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

Annual report to partners 2015-2016

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 2015-2016 financial year.

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 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 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 2015-2016, including these major collections:

- 2016 Federal Election Campaign

  This collection consists of six sub collections covering candidate, political party, media, interest and lobby group and electoral research websites. Altogether around 855 websites were archived.

- Great Barrier Reef

  This collection led by the State Library of Queensland includes 56 websites from conservation, government and tourism sectors.

[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.

- Other collections which were commenced or continued to be curated collaboratively included:
    - Agricultural Shows and Show Societies
    - ANZAC Centenary
    - Australian Feature Production Films and Documentaries
    - Notable Australian Companies
    - Iconic Australian Brands

2. *Growth of the Archive*

2.1 Size and annual growth of the PANDORA Archive

The PANDORA Archive maintained consistent growth in 2015-2016. The percentage growth rate for Titles and Instances was of a similar magnitude to the previous financial year being steady in respect to growth in titles collected and up 1% over the previous year in respect to instances collected. The amount of data collected, measured in terabytes, continues to increase growing 39% this financial year compared with 35% 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 2016	30 June 2015	Growth 2015-2016
Titles	46,904	42,424	(10 %)
Instances	128,797	112,652	(14 %)
Terabytes	22.41	16.14	(39%)

Government publications remain a substantial component of the collecting focus and comprise approximately 51 % of the titles in the Archive. In the 2015-2016 financial year 35% 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.

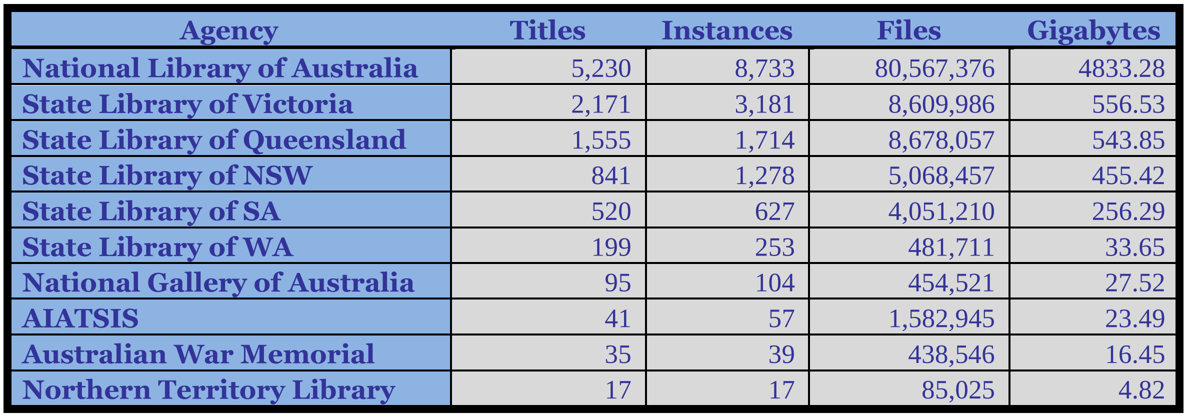
2.2 Statistics for annual participant contributions

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. In order to make the chart more useful it has been sorted based on the contribution of Instances archived.

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 431 MB up from 330 MB last year.

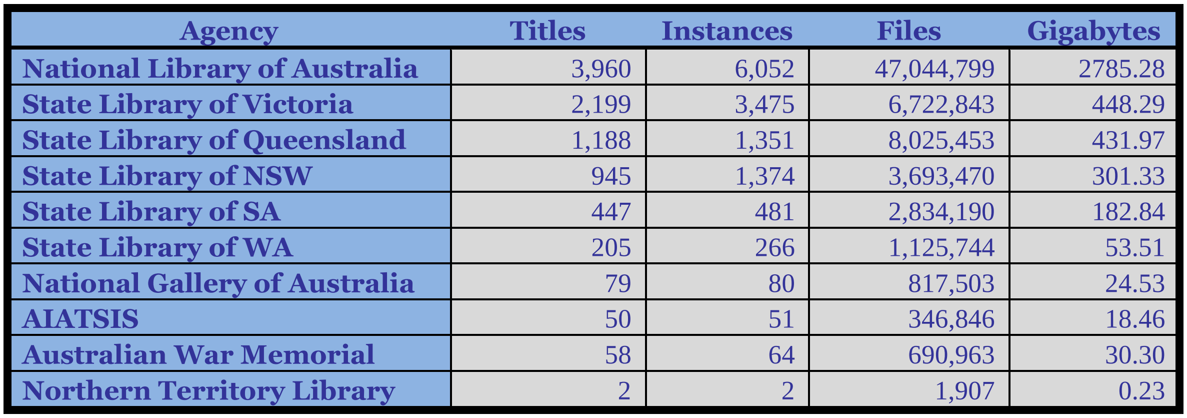
The third chart shows the percentage variation from the previous financial year for each agency for each measure, most notably indicating an across-the-board increase in the size of the instances archived.

2015-2016 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
AIATSIS	41	57	1,582,945	23.49
Australian War Memorial	35	39	438,546	16.45
Northern Territory Library	17	17	85,025	4.82

2014-2015 (previous) 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 Queensland	1,188	1,351	8,025,453	431.97
State Library of NSW	945	1,374	3,693,470	301.33
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
AIATSIS	50	51	346,846	18.46
Australian War Memorial	58	64	690,963	30.30
Northern Territory Library	2	2	1,907	0.23

Percentage change in contributions between 2014-2015 and 2015-2016 financial years

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	32%	44%	71%	74%
State Library of Victoria	-1%	-8%	28%	24%
State Library of Queensland	31%	-7%	8%	26%
State Library of NSW	-11%	16%	37%	51%
State Library of SA	16%	30%	43%	40%
State Library of WA	-3%	-5%	-57%	-37%
National Gallery of Australia	20%	30%	-44%	12%
AIATSIS	-18%	12%	356%	27%
Australian War Memorial	-40%	-39%	-37%	-46%
Northern Territory Library	750%	750%	4359%	1996%

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 that support the collection of Australian web resources.

3.1 Extension of Legal Deposit to electronic materials

On 17 February 2016 amendments to the Copyright Act 1968 came into effect which extended the legal deposit provisions in the Act to electronic materials including online materials. Consequently the Library undertook infrastructure development to manage the deposit of electronic materials. Specifically this involved the release of an ‘eDeposit’ system. Because web archiving is a long-standing operational activity in the Library less work was required to transition immediately to collecting under legal deposit provisions; while the changes provide the opportunity for future development to web archiving infrastructure to enhance collecting potential.

3.2 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.

Some minor changes to PANDAS that were completed to enable a smooth transition to incorporate workflows to accommodate the extension of legal deposit to electronic materials that came into force in February 2016. Some changes included: the ability for NLA curators to harvest and archive material without recording publisher permissions; changes to permission letter templates to reflect the purpose of permissions relating only to access (not collecting); a flag to indicate a title collected under legal deposit provisions; and, the automatic application of an on-site access restriction to titles archived without an access permission. These changes only applied to the NLA working view of PANDAS. Partner agencies workflows was not affected.

3.3 Australian web domain harvest

In the first quarter of 2016 the Library conducted the 11th large-scale harvest of the Australian web domain. This was the first whole Australian domain harvest conducted since the commencement of legal deposit on 17 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 web archiving.

The harvest was run during February and March 2016 (commencing after the implementation of electronic legal deposit) and more than 690 million unique documents were captured, amounting to 53.1 terabytes of data from around two and a half million hosts.

Following this harvest the combined total for all eleven Australian domain harvests has now reached nearly 8 billion files amounting to around 380 terabytes of data. This figure includes a data extract of Internet Archive content for the period 1996-2004 obtained in 2015 (amounting to 448,285,899 files or 6.7 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
2016	690m	2,440,805	53.1

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.

3.4 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 2016. Currently around 234 million files or 24 terabytes of data is delivered through the AGWA.

Work continued on the development of some tools to support the AGWA workflows. A simple file harvester called 'Butterflynet' was implemented which provides staff with a 'one-click' facility to collect simple PDF documents that can be indexed and added in real-time to the AGWA collection. A version of Ilya Kreymer's 'Webrecorder', a headless browser harvester, was also implemented providing senior curators the ability to do patch-up harvesting to collect content, such as JavaScript files, that could not be collected by the harvest crawler. In addition, a seed list management tools was also added to suite of applications being prototyped in this development environment locally referred to as 'Bamboo'.

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

Google Analytics reporting is now used to record usage of the web archive content. Consequently, actual numbers are inconsistent with and cannot be compared with previous reporting.

4.1 User views of the PANDORA Archive

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 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 13 % 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 2015-2016.

	Archived Title	Participant Responsible	Live site	Page views
1	**First families 2001**	SLV	No	209,135
2	**National ANZAC Centre**	SLWA	Yes	170,329
3	**Sydney Centre for Studies in Caodaism**	NLA	Yes	146,883
4	**punters.com.au**	NLA	Yes	143,063
5	**Convergence Review**	NLA	No	138,889
6	**Life on the goldfields**	SLV	No	132,537
7	**Footypedia**	NLA	No	120,163
8	**cultureandrecreation.gov.au**	NLA	No	104,193
9	**Full Points Footy**	SLV	No	101,680
10	**Australian mining : news website**	NLA	Yes	101,614
11	**Antipodean SF**	NLA	No	97,515
12	**Walking for Country**	AIATSIS	Yes	85,572
13	**Simplicity Institute**	NLA	Yes	77,321
14	**Digger history**	AWM	No	77,230
15	**State Library of NSW website**	SLNSW	Yes	76,358
16	**The Spirits of Gallipoli**	NLA	Yes	73,663
17	**Centenary of Federation**	NLA	No	73,089
18	**ARIA report**	SLNSW	Yes	69,374
19	**Senator Kate Lundy**	NLA	No	67,770
20	**Using Lives: essays in Australian biography and history**	NLA	Yes	65,523

5. *Promoting the Archive*

5.1 Presentations, representations and papers

Presentations given by National Library Web Archiving staff during the 2015-2016 financial year included:

- Paul Koerbin was invited to participate in a panel presentation on the future of national web archiving at the International Internet Preservation Consortium (IIPC) General Assembly and Conference in Reykjavik, Iceland in April 2016. Dr Koerbin's participation was made possible by the generous travel support provided by Brewster Kahle and the Internet Archive.

- Paul Koerbin gave a presentation on the Australian Government Web Archive to National Library Petherick Readers annual meeting in December 2015.

- Paul Koerbin wrote a brief article on the Australian Government Web Archive for the Independent Scholars Association of Australia (ISAA) newsletter.

5.2 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 2015-2016, including:

- Staff from the Digital Transformation Office (9 March 2016)
- Staff from the Government Records Service of Hong Kong (31 May 2016)

In December 2015 Helen Hockx-Yu, Director of Global Web Services at the Internet Archive, visited the Library to find out more about our web archiving developments for a report on national web archiving programs that she was working on.

In November 2015 two staff from the Australian War Memorial spent two days at the Library receiving a PANDAS training refresh and some assistance with quality assurance work and technical fixes.

6. Concluding summary

Some of the highlights of 2015-2016 include:

- Continuing steady growth of the PANDORA Archive content at 10% for titles, 14% for archived instances and 39% in the growth of the data (section 2.1).
- Completion of the 2016 large scale harvest of the Australian web domain, the eleventh such bulk collection of .au web content since 2005 (section 3.3).
- Changes to the PANDAS workflow system to support new collecting workflows following the extension of legal deposit to online publications from February 2016 (sections 3.1 and 3.2).
- Continued development of prototype tools to support in house harvesting for AGWA (section 3.4).
- Representation at the International Internet Preservation Consortium (IIPC) General Assembly and Conference in Reykjavik, Iceland, in April 2016 (section 5.1).