online at http://pandora.nla.gov.au/documents/auscrawls.pdf

8. Concluding summary

Some of the highlights of 2007-2008 include:

- Content of the Archive grew by 26.1 %[4] (section 2);

- The third version of PANDAS was put fully into production providing enhanced workflows and more stable system for web archiving staff in all PANDORA participant agencies . PANDAS version 3 also provides functionality to manage the PANDORA subject listings through the user interface and a revised subject list was developed and implemented in consultation with participant agencies (section 3.1);

- The third large scale harvest of the Australian web domain was undertaken in August – September 2007. This resulted in an archival collection of 516 million files amounting to 18.47 terabytes of data. The combined total for all three Australian domain harvests has now reached 1.3 billion files amounting to 44.2 terabytes of data (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 and produced significant outcomes in digital preservation including the AONS II software, the Australian METS Profile and a Digital Preservation Workflow system (section 5); and,

- Active participation in the International Internet Preservation Consortium continued with the National Library leading IIPC Preservation Working Group and the hosting of the 2008 IIPC General Assembly in April 2008 (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