Annual report to partners 2017-2018

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 2017-2018 financial year is provided in accordance with the National Library's obligation as stated in section 6.2 (k) of the Memorandum of Understanding 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 2017-2018 adding value through the curation of selected content. Notable collections worked on during the year include:

- *Same sex marriage debate and postal survey* – 122 websites were collected relating to the debate around marriage equality and same sex marriage and the Commonwealth Government's postal survey on the issue. This collection includes many statements and open letters that were published online during the lead up to the postal survey in late November 2017.

- *Federal Budget 2018* – as has been the practice for the past few years, a large collection of over 20 websites and documents relating to the Federal Budget process, including media, analysis and submissions, was completed.

[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.

- *2018 Winter Olympic and Paralympic Games* – 21 websites relating to Australian participation in the Winter Olympics in Pyeongchang, South Korea, in February 2018 were collected. The National Library also contributed website URLs to a collaborative international collection for the Olympics coordinated by the International Internet Preservation Consortium.

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 2017-2018 particularly when measured by data collected. The percentage growth rate for Titles was a quarter of a percent higher than the previous year at 8.25%; and the percentage growth for Instances was of a similar magnitude to last year at around 14 %. The amount of data collected, measured in terabytes, continues to increase growing at around 28 % this financial year which is down somewhat from last year's growth of 40%.

	30 June 2018	30 June 2017	Growth 2017-2018
Titles	54,781	50,605	(8.25 %)
Instances	168,544	147,309	(14.35 %)
Terabytes	40.65	31.57	(28.76 %)

Government publications remain a substantial component of the collecting focus and currently comprise approximately 48 % of the titles in the Archive. In the 2017-2018 financial year 34% 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 third chart shows the percentage variation in contribution from the previous financial year for each agency for each measure. There was a significant increase by the National Library in the contribution of titles (from a 6% increase last year to 23% this year) and instances (up 7% from the previous year) – arising from the ability to collect provided by legal deposit – and a small increases in contribution of titles and instances by the State Library of Victoria. All other partners showed a decrease in the rate of contribution of titles and instances; and all partners showed a decrease in the amount of data collected which is a significant change from the previous year.

2017-2018 financial year contributions by participant agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	6,826	13,395	114,888,444	7208.96
State Library of Victoria	2,593	3,844	8,995,347	729.48
State Library of Queensland	1,511	1,596	7,366,937	524.55
State Library of NSW	660	1,109	4,532,890	372.03
State Library of SA	561	644	3,638,634	304.52
State Library of WA	150	241	374,958	35.57
National Gallery of Australia	92	93	417,649	33.27
Australian War Memorial	41	43	445,772	24.47
AIATSIS	22	22	74,542	9.18
Northern Territory Library*	16	16	52,190	6.66

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

2016-2017 (previous) 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.

Percentage change in contributions by contributing partners between the 2016-2017 and 2017-2018 financial years

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	23%	27%	-0.4%	7%
State Library of Victoria	10%	4%	-2%	13%
State Library of Queensland	-2%	-3%	-17%	-52%
State Library of NSW	-12%	-10%	-3%	-4%
State Library of SA	-3%	-19%	-20%	-5%
State Library of WA	-45%	-32%	-43%	-16%
National Gallery of Australia	-10%	-15%	-24%	-5%
Australian War Memorial	-9%	-9%	-36%	-21%
AIATSIS	-24%	-33%	-47%	-45%
Northern Territory Library	n/a	n/a	n/a	n/a

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 2017-2018 financial year has continued to be 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 further development of web collecting infrastructure, including PANDAS (the PANDORA Digital Archiving System) and the Heritrix based bulk harvesting infrastructure (known as 'Bamboo' or 'Digital Library Web') used for the Australian Government Web Archive.

3.1 Development of the 'Trove web archive' zone

The Library has been working on a new technical and policy infrastructure that will allow the Library to provide access to its entire corpus of web archive collections – including PANDORA, AGWA and the large domain harvest collections – through a single search interface under Trove. This development work includes a new interface for the delivery of web archive content that will replace both the PANDORA and the AGWA delivery systems.

Work on this project continued through 2017-2018, including:

- Refinement of the indexing the vast amount of content that forms the domain harvest collections (i.e. more than 4 billion text documents full-text indexed) so search results are ranked to ensure high value content is delivered;
- Refining a new access control tool to manage both access and discovery restrictions (together with the migration of PANDORA restrictions to the new tool); and,
- Work to identify, assess, document and mitigate legal and social 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, addressing performance issues and continuing world to mitigate risk associated with releasing domain harvest content.

3.2 Australian web domain harvest

In the first quarter of 2018 the Library conducted the 13th large-scale harvest of the Australian web domain. This was the third 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 the period from March to May 2018 and more than 980 million unique documents were captured, amounting to 77 terabytes of data from more than three million hosts.

Following this harvest, the combined total for all 13 Australian domain harvests has now reached nearly 10 billion files amounting to around 528 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
2018	986m	3,030,348	77.9

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 2018. Currently around 506 million files or 56 terabytes of data are delivered through the AGWA.

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 through the AGWA's dedicated access portal 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 2017 – 2018

Total page views	Number of users	Average views per month	Average pages viewed per visit
1,746,568	317,722	145,547	3.85

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

4.2 User views of the Australian Government Web Archive

Usage in 2017 – 2018

Total page views	Number of users	Average views per month	Average pages viewed per visit
679,348	91,609	56,612	4.99

Usage in 2016 – 2017

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

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 location. This percentage is unchanged from that 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 2017-2018.

	Archived Title	Participant Responsible	Live site	Page views
1	**Australian policy online**	NLA	Yes	254,687
2	**ATSIC publications**	NLA	No	216,015
3	**Racing and sports: Melbourne Cup**	SLV	Yes	202,144
4	**Antipodean SF**	NLA	No	168,012
5	**First families 2001**	SLV	No	146,714
6	**Douglas Snelling**	NLA	Yes	134,410
7	**National ANZAC Centre**	SLWA	Yes	126,265
8	**The Conversation (G20)**	NLA	No	120,021
9	**Deputy Prime Minister & Minister for Agriculture and Water Resources**	NLA	No	107,643
10	**Australian Federal Attorney-General**	NLA	Yes	104,811
11	**Nicola Marsh**	NLA	Yes	103,966
12	**Media spy**	NLA	Yes	101,909
13	**Sydney Centre for Studies in Caodaism**	NLA	Yes	98,680
14	**Culture.gov.au : Australia's cultural portal**	NLA	No	95,087
15	**Gravesecrets at your fingertips [cemetaries]**	SLSA	Yes	92,931
16	**Life on the goldfields**	SLV	No	92,315
17	**Victorian essential learning standards**	SLV	No	78,497
18	**Sensoria : a journal of mind, brain and culture**	NLA	No	74,581
19	**Footypedia**	NLA	No	74,525
20	**ARIA report**	SLNSW	Yes	73,853

5. *Promoting the Archive*

5.1 Presentations, representations and papers

Presentations given by National Library Web Archiving staff during the 2017-2018 financial year included:

- Paul Koerbin was invited to join the Organising Committee for the 2018 International Internet Preservation Consortium (IIPC) Web Archiving Conference being organized and hosted by the National Library of New Zealand. Dr Koerbin was also appointed co-chairperson of the Conference Programme Committee.
- Russell Latham (National Library of Australia), Maxine Fisher (State Library of Queensland) and Peter Jetnikoff (State Library of Victoria) all had papers relating to PANDORA accepted for presentation at the 2018 IIPC Web Archiving Conference (to be held in November 2018 in Wellington).
- Paul Koerbin gave a presentation on PANDORA to an AIATSIS Collections meeting in March 2018.
- In September 2017, Paul Koerbin participated in a web conference meeting with a number of senior managers from the National Library of Singapore. The web conference covered a broad range of topics covering the National Library of Australia's experience in web archiving practice.

5.2 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 1,200 followers.

6. *Concluding summary*

Some of the highlights of 2017-2018 include:

- Continuing steady growth of the PANDORA Archive content at 8.25 % for titles, 14.35 % for archived instances and 28.76 % growth of the data collected (section 2.1).
- Completion of the 2018 large scale harvest of the Australian web domain, the 13th such bulk collection of .au web content since 2005 (section 3.2); adding 77 terabytes of data or nearly 1 billion files to the web archive collection.
- Continued work towards the release of a new Trove discovery and delivery service that will provide access to the Library's entire web archive holdings (section 3.1).
- Steady use of the PANDORA Archive with an average of around 145,000 views per month (section 4.1).
- Engagement with the international web archiving community by PANDORA participants through the acceptance of three papers for 2018 IIPC Web Archiving Conference to be held in Wellington in November 2018 (section 5.1).