Annual report to partners 2013-2014

Contents

1. *PANDORA participants working together*

PANDORA, Australia's Web Archive (http://pandora.nla.gov.au/) is a selective archive of Australian online publications and websites which is built collaboratively by the National Library of Australia, all of the mainland state libraries, the Northern Territory Library, the Australian War Memorial, the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) and the National Gallery of Australia. This is a report to contributing partners on activities and developments in the 2013-2014 financial year.

1.1 Consultation mechanisms

The National Library continued to inform other PANDORA participants about the operation of PANDORA through the two email discussion lists, the PANDORA Wiki and a fortnightly newsletter distributed through email and the Wiki.

1.2 Reports

Each month, a report on the growth of the Archive and usage statistics is sent to the email discussion list. 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 National 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 website 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.

A report (this report) on progress, activities and trends to the Chief Executive Officers of partner agencies is prepared annually and is also made available on the PANDORA website Partners page http://PANDORA.nla.gov.au/partners.html.

1.3 Collaborative collecting

Two major collaborative collections were undertaken jointly by PANDORA partners in 2013-2014. The September 2014 federal election campaign collection consisting of more than 560 websites; and the ANZAC Centenary collection, currently with over 30 websites and which is projected to continue growing as partners select and add new websites over the next four years of commemoration of the 1914-18 World War. These collaborative collections can be publicly accessed at the addresses indicated below. The federal election campaign being a large collection is split over six sub collections covering candidates, lobby groups, media, political parties and research sites as well as the subsequent Western Australian Senate re-election.

- ANZAC Centenary
  http://pandora.nla.gov.au/col/11341

[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.

- 2013 Australian federal election campaign

  http://pandora.nla.gov.au/col/12283
  http://pandora.nla.gov.au/col/12291
  http://pandora.nla.gov.au/col/12299
  http://pandora.nla.gov.au/col/12300
  http://pandora.nla.gov.au/col/12301
  http://pandora.nla.gov.au/col/12302
  http://pandora.nla.gov.au/col/12862

2. *Growth of the Archive*

2.1 Size and annual growth of the PANDORA Archive

The PANDORA Archive maintained steady growth in 2013-2014, with the percentage growth rate for Titles and Instances of a similar magnitude to the previous financial year; while the amount of data collected, measured in terabytes, continues to increase growing 37% this financial year compared with 28.7% last financial year. Again, the growth rate in data size is the standout, highlighting the increasing complexity and size of many of the websites being collected by some agencies.

	30 June 2013	30 June 2014	Growth 2013-2014
Titles	34,694	38,535	3,841 (11%)
Instances	86,977	99,390	12,143 (14%)
Terabytes[2]	8.74	11.94	3.2 (37%)

Government publications remain a substantial component of the collecting focus and comprise approximately 55 % of the titles in the Archive.

2.2 Statistics for annual partner contributions

The following chart shows the contribution to PANDORA of each participating agency for 2013-2014 and the previous financial year for comparison. The contributions are measured by the number of titles archived, the number of instances archived and what this constitutes in the number of files and data size measured in gigabytes. In order to make the chart more useful it has been sorted based on the contribution of Instances archived.

The statistics record contributions by title, actual archived instances (for which there could be multiple for a single title), files and data size. It is possible to discern different approaches from different agencies, for example some agencies have a close match between titles and instances reflecting one-off harvests or long schedules (e.g. annual) for repeat harvests. Other agencies do a larger proportion of re-harvesting of titles during the year. The relationship between instances and data size shows some agencies doing a larger number of smaller harvests while the average instances size is around 200MB. The third chart shows the percentage variation from the previous financial for each agency for each measure, most notably indicating an across-the-board increase in the size of the instances archived.

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

2013-2014 financial year contributions by partner agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	3,759	6,167	45,085,336	2211.84
State Library of Victoria	2,078	2,825	5,880,273	408.70
State Library of NSW	887	1,183	2,551,158	166.07
State Library of Queensland	890	979	4,218,612	218.71
State Library of SA	513	655	3,482,975	184.10
State Library of WA	196	351	465,847	27.28
AIATSIS	62	65	182,768	18.93
National Gallery of Australia	56	59	298,310	9.82
Australian War Memorial	30	33	88,966	7.35
NFS Archive[3]	5	5	143,980	2.20
Northern Territory Library	2	2	4,821	0.97

2012-2013 financial year contributions by partner agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	2,986	4,874	29,967,158	1351.68
State Library of Victoria	1,789	2,153	5,026,534	283.11
State Library of NSW	987	1,335	1,487,100	83.41
State Library of Queensland	795	976	1,712,218	98.37
State Library of SA	456	655	1,685,180	96.70
State Library of WA	225	314	449,237	27.42
NFS Archive	61	63	427,050	26.20
AIATSIS	56	59	210,269	15.45
National Gallery of Australia	40	40	78,481	3.58
Australian War Memorial	23	24	35,396	3.71
Northern Territory Library	18	18	53,684	4.94

Percentage change between 2012-2013 and 2013-2014 financial years

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	26%	27%	50%	64%
State Library of Victoria	16%	31%	17%	44%
State Library of NSW	-10%	-11%	72%	99%
State Library of Queensland	12%	0%	146%	122%
State Library of SA	13%	0%	107%	90%
State Library of WA	-13%	12%	4%	-1%
NFS Archive	-92%	-92%	-66%	-92%
AIATSIS	11%	10%	-13%	23%
National Gallery of Australia	40%	48%	280%	174%
Australian War Memorial	30%	38%	151%	98%
Northern Territory Library	-89%	-89%	-91%	-80%

[3] The National film and Sound Archive ceased contributing as a PANDORA Partner during the course of the 2013-2014 financial year.

3. *Development of the Web Archive*

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

3.1 Development of PANDAS

PANDAS (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 websites to PANDORA. This does not include cataloguing, which is carried out in separate local systems.

No major development of PANDAS is currently being undertaken as the Library plans to redevelop its web archiving systems as part of its ongoing Digital Library Infrastructure Replacement Project. However, in 2013-2014, some minor enhancements were made to improve workflow efficiencies.

An enhancement implemented was an instance history display that shows details about previous harvests including the dates of gathering, processing and archiving/deleting, together with operator details and the size of harvests.

Another enhancement was the incorporation of a permission request system which allows curators to create, edit and send emails directly from PANDAS. Currently this is only functional for the National Library curators.

3.2 Australian web domain harvest

In the first quarter of 2014 the Library conducted the ninth large scale harvest of the Australian web domain.

As with the previous harvests conducted annually since 2005 the National Library contracted the Internet Archive to undertake the whole domain harvest crawl. The Internet Archive has extensive experience in this form of web archiving.

The harvest was run during February 2014 and around 953 million unique documents were captured, amounting to 31.93 terabytes of data from around seven million hosts. Following this harvest the combined total for all nine Australian domain harvests has now reached 6.3 billion files amounting to around 237 terabytes of data.

The following table shows the amount of content collected for each of the domain harvests conducted to date.

Domain Harvest	Unique files	Hosts	Size (TB)
2005	185 m	811,523	6.69
2006	596 m	1,046,038	19.04
2007	516 m	1,247,614	18.47
2008	1 billion	3,038,658	34.55
2009	756 m	1,074,645	24.28
2011	660 m	1,346,549	30.71
2012	1 billion	1,467,158	41.88
2013	660 m	1,690,232	29.17
2014	953 m	7,046,168	31.93

In the absence of legal deposit provisions for online publications and websites at the Commonwealth level, the access that the National Library can provide to the whole domain harvest remains limited and they are not currently available to the general public. Unlike the selective Archive, we have not been able to negotiate prior permission individually with publishers to provide access to the collected content.

3.3 Collecting Commonwealth Government online publications – the 'Australian Government Web Archive'

In March 2014 the Library released a production version of the Australian Government Web Archive (AGWA). The AGWA provides access to freely available Commonwealth Government web materials collected by the Library through bulk harvesting processes. The AGWA is a distinct and separate web archive from PANDORA. Initially content for AGWA was harvested by the Internet Archive as an additional collection to the annual domain harvests. The first such collection was undertaken in early 2011 and new content has been progressively added from other Internet Archive harvests and through archival harvests undertaken in-house at the Library. We are also in the process of adding historic archival content to the AGWA. This includes content from earlier whole domain harvests covering the period 2005 onwards.

The AGWA application, designed to deliver bulk harvested content collected using the Heritrix archival web harvester, was developed within the web archiving operational unit by Dr Mark Pearson. In part the purpose of this development was to enable the Library to gain experience working with web archiving infrastructure for enterprise scale collecting using the Heritrix crawler, the Wayback URL indexing and delivery system and SOLR full text indexing. This exercise was undertaken in part also as an important preliminary project to further development of the web archiving infrastructure and systems as part of the Digital Library Infrastructure Replacement Project.

Currently the AGWA stands outside the Library's main discovery service Trove and can be found at the following location: http://webarchive.nla.gov.au/gov/

4. *Focus on users*

4.1 User page views of the PANDORA Archive

Web usage statistics PANDORA are available from the Library's website at:
http://stats.nla.gov.au/cgi-bin/report_index.cgi?report=PANDORA

Usage in 2013 – 2014

Total page views	Average per month	Month of highest use	Month of lowest use
82,516,835	**6,876,403**	**10,037,646 (Feb. 2014)**	**4,983,475 (March 2014)**

4.2 Most viewed titles (websites) in the PANDORA Archive

Around 7 % of the titles archived in PANDORA are recorded in PANDAS as being no longer online at the original 'live' site. Since this figure relies on curators recording this fact, the actual figure is probably somewhat higher; and even sites that are still 'live' may not continue to include content that was harvested earlier for the Archive. A high percentage of the most used sites in PANDORA are ones that are no longer available as live websites. The table below shows the top 10 sites accessed in 2013-2014.

Archived Title	Partner Responsible	Live site	Page views
First families 2001	SLV	No	362,422
cultureandrecreation.gov.au	NLA	No	302,826
Sydney Centre for Studies in Caodaism	NLA	Yes	284,308
Digger history	AWM	No	215,451
Nova : science in the news	NLA	Yes	210,594
Life on the goldfields	SLV	No	190,929
Centenary of Federation	NLA	No	144,780
Tony Abbott – Liberal for Warringah	NLA	No	143,390
Federal Minister for Fisheries, Forestry	NLA	No	126,190
Footypedia	NLA	No	102,143

5. *Promoting the Archive*

5.1 Publications and public presentations

Presentations given and papers published by National Library Web Archiving staff during the 2013-2014 financial year included the following:

- *The Australian Government Web Archive*. A presentation by Dr Paul Koerbin at the Australian Government's Office of the Australian Information Commissioner's 'Information Contact Officer Meeting', 16 May 2014, Canberra. Available at: http://www.oaic.gov.au/about-us/working-with-others/information-contact-officer-network/icon-meeting-friday-16-may

5.2 Social media and the web archiving blog

During 2013-2014 the following web archiving blog posts were published:

- *Web archiving in a fast moving world* by Russell Latham, 2 July 2013
- *Archiving online election campaigns* by Paul Koerbin, 9 August 2013
- *Web archiving – an antidote to 'present shock'?* by Paul Koerbin, 18 March 2014
- *To know, to utter, to argue … and to archive and access* by Paul Koerbin, 27 May 2014
- *2013 Federal Election web collecting* by Russell Latham, 11 June 2014

In April 2014 the Library established the @NLAPandora Twitter account to allow the PANDORA managers (Paul Koerbin and Russell Latham) to tweet about the Library's web archiving activities and to engage with users directly through social media. At the time of writing @NLAPandora had posted more than 160 tweets and had around 250 followers.

5.3 Presentations to visitors to the National Library

The National Library regularly hosts visitors from other libraries and organisations. Presentations on PANDORA, web archiving and PANDAS were provided to visitors to the Library during 2013-2014, including:

- A large delegation of librarians from Indonesia;
- Digital collecting staff from the National Library of New Zealand;
- Kirsten McCormick, a researcher on a visiting fellowship from Scotland;
- A delegation from Charles Sturt University;
- Amanda Lawrence from the Grey Literature Strategies Project; and
- Isaac Gillman, Associate Professor and Scholarly Communication & Publishing Services Librarian at Pacific University, Oregon, USA, visiting Australia on an Endeavour Fellowship.

6. *Concluding summary*

Some of the highlights of 2013-2014 include:

- Continuing steady growth of the Archive content at 11% for titles, 14% for archived instances and 37% in the growth of the data (section 2.1).
- Completion of the 2014 large scale harvest of the Australian web domain, the ninth such bulk collection of .au web content since 2005 (section 3.2).
- The implementation of a new web archive for Australian Commonwealth Government websites called The Australian Government Web Archive (section 3.3).
- Continued engagement through social media and the web archiving blog (section 5.2).