PANDORA
AUSTRALIA'S WEB ARCHIVE
National Library of Australia and Partners

Annual report to partners 2016-2017

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 Australian War Memorial, the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) and the National Gallery of Australia. This report to contributing participants on activities and developments in the 2016-2017 financial year is provided in accordance with the National Library's obligation as stated in section 6.2 (k) of the Memorandum of Understand with participant agencies.

1.1 Consultation mechanisms

The National Library continued to inform other PANDORA participants about the operation of PANDORA through an email discussion list, the PANDORA Wiki and a semi-regular newsletter distributed through email and the Wiki.

1.2 Reports

Each month, the Library distributes a report on the growth of the Archive and usage statistics to the PANDORA email discussion list. This report includes a list of the ten most popular (most viewed) sites for the month and which agency is responsible for the selection.

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. The Library publishes these lists 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 email discussion list.

This report on progress, activities and trends to the Chief Executive Officers of active participant agencies is prepared annually. It is made available on the PANDORA website partners page http://PANDORA.nla.gov.au/partners.html where it can be viewed along with all previous reports from 2004-2005.

1.3 Adding value – notable collections

A number of collections were developed, formed or extended during the 2016-2017 adding value through the curation of selected content. Notable collections worked on during the year include:

- 2016 Federal Election Campaign

  This collection was built collaborative by PANDORA participants to capture material for the July 2016 federal election campaign. The collection consists of six sub-collections covering candidates, political party, media, interest and lobby group and electoral research websites. Altogether, around 857 websites were archived.

- 2016 Rio Olympic Games

[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 includes 33 websites documenting Australian involvement in the Games, covering advertising, media, Olympians, Paralympians and official team and committee websites.

- Australian Broadcasting Corporation (ABC)

  This collection brings together nearly 300 web pages, mini-sites etc. from the ABC collected over a number of years. It includes material related to the arts, news and politics, radio and television shows, sport and feature items; as well as ABC publications and web pages commemorating people and events.

- The Library now collects a number of online news sites daily including:
  - The Sydney Morning Herald (since June 2009)
  - The Conversation (since May 2013)
  - The Guardian Australia (since February 2016)
  - The Canberra Times (since February 2016)
  - News.com.au (since February 2016)
  - Huffington Post Australia (since March 2016)
  - The Age (since May 2016)
  - The Australian Financial Review (since March 2017)
  - ABC News (since May 2017)

2. *Growth of the Archive*

2.1 Size and annual growth of the PANDORA Archive

The PANDORA Archive maintained a consistent high level of growth in 2016-2017 particularly when measured by data collected. The percentage growth rate for Titles was slightly down on the previous year (by about 2 %) and Instances was of a similar magnitude at around 14 % growth. The amount of data collected, measured in terabytes, continues to increase growing at nearly 41 % this financial year compared with 39% last financial year.

	30 June 2017	30 June 2016	Growth 2016-2017
Titles	50,605	46,911	(7.9 %)
Instances	147,399	128,842	(14.4 %)
Terabytes	31.57	22.43	(40.7 %)

Government publications remain a substantial component of the collecting focus and currently comprise approximately 49 % of the titles in the Archive. In the 2016-2017 financial year 37% of new titles registered were government titles. The lower percentage than the historic average for collecting government publications is most probably because the National Library is increasingly using the Australian Government Web Archive and its new ‘eDeposit’ service to collect Commonwealth Government web and digital material.

2.2 Statistics for annual participant contributions

The first two charts shows the contribution to PANDORA of each participating agency for the current and previous financial years for comparison. The contributions are measured by the number of titles archived, the number of instances archived, the number of files collected and data size measured in gigabytes. The charts are arranged in order based on the contribution of Instances archived in the current financial year.

The charts suggest different approaches to collecting by participant 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 as shown by the difference between titles and instances. The relationship between instances and data size shows some agencies are doing a larger number of smaller harvests; while the average instances size collected this financial year is 517 MB up from 431 MB 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.

2016-2017 financial year contributions by participant agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	5,545	10,521	115,336,355	6737.92
State Library of Victoria	2,360	3,687	9,158,934	647.64
State Library of Queensland	1,539	1,644	8,900,235	1095.68
State Library of NSW	751	1,238	4,664,175	388.89
State Library of SA	579	793	4,557,938	319.72
State Library of WA	272	354	654,467	42.12
National Gallery of Australia	102	109	547,356	35.17
Australian War Memorial	45	47	699,125	30.91
AIATSIS	29	33	139,516	16.79
Northern Territory Library*	13	13	12,823	2.05

*Harvests for the NTL were completed by the NLA as the NTL currently remains an inactive participant.

2015-2016 (previous) financial year contributions by participant agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	5,230	8,733	80,567,376	4833.28
State Library of Victoria	2,171	3,181	8,609,986	556.53
State Library of Queensland	1,555	1,714	8,678,057	543.85
State Library of NSW	841	1,278	5,068,457	455.42
State Library of SA	520	627	4,051,210	256.29
State Library of WA	199	253	481,711	33.65
National Gallery of Australia	95	104	454,521	27.52
Australian War Memorial	35	39	438,546	16.45
AIATSIS	41	57	1,582,945	23.49
Northern Territory Library	17	17	85,025	4.82

Percentage change in contributions between 2015-2016 and 2016-2017 financial years

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	6%	20%	43%	39%
State Library of Victoria	9%	16%	6%	16%
State Library of Queensland	-1%	-4%	3%	101%
State Library of NSW	-11%	-3%	-8%	-15%
State Library of SA	11%	26%	13%	25%
State Library of WA	37%	40%	36%	25%
National Gallery of Australia	7%	5%	20%	28%
Australian War Memorial	29%	21%	59%	88%
AIATSIS	-29%	-42%	-91%	-29%
Northern Territory Library	-24%	-24%	-85%	-57%

3. *Development of the Web Archive*

The National Library is committed to the ongoing development of the policy, procedures and technical infrastructure that support both the collection of Australian web resources and improves the discovery and delivery of the web archive content.

The focus of web archiving development over the 2016-2017 financial year has been on improving access to the web archive collections through the development of a new discovery and delivery system (see 3.1). There has been no development of the collecting infrastructure, including PANDAS (the PANDORA Digital Archiving System).

3.1 Development of the 'Trove web archive' zone

The final stage of the Library's Digital Library Infrastructure Redevelopment project (which concluded in June 2017) focused on developing a new infrastructure that will allow the Library to provide access to all its web archive collections, including PANDORA, AGWA and the large domain harvests, through a single search interface under Trove. In addition, this development work included a new interface for the delivery of web archive content that will replace both the PANDORA and the AGWA delivery systems.

Work completed by the end of June 2017 includes:

- Building a new delivery interface that will appear under Trove as the archived websites zone (replacing the existing PANDORA archived websites zone);
- Development of the backend infrastructure for management of the different web archive collections so as to allow them to be indexed and delivered as a single collection;
- Indexing the vast amount of content that forms the domain harvest collections (i.e. more than 4 billion text documents full-text indexed);
- Building a new access control tool to manage both access and discovery restrictions (together with the migration of PANDORA restrictions to the new tool);
- Undertaking initial user acceptance testing of the delivery system; and,
- Work to identify, assess and document risks associated with releasing the whole domain content.

More work remains to be done before the new delivery and discovery system can be released into production. This includes further user testing, bug fixing and index enhancements to mitigate risk associated with releasing domain harvest content.

3.2 Australian web domain harvest

In the first quarter of 2017 the Library conducted the 12th large-scale harvest of the Australian web domain. This was the second Australian domain harvest conducted since of legal deposit legislation was extended to online electronic material in February 2016.

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 large scale web archiving.

The harvest was run during February and March 2017 and more than 900 million unique documents were captured, amounting to 62 terabytes of data from more than four million hosts.

Following this harvest, the combined total for all 12 Australian domain harvests has now reached nearly 9 billion files amounting to around 450 terabytes of data. This figure includes additional data extracts obtained from the Internet Archive for content for the period 1996-2004 (for content prior to the commencement of custom .au domain harvests) and data for the 2010 calendar year (to fill a gap resulting from a domain harvest scheduling change between 2009 and 2011).

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)
1996-2004 data extraction	448m	n/a	6.7
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
2010 data extraction	100m	n/a	4.1
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
2016	690m	2,440,805	53.1
2017	900m	4,380,947	62.0

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

3.3 Collecting Commonwealth Government online publications

The Library added a substantial amount of content to its 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 2017. Currently around 455 million files or 48 terabytes of data is delivered through the AGWA.

The Library completed a small work package to index content on state government and territory domains collected as part of the annual Australian domain harvests. This indexing work provides for state government content to be made accessible through the AGWA portal.

Until the new Trove web archive delivery system is released into production (see 3.1), the AGWA remains outside the Trove discovery system. Content is openly accessible at the following location: http://webarchive.nla.gov.au/gov/

4. *Focus on users*

The Library uses Google Analytics reporting to record usage of the web archive content for both the PANDORA Archive and the Australian Government Web Archive (AGWA). The usage statistics for the previous financial year are included to provide a comparison. The figures show that while the PANDORA page views are fewer this year than the previous year, the number of users rose by nearly 8 %. Figures show a marked increase in the usage of the AGWA with a 40 % increase in page views and 43 % increase in users.

4.1 User views of the PANDORA Archive

Usage in 2016 – 2017

Total page views	Number of users	Average views per month	Average pages viewed per visit
1,756,602	316,338	146,383	4.01

Usage in 2015 – 2016

Total page views	Number of users	Average views per month	Average pages viewed per visit
1,836,961	293,580	153,080	4.4

4.2 User views of the Australian Government Web Archive

Usage in 2016 – 20176

Total page views	Number of users	Average views per month	Average pages viewed per visit
609,792	98,214	50,816	4.67

Usage in 2015 – 2016

Total page views	Number of users	Average views per month	Average pages viewed per visit
435,506	68,480	36,292	4.76

4.3 Most viewed titles (websites) in the PANDORA Archive

Around 16 % of the titles archived in PANDORA are recorded in PANDAS as being no longer online at the original 'live' site. This is an increase of 3 % over the percentage reported last year. 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 2016-2017.

	Archived Title	Participant Responsible	Live site	Page views
1	**ARIA report**	SLNSW	Yes	394,125
2	**John Kaye – Greens MP**	SLNSW	No	323,679
3	**Young and Well Cooperative Research Centre**	NLA	No	285,542
4	**Antipodean SF**	NLA	No	194,450
5	**First families 2001**	SLV	No	180,177
6	**National ANZAC Centre**	SLWA	Yes	169,351
7	**Life on the goldfields**	SLV	No	140,625
8	**Sydney Centre for Studies in Caodaism**	NLA	Yes	133,715
9	**cultureandrecreation.gov.au**	NLA	No	126,364
10	**Reviews in Australian studies**	NLA	No	108,162
11	**Victorian essential learning standards**	SLV	No	103,469
12	**National Library of Australia staff papers**	NLA	Yes	91,648
13	**Cablog : a cabbie's blog**	NLA	No	90,674
14	**Sydney Morning Herald (November 2012)**	NLA	No	90,240
15	**GamesInfo**	NLA	Yes	88,476
16	**Brady family tree in Western Australia**	NLA	Yes	84,179
17	**Significance 2.0 guide to assessing collections**	NLA	No	83,228
18	**South Land to New Holland: Dutch charting**	NLA	No	80,411
19	**Digger history**	AWM	No	75,230
20	**The Spirits of Gallipoli**	NLA	Yes	75,187

5. Promoting the Archive

5.1 Presentations, representations and papers

Presentations given by National Library Web Archiving staff during the 2016-2017 financial year included:

- Paul Koerbin's chapter for the book *Web 25: histories from the first 25 years of the World Wide Web*, edited by Niels Brügger (Aarhus University), was published by Peter Lang Publishing, New York, in 2017. Dr Koerbin's chapter is titled 'Revisiting the World Wide Web as artefact: case studies in archiving small data for the National Library of Australia's PANDORA Archive'.

- Paul Koerbin gave a mentor reviewed paper titled 'Operational challenges and innovation for national web archiving' at the 2017 ALIA Information Online Conference, Sydney, February 2017. The paper from the conference is published online at: https://informationonline.alia.org.au/content/operational-challenges-and-innovation-national-web-archiving

- Paul Koerbin gave a floor presentation at the Canberra Museum and Art Gallery as part of CMAG's exhibition Memory of the World in Canberra in March 2017. The exhibition featured items and collections listed on the Australian Memory of the World Register. In 2004, the PANDORA Archive was included on the Australian Memory of the World list as the 14th registration.

5.1 Social media

The Library's senior PANDORA curators used the @NLAPandora Twitter account for timely promotion of content from both the PANDORA Archive and the Australian Government Web Archive; and to engage directly with comments and questions. The @NLAPandora account has over 1,000 followers.

6. Concluding summary

Some of the highlights of 2016-2017 include:

- Continuing steady growth of the PANDORA Archive content at 7.9 % for titles, 14.4 % for archived instances and 40.7 % growth of the data collected (section 2.1).
- Completion of the 2017 large scale harvest of the Australian web domain, the 12th such bulk collection of .au web content since 2005 (section 3.3).
- Substantial work towards the completion of a new Trove discovery and delivery service that will provide access to the Library's entire web archive holdings (section 3.1).
- An 8 % increase in users of the PANDORA Archive and 43 % increase in users of the Australian Government Web Archive (section 4).
- A chapter devoted to the significance of the PANDORA Archive included the international publication *Web 25: histories from the first 25 years of the World Wide Web* (section 5.1).