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

Annual report to partners 2018-2019

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 2018-2019 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

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 Partner teleconferences

The Library organised 'Zoom' teleconferences with Partners in February and July 2019. These teleconferences aimed to provide an update on developments, including the release of the new Trove Web Archive service, as well as providing an opportunity for discuss and questions.

1.3 Notable collaborative collections

A number of collections were developed, formed or extended during the 2018-2019 period adding value through the curation of selected content. Notable collections collectively worked on during the year include:

- The May 2019 Federal Election campaign collection. This project collected around 1,086 websites relating to the election campaign. This included party and candidate sites, media and commentary, selected Twitter accounts, lobby groups and electoral study and research websites. The size of collection was assessed as a 12% increase in content over the previous election campaign collecting in 2016.
- Restaurants and Cafes collection. This collection aimed to capture a broad range of sites from both city and regional areas that document dining culture from the casual to the formal. More than 100 restaurant and café websites from around Australia were archived.

[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.

2. *Growth of the Web Archive*

2.1 Size and annual growth of the PANDORA Archive

The PANDORA Archive maintained a consistent high level of growth in 2018-2019 consistent with recent, post legal deposit, years. The percentage growth rate for Titles at 9.88% was 1.63% higher than the previous year while the percentage growth rate for Instances at 14.33% was unchanged. The amount of data collected, measured in terabytes grew at nearly 12% which, while still strong was down significantly on the growth rate of the previous two years.

	30 June 2019	30 June 2018	Growth 2018-2019
Titles	60,197	54,781	(9.88 %)
Instances	192,709	168,544	(14.33 %)
Terabytes	45.5	40.65	(11.93 %)

Government publications remain a substantial component of the collecting focus and currently comprise approximately 45.5 % of the titles in the Archive. In the 2018-2019 financial year, 24% of new titles registered and archived were government titles. This is a lower percentage than the historic average for collecting government publications which reflects the National Library's focus on bulk harvesting of government content outside the PANDORA infrastructure and a growing uptake by government publishers of the National eDeposit (NED)' service.

2.2 Statistics for annual participant contributions

The first two charts below show the contribution to PANDORA of each participating agency for the current and previous financial years for comparison.

The third chart shows the percentage variation in contribution from the previous financial year for each agency for each measure. A number of partners recorded an increase in the contribution of titles and instances while there was an across the board and significant decrease in the amount of data collected (as measured in terabytes and number of files). The may be the result of a focus on more document-like material rather than large websites or more targeted scoping of harvests, or a combination of both.

2018-2019 financial year contributions by participant agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	8,879	16,221	45,919,879	3768.32
State Library of Victoria	2,964	3,919	3,909,612	397.86
State Library of Queensland	1,348	1,416	3,139,399	269.93
State Library of NSW	1,062	1,617	2,396,598	265.29
State Library of SA	537	567	2,060,148	181.88
State Library of WA	198	246	261,715	29.63
National Gallery of Australia	110	113	157,561	19.24
Australian War Memorial	35	36	211,975	15.51
AIATSIS	3	3	4,281	0.47
Northern Territory Library*	1	1	125	0.16

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

2017-2018 (previous) 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.

Percentage change in contributions by contributing partners between the 2017-2018 and 2018-2019 financial years

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	30%	21%	-60%	-48%
State Library of Victoria	14%	2%	-56%	-45%
State Library of Queensland	-10%	-11%	-57%	-49%
State Library of NSW	61%	46%	-47%	-29%
State Library of SA	-4%	-12%	-43%	-40%
State Library of WA	32%	2%	-30%	-17%
National Gallery of Australia	20%	22%	-62%	-42%
Australian War Memorial	-15%	-16%	-52%	-38%
AIATSIS	-86%	-86%	-94%	-95%
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.

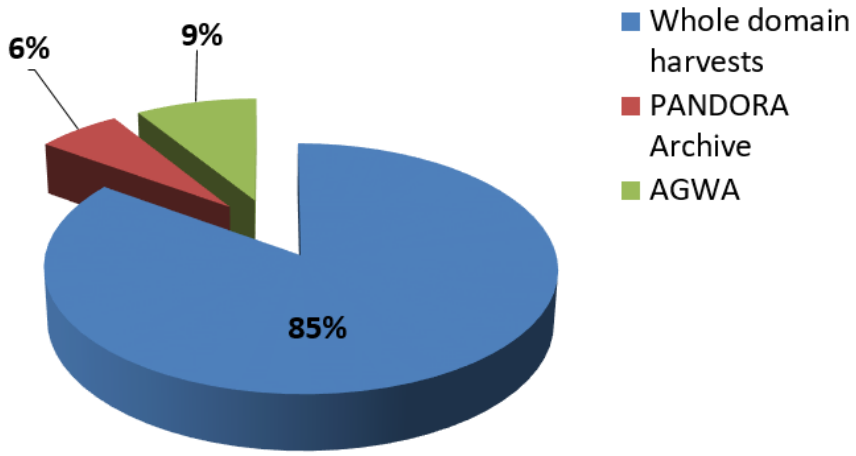
The focus of web archiving development over the 2018-2019 financial year the completion and implementation of the Trove discovery service for the Australian Web Archive (see 3.1).

3.1 The Australian Web Archive and Trove discovery and delivery service

On 5 March 2019 the Library released a new service called the Australian Web Archive (AWA) through the Trove discovery service. The new Trove service is designated the 'Australian Web Archive' because it includes the content from all three web archive collections maintained at the Library: the PANDORA Archive; the Australian Government Web Archive; and the entire corpus of domain harvest collections.

The Trove service provides a new single point of access to PANDORA and AGWA content and, most importantly, access for the first time to all the content from the domain harvest collection. The domain harvest collection accounts for around 85% of the entire AWA content.

The following graph shows the relative percentage of the size of the collections that comprise the Australian Web Archive.



Total web archive data delivered through the AWA:

- Around 600 TBs
- More than 10 billion files

3.2 Australian web domain harvest

In the first quarter of 2019 the Library conducted the 14th large-scale harvest of the Australian web domain. This was the fourth Australian domain harvest conducted since 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 the infrastructure, expertise and experience in this method of large scale web archiving.

The harvest was run during the period from March to May 2019 and more than 818 million unique documents were captured, amounting to 75 terabytes of data from two and a half million host domains.

Following this harvest, the combined total for all 14 annual Australian domain harvests has now passed 10 billion files amounting to around 600 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	448 m	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	100 m	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	566 m	2,580,521	42.1
2016	690 m	2,440,805	53.1
2017	900 m	4,380,947	62.0
2018	986 m	3,030,348	77.9
2019	818 m	2,520,041	75.7

3.3 Collecting Commonwealth Government online publications

With the release of the Australian Web Archive in March 2019 content collected for the Australian Government Web Archive (AGWA) became part of that service delivered through Trove. Consequently, the separate AGWA portal was closed in July 2019. All links to the old AGWA service will now redirect to the AWA. The AGWA portal was always conceived of as a prototype service and the move to integrate the content with the other web archive collections (PANDORA and the domain harvests) fulfils the strategic objective for the service.

The Library continues to run 'in-house' bulk harvests of Commonwealth Government websites roughly four times a year collecting around 6 TBs (or 50 million files) of government content per annum. Custom bulk harvests collected over 2 million government PDF documents in the 2018-2019 financial year.

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 new Trove Australian Web Archive. The 2018-2019 financial year was a transitional period moving from PANDORA and AGWA delivery to the single delivery service the Trove. The PANDORA Archive and AGWA access pages were both available during this period of transition. Figures for the Trove Australian Web Archive are for the period from 4 March 2019, the date that the service came into operation. Separate figures for PANDORA are

reported up to 3 March 2019 and from 4 March to 30 June 2019 separately. Because of these changes, comparative figures for previous years have not been provided.

4.1 User views of the PANDORA Archive

Usage for the period 1 July 2018 to 3 March 2019 (pre-Trove Web Archive release)

Total page views	Number of users	Average views per month	Average pages viewed per visit
1,040,708	178,154	130,088	4.15

Usage for the period 4 March 2019 to 30 June 2019 (post-Trove Web Archive release)

Total page views	Number of users	Average views per month	Average pages viewed per visit
340,422	75,385	85,105	3.52

4.2 User views of the Trove Australian Web Archive

Usage for the period 4 March 2019 (release date) to 30 June 2019

Total page views	Number of users	Average views per month	Average pages viewed per visit
318,953	27,844	79,738	7.93

4.3 User views of the Australian Government Web Archive

Usage in 2018 – 2019

Total page views	Number of users	Average views per month	Average pages viewed per visit
656,858	88,491	54,738	4.71

4.4 Most viewed titles (websites) in the PANDORA Archive

Around 15 % of the titles archived in PANDORA are recorded in PANDAS as being no longer online at the original 'live' site location. Since this figure relies on curators recording this fact, the actual figure is certainly somewhat higher; and even sites that are still 'live' may not continue to include content that was harvested earlier for the Archive. Around half of the most used sites in PANDORA are recorded as no longer available as live websites. The table below shows the top 20 sites accessed in 2018-2019.

	Archived Title	Participant Responsible	Live site	Page views
1	**Johnny's Pages: Old S.A.R. Shunter's Memories**	SLSA	No	259,806
2	**Antipodean SF**	NLA	No	123,029
3	**Footypedia**	NLA	No	117,572
4	**National Treasures from Australia's Great Libraries**	NLA	No	104,792
5	**Honouring Anzacs**	NLA	Yes	88,511
6	**First families 2001**	SLV	No	87,847
7	**Reconciliation Australia**	AIATSIS	Yes	82,306
8	**Sydney Morning Herald (October 2009)**	NLA	No	78,148
9	**National ANZAC Centre**	SLWA	Yes	67,811
10	**Communicable diseases intelligence**	NLA	Yes	66,524
11	**2009 Victorian Bushfires Royal Commission**	SLV	No	64,953
12	**Sydney Centre for Studies in Caodaism**	NLA	Yes	54,495
13	**Australian Federal Attorney-General**	NLA	Yes	49,819
14	**Life on the goldfields**	SLV	No	46,215
15	**ANZAC Centenary: the Bendigo story**	SLV	No	42,327
16	**ARIA report**	SLNSW	Yes	41,417
17	**State Library of Victoria**	SLV	Yes	40,536
18	**Gravesecrets at your fingertips [cemetaries]**	SLSA	Yes	39,064
19	**The Spirits of Gallipoli**	NLA	Yes	38,665
20	**Daily weather observations**	NLA	Yes	36,257

5. International relations

5.1 International Internet Preservation Consortium (IIPC)

- Paul Koerbin was the co-chair of the Programme Committee for the 2018 IIPC Web Archiving Conference held in Wellington, New Zealand in November 2018. A small number of Pandora partner agency staff from the National Library and state libraries of Victoria, Queensland, South Australia and New South Wales managed to attend the Conference.
- In April 2019 the National Library was elected to the IIPC Steering Committee. The Library was a founding member of the IIPC in 2003 but had not had a position on the Steering Committee since 2009. Membership of the Steering Committee is for a three year period until mid-2022.
- Dr Koerbin was elected Vice-Chair of the IIPC, one of the three officer roles in the Consortium, for the 2020 calendar year.

6. Promoting the Archive

6.1 Presentations, representations and papers

Major presentations and papers during the 2018-2019 financial year included:

- Russell Latham (National Library of Australia), Maxine Fisher (State Library of Queensland) and Peter Jetnikoff (State Library of Victoria) all presented papers relating to PANDORA at the 2018 IIPC Web Archiving Conference held in Wellington, New Zealand in November 2018. The papers presented were:
    - Maxine Fisher: Web archiving Australia's Sunshine State: from vision to reality
    - Peter Jetnikoff: Curating dissent at the State Library of Victoria
    - Russell Latham: 'Who by fire'" lifespans of websites from a web archive perspective
- Paul Koerbin was invited to write a chapter for a forthcoming book edited by Daniel Gomes (Portuguese Web Archive) titled 'The Past Web' to be published by Springer. Dr Koerbin's chapter is titled 'National web archiving in Australia: representing the comprehensive'.
- In August 2018, Paul Koerbin gave a presentation on legal deposit and web archiving to a delegation from the National Library of Indonesia.

6.2 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 1,300 followers. The Library's creation of the hashtag #WebArchiveWednesday was taken up by the International Internet Preservation Consortium members and has not become an established promotional tag for the web archiving community internationally.

7. Concluding summary

Some of the highlights of 2018-2019 include:

- Continuing steady growth of the PANDORA Archive content at 9.88 % for titles, 14.33 % for archived instances and 11.93 % growth of the data collected (section 2.1).
- Release of the new Trove Australian Web Archive service providing access to all web archive collections including PANDORA, AGWA and all domain harvests (section 3.1).
- Completion of the 2019 large scale harvest of the Australian web domain, the 14th annual bulk collection of .au web content since 2005 adding 75 terabytes of data or more than 800 million files to the web archive collection (section 3.2).
- The National Library elected to the Steering Committee of the International Internet Preservation Consortium (IIPC) and Dr Koerbin elected to the role of Vice-Chair (section 5.1).
- Strong representation at the 2018 IIPC Web Archiving Conference in Wellington by PANDORA partners (section 6.1)