Annual report to partners 2008-2009

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 2008-2009 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 made available 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 2008	30 June 2009	Growth 2008-09
Titles	19,308	22,464	3,156 (16.34%)
Instances	38,183	46,591	8,408 (22.02%)
Terabytes [2]	2.19	3.02	0.83 (37.89%)
Usage (page views)	7,295,996	3,861,089	-3,434,907 (-52.92%)

The Archive continued to show content growth during 2008-2009, with good percentage increases for the number of titles and number of instances. The size of the Archive is now more than 3 terabytes of data. The continued very strong growth in data size is due in part to the amount of large multimedia files that now being collected with many websites.

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 each subsequent gathering.

2 This figure does not include the preservation and other master and back up copies.

The apparent significant decrease in the number of page views is due to changes made in September 2008 to the way page views across all NLA web services are counted. Accurate trends should be able to be seen later in 2009.

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 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 collection of Australian web resources.

3.1 Development of PANDAS

PANDAS (the PANDORA Digital Archiving System) is the web-based workflow management system 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 2008-2009 the Library completed some minor improvements to PANDAS, primarily in the delivery of reports.

During 2008-2009 the Library moved PANDORA onto Ruby on Rails a new platform to improve the speed of delivery of content to users and allow for other enhancements. One of these was to allow for a new search trend functionality that can be viewed at http://pandora.nla.gov.au/search-trends/.

Another enhancement was the addition of a number of new subject headings in the PANDORA database to improve user browsing of the Archive. For three months the Library employed a graduate recruit to investigate ceased titles and assign the new subject headings to them. A report on all ceased titles for each agency and instructions on how they could best amend them was distributed to all agencies.

The National Library continues to develop innovative search capabilities and in May 2009 made available the prototype of a new single business discovery service http://sbdsproto.nla.gov.au. This service offers the potential for improved PANDORA search functionality, including the ability to find archived web content along with other formats in a single search.

3.2 Whole domain harvests

In the second half of 2008 the Library conducted the fourth and most ambitious of its annual large scale harvests of the Australian web domain. Despite the advantages of the selective approach to archiving, its shortcomings 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 earlier harvests, in 2008 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 fourth annual harvest had the target of collecting 1 billion 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.

The harvest was conducted between 18 July and 22 September 2008 and 1,000,618,888 unique documents were captured amounting to 34.55 terabytes of data. The Internet Archive indexed the contents of the harvest and shipped it to the National Library in late 2008. The combined total figures for Australian web content captured by the four Australian domain harvests between 2005 and 2008 equates to 2.3 billion files and 78.75 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 collections remains limited and they are not currently available to the general public. Unlike the selective Archive, we are not able to negotiate prior permission individually with publishers to provide access to the archived contents.

The Library is preparing to undertake the fifth annual whole domain harvest commencing in September 2009.

4. Focus on users

Once again this year an analysis of usage of the Archive during the previous financial years, 2006-2007 and 2007-2008, was undertaken.

The analysis showed an apparent large drop in usage during the 2008-2009 financial year over previous years. The reason for this is due to a change in reporting mechanisms, thus it is not clear whether outside of this reporting change there has been a real drop in usage or not.

Usage in 2008 - 2009

Total pageviews	Average per month	Month of highest use	Month of lowest use
3,861,089	321,757	September – 516,286	December – 249,755

Usage in 2007 - 2008

Total pageviews	Average per month	Month of highest use	Month of lowest use
7,295,996	607,999	January - 1,084,499	July -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 – 650,553	February – 375,962

As another measure of Archive usage, according to Google, as at the end of the 2008-2009 financial year there are over 14,000 external links to the PANDORA website of which around 17% point to the PANDORA home page.

While only 4 per cent of the titles archived in PANDORA have completely disappeared from the live Web, 64% (between six and seven) of the ten most heavily-used titles in the period July 2008 to June 2009 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 24 responses to the PANDORA User Survey questionnaire between July 2008 and June 2009. This small and declining number of responses from such a large number of users cannot reasonably be used to indicate the wider PANDORA usage. The survey did indicate however that for those who answered the survey the primary purpose for visiting the Archive was academic or professional research (62%). Family history also remains popular (12%) although this is slightly less than the percentage reported from last year's survey results.

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 closely with preservation staff in the Web Archiving and Digital Preservation Branch further analysis of options for the future display of web archive content are being investigated. A discrete copy of the presentation copy of PANDORA has been created and testing and analysis of content and file types is being undertaken by digital preservation staff. It is understood that all content previously archived is still accessible with current browsers and plug-ins.

6. International relations

During 2008-2009 the Library continued its active participation in the International Internet Preservation Consortium (IIPC).

Paul Koerbin was a founding owner of a new IIPC mail list for international web curators at http://netpreserve.org/about/curator.php

During 2008-2009 Colin Web led the IIPC Preservation Working Group (PWG) and was a member of the IIPC Steering Committee.

During 2008-2009 IIPC meetings were held in:

- Aarhus, Denmark, September 2008. Colin Webb prepared an agenda, notes and chairing arrangements for a one day brainstorming session on preservation issues and undertook follow up work but did not attend the meeting in person.

- London, UK, 2 October 2008. This meeting was held in conjunction with the IPRES 2009 conference. Colin Webb travelled to London and led a one day meeting of the Preservation Working Group (PWG) and participated in a Steering Committee meeting. The PWG established a prioritised work program for the next 12 months. A number of work packages were proposed with the Library to participate in a small number and Colin to monitor progress of all. In addition to attending the IPRES conference, Colin also participated in an international meeting on the future of the GDFR (Global Digital Format Registry) and agreed to lead development of a new governance model for a unified global format registry (UDFR).

- Ottawa, Canada, May 4-7, 2009. The IIPC General Assembly. Colin Webb participated in the PWG meeting by telephone/web link up giving presentations and reporting on two work packages that the Library is leading on preservation strategies and an environmental scan.

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. An updated version of the PANDORA Brochure was also printed in May 2009. The Fact Sheet is available online at: http://pandora.nla.gov.au/overview.html#factsheet

7.2 Papers and articles

A small number of papers and articles were published or presented during the year, including:

- A presentation by Paul Koerbin on web archiving and copyright issues for the WIPO International Workshop on Digital Preservation and Copyright held at the World Intellectual Property Organization in Geneva on 15 July 2008.

- *Web Archiving in a Web 2.0 World.* A paper written and presented by Edgar Crook at the ALIA Biennial Conference, Alice Springs, on 2 September 2008, available online at http://pandora.nla.gov.au/pan/13910/20080930-1156/conferences.alia.org.au/alia2008/pdfs/124.TT.pdf

- A presentation by Paul Koerbin titled *Web Archiving: Another Dimension of Cultural Acquisition* at the ALIA Acquisition Seminar in Canberra on 17 October 2008

- An opinion piece written by Paul Koerbin for the ABC News Opinion and Analysis website in May 2009 available at: http://www.abc.net.au/news/stories/2009/05/07/2562951.htm

- A book chapter titled 'Issues in business planning for archival collections of web materials' written by Paul Koerbin for a forthcoming book on business planning for digital libraries edited by Mel Collier for the Leuven University Press.

8. Review of Web Archiving

During the 2008-2009 financial year, Paul Koerbin completed a review of web archiving at the National Library. The final 45 page report of the review included 30 recommendations. Recommendations were broadly focused on:

- Collecting principles and scope: recommending a move away from detailed selection guidelines in favour of a more dynamic curatorial approach based on concisely stated policy and collection intentions that are aligned with broader Library objectives.

- Collecting methods: recommending the use of a variety of collecting methods including selective, domain and seed-list based methods to achieve best outcomes for specific materials. Timeliness is seen as a critical driving factor in aligning collection methods to target materials.

- Technology and tools: recommending that resources and effort should be put into infrastructure to allow experimentation, trialling and deployment of a broader range of technologies and tools (including IIPC supported tools and standards) to support a broader methodological base.

- Collaboration: recommending a closer working relationship among areas within the Library with a curatorial stake in web materials; and investigating the feasibility of engaging communities of interest outside the Library (known and 'crowdsourced'), at least in nominating and developing seed lists and possibly extending to the deposit of web materials.

- Access: recommending that the Library formulate and provide a clear public position on access objectives for the archived collections; and investigating ways of enhancing access and looking for opportunities to develop innovative ways in which the archival content can be discovered.

9. Concluding summary

Some of the highlights of 2008-2009 include:

- Continued strong growth of the Archive (section 2);

- Completion of the *Review of Web Archiving* report [section 8).

- Completions of the largest single harvest of the Australian web domain, undertaken in the second half of 2008 and resulting in an archival collection of over 1 billion files (section 3.2).

- Maintaining an active role in the International Internet Preservation Consortium continued with the National Library continuing to lead the IIPC Preservation Working Group (section 6).