Annual report to partners 2014-2015

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 participants on activities and developments in the 2014-2015 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 regular 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 participant 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 participants are advised of their availability via a message to the two email discussion lists.

This report on progress, activities and trends to the Chief Executive Officers of participant 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

PANDORA participant agencies have contributed to a number of collaborative collections in 2014-2015, including:

- ANZAC Centenary Collection

  This collection consists of more than 100 titles contributed by almost all agencies. It contains a number of sub-category collections, including: Media, Veterans' Organisations and Community websites.

- Personal stories of Australian's at war

[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 It includes the gathering of a monograph that has been archived once only, the first gathering of a serial title or integrating title (for example, a web site that changes over time), and all subsequent gatherings.

This collection complements the ANZAC Centenary Collection. The websites archived in the collection typically include transcripts of diaries or letters from those who went to war.

- G20 Australia (2014)

This collection is a specific collaboration between the State Library of Queensland and the National Library to capture material relating to the G20 Summit held in Brisbane in November 2014.

2. *Growth of the Archive*

2.1 Size and annual growth of the PANDORA Archive

The PANDORA Archive maintained steady growth in 2014-2015. The percentage growth rate for Titles and Instances was of a similar magnitude to the previous financial year being down just 1%; while the amount of data collected, measured in terabytes, continues to increase growing 35% this financial year compared with 37% 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 2015	30 June 2014	Growth 2014-2015
Titles	42,424	38,535	(10 %)
Instances	112,652	99,390	(13 %)
Terabytes[2]	16.14	11.94	(35 %)

Government publications remain a substantial component of the collecting focus and comprise approximately 53 % of the titles in the Archive.

2.2 Statistics for annual participant contributions

The following chart shows the contribution to PANDORA of each participating agency for 2014-2015 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 collected this financial year is 330 MB up from 200MB last year. The third chart shows the percentage variation from the previous financial year 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.

2014-2015 financial year contributions by participant agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	3,960	6,052	47,044,799	2785.28
State Library of Victoria	2,199	3,475	6,722,843	448.29
State Library of NSW	945	1,374	3,693,470	301.33
State Library of Queensland	1,188	1,351	8,025,453	431.97
State Library of SA	447	481	2,834,190	182.84
State Library of WA	205	266	1,125,744	53.51
National Gallery of Australia	79	80	817,503	24.53
Australian War Memorial	58	64	690,963	30.30
AIATSIS	50	51	346,846	18.46
Northern Territory Library	2	2	1,907	0.23

2013-2014 financial year contributions by participant 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
Northern Territory Library	2	2	4,821	0.97

Percentage change between 2013-2014 and 2014-2015 financial years

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	5%	-2%	4%	26%
State Library of Victoria	6%	23%	14%	10%
State Library of NSW	7%	16%	45%	81%
State Library of Queensland	33%	38%	90%	98%
State Library of SA	-13%	-27%	-19%	-1%
State Library of WA	5%	-24%	142%	96%
National Gallery of Australia	41%	36%	174%	150%
Australian War Memorial	93%	94%	677%	312%
AIATSIS	-19%	-22%	90%	-2%
Northern Territory Library	0%	0%	-60%	-76%

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.

In the second quarter of 2015 planning commenced in respect to some minor changes to PANDAS that are required to enable a smooth transition to incorporate workflows to accommodate the extension of legal deposit to electronic materials that became law in July 2015 and which will come into force in February 2016.

Most of the web archive development work in 2014-2015 focused on the Australian Government Web Archive (see section 3.3. below).

3.2 Australian web domain harvest

In the first quarter of 2015 the Library conducted the tenth 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 March and April 2015 and around 566 million unique documents were captured, amounting to 42 terabytes of data from around two and a half million hosts.

In addition to the 2015 domain harvest an extract of the .au data for the years 1996 to 2004 collected by the Internet Archive was also acquired. This means that the Library's Australian (.au) domain harvest collection now include snapshots covering the entire period from 1996 to the present. The 1996-2004 data content amounts to around 448 million files or 6.7 terabytes of data.

Following this harvest the combined total for all ten Australian domain harvests has now reached seven (7) billion files amounting to around 326 terabytes of data.

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

Domain Harvest	Unique files	Hosts crawled	Size (TB)
2005	185 m	811,523	8.0
2006	596 m	1,046,038	21.3
2007	516 m	1,247,614	20.5
2008	1 billion	3,038,658	39.5
2009	756 m	1,074,645	34.8
2011	660 m	1,346,549	35.2
2012	1 billion	1,467,158	47.1
2013	660 m	1,690,232	43.7
2014	953 m	7,046,168	27.7
2015	566m	2,580,521	42.1

NB: data has been amended since last year's reporting to reflect raw uncompressed data rather than the previously reported compressed data.

Content from the Australian domain harvests is not currently made available to the public with the exception of Commonwealth Government websites which are accessible through the Australian Government Web Archive.

3.3 Collecting Commonwealth Government online publications

Substantial content has been added to the Library's second web archive service, the Australian Government Web Archive (AGWA), over the past year. This includes a number of harvests run 'in-house' as well as content extracted from the Australian domain harvests supplied by the Internet Archive. This means that content accessible through the AGWA now covers the period 1996 to 2015. Currently around 145 million files or 15 terabytes of data is delivered through the AGWA.

Work has progressed on developing a dashboard to manage some of the AGWA workflows. While harvests are still run using the native Heritrix console and XML configuration file, the AGWA dashboard – known as 'Bamboo' – supports simple curator workflows to index and add harvested content to the production service. Work is also underway to provide Library staff with a 'one-click' facility to collect simple PDF documents that can be added in real-time to the AGWA collection.

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

Usage in 2014 – 2015

Total page views	Average per month	Month of highest use	Month of lowest use
84,773,561	7,064,463	9,801,045 (June 2015)	4,945,533 (Feb. 2015)

There was a 2.73% increase in page views in 2014-2015 over the previous year.

4.2 Most viewed titles (websites) in the PANDORA Archive

Around 7.5 % 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 20 sites accessed in 2014-2015.

	Archived Title	Participant Responsible	Live site	Page views
1	**Electoral Commission SA**	SLSA	Yes	635,596
2	**First families 2001**	SLV	No	371,555
3	**Sydney Centre for Studies in Caodaism**	NLA	Yes	271,302
4	**Life on the goldfields**	SLV	No	228,406
5	**Digger history**	AWM	No	214,019
6	**Antipodean SF**	NLA	No	200,215
7	**Australia in the Asian century**	NLA	No	164,155
8	**cultureandrecreation.gov.au**	NLA	No	159,309
9	**Gravesecrets at your fingertips! Cemeteries**	SLSA	Yes	155,321
10	**Footypedia**	NLA	No	155,289
11	**Publications – Commission for Children and Young People and Child Guardian**	NLA	No	143,713
12	**The Spirits of Gallipoli**	NLA	Yes	140,753
13	**ARIA report**	SLNSW	Yes	119,638
14	**State Library of NSW website**	SLNSW	Yes	118,964
15	**National ANZAC Centre**	SLWA	Yes	108,103
16	**Nova : science in the news**	NLA	No	108,074
17	**Centenary of Federation**	NLA	No	102,828
18	**Digital switchover : are you ready for digital TV?**	NLA	No	101,672
19	**South Land to New Holland : Dutch charting of Australia 1606-1756**	NLA	No	98,328
20	**Canberra Roller Derby League**	NLA	Yes	96,508

5. *Promoting the Archive*

5.1 Publications and public presentations

Presentations given and papers published by National Library Web Archiving staff during the 2014-2015 financial year included the following:

- Russell Latham gave a presentation to the 2014 ALIA National Conference entitled: 'The online campaign: building the 2013 Australian Federal Election collection'. This peer reviewed paper was also published as a NLA staff paper and is available at: http://www.nla.gov.au/our-publications/staff-papers/the-online-campaign-building-the-2013-australian-federal-election.

- Alison Dellit gave a presentation to the 2014 ALIA National Conference entitled: 'Web Archiving – Australian Government Web Archive'.

- Paul Koerbin gave a lecture on government web archiving to Australian Public Service ICT Graduates at the University of Canberra in June 2015.

5.2 Media, social media and the web archiving blog

In May 2015 Paul Koerbin was interviewed for the ABC's *Future Tense* program titled 'Archaeology and artefacts: current threats, future possibilities'. The program was broadcast on 26 July 2015, see:
http://www.abc.net.au/radionational/programs/futuretense/archaeology3a-current-threats2c-future-possibilities/6633840

During 2014-2015 the following web archiving blog posts were published:

- *Archiving Floriade – preserving ephemeral events* by Paul Koerbin, 25 September 2014

- *The Australian Government Web Archive* by Paul Koerbin, 11 February 2015

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 370 tweets and had around 550 followers.

5.3 Presentations to visitors to the National Library

The National Library regularly hosts visitors from other libraries and organisations. Formal presentations on PANDORA, web archiving and PANDAS were provided to visitors to the Library during 2014-2015, including:

- A delegation of librarians from Indonesia (25 Nov. 2014)

6. *Concluding summary*

Some of the highlights of 2014-2015 include:

- Continuing steady growth of the Archive content at 10% for titles, 13% for archived instances and 35% in the growth of the data (section 2.1).
- Completion of the 2015 large scale harvest of the Australian web domain, the tenth such bulk collection of .au web content since 2005 (section 3.2).
- A data extract of .au domain from Internet Archive were acquired for the years 1996-2004 (section 3.2).
- Promoting PANDORA and the Australian Government Web Archive at conferences within Australia (section 5.1).
- Continued engagement through social media and the web archiving blog (section 5.2).