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

Annual report to partners 2019-2020

Contents

1. *PANDORA participants working together*

PANDORA, refers to the collaborative selective web archiving program led by the National Library of Australia (NLA) with participating agencies: the state libraries of Victoria (SLV), New South Wales (SLNSW), Queensland (SLQ), South Australia (SLSA) and Western Australia (SLWA), the Library & Archives NT (LANT), the Australian War Memorial (AWM), the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) and the National Gallery of Australia NGA).

This report to contributing participants on activities and developments in the 2019-2020 financial year is made available 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 an ad hoc 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 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.

1.3 Notable PANDORA collections

A number of collections were developed, formed or extended during the 2019-2020 period adding value through the curation of selected content. Notable collections worked on during the year include:

- *Bushfires, Australia (2019-2020)* – a collection of just over 140 websites documenting the catastrophic bushfires in eastern and southern Australia in the summer of 2019-2020.
- *Coronavirus (COVID-19) Pandemic and Australia, 2020* – large collection documenting all aspects and effects of the pandemic and responses to it in Australia. The collection includes 18 sub-collections covering individual state responses as well as effects on culture, tourism, the economy etc. More than 1600 websites archived (as at October 2020), many regularly and frequently, making it the largerst ever PANDORA collection.
- *Endeavour 1770 – Encounters 2020 (250 year anniversary of Cook's Endeavour encounters in Australia)* – a collection of just over 30 websites documenting events and views in respect to the commemoration of Cook's landing on the Australian continent 250 years ago.

[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 2019-2020 consistent with recent years following the extension of legal deposit to online materials. The percentage growth rate for Titles at 8.19% was 1.69% lower than the previous year while the percentage growth rate for Instances at 17.24% was 2.91% higher than last year. The amount of data collected, measured in terabytes grew at nearly 17% which is a 5% increase over the previous year.

	30 June 2020	30 June 2019	Growth 2019-2020
Titles	65,128	60,197	(8.19 %)
Instances	225,934	192,709	(17.24 %)
Terabytes	53.13	45.5	(16.76 %)

Government publications remain a substantial component of the collecting focus and currently comprise approximately 45 % of the titles in the Archive. In the 2019-2020 financial year, 40% of new titles registered and archived were government titles. This is significantly higher than the previous year (24%) and reflects extensive selection of government online content in respect to the COVID-19 pandemic in the first half of 2020.

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. There does not appear to be any pattern across agencies in the variation of contribution this year except that among the more active agencies there was a notable growth in data collected. Generally, however, this chart would seem to refelct individual approaches to and capacity for collecting from agency to agency.

2019-2020 financial year contributions by participant agency

Agency	Titles	Instances	Files	Gigabytes
National Library of Australia	9,898	24,466	62,443,031	6021.12
State Library of Victoria	2,976	4,400	4,371,853	556.16
State Library of Queensland	1,361	1,736	3,754,536	350.14
State Library of NSW	591	1,230	2,508,091	455.83
State Library of SA	596	710	1,864,351	366.45
State Library of WA	215	424	179,469	24.46
National Gallery of Australia	72	72	110,487	11.50
Australian War Memorial	35	35	104,682	8.53
AIATSIS	36	39	67,755	8.48
Library & Archives NT*	3	4	1,830	0.26

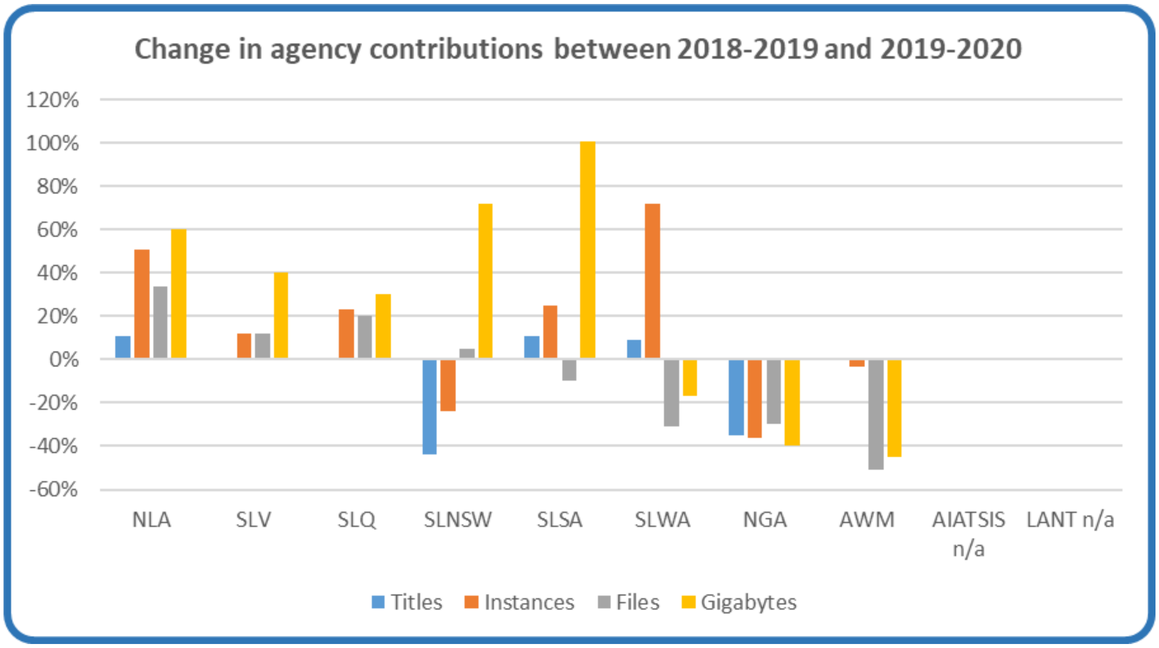
*Formerly the Northern Territory Library (NTL).

2018-2019 (previous) 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
Library & Archives NT*	1	1	125	0.16

*Formerly the Northern Territory Library (NTL).

Percentage change in contributions by contributing partners between the 2018-2019 and 2019-2020 financial years

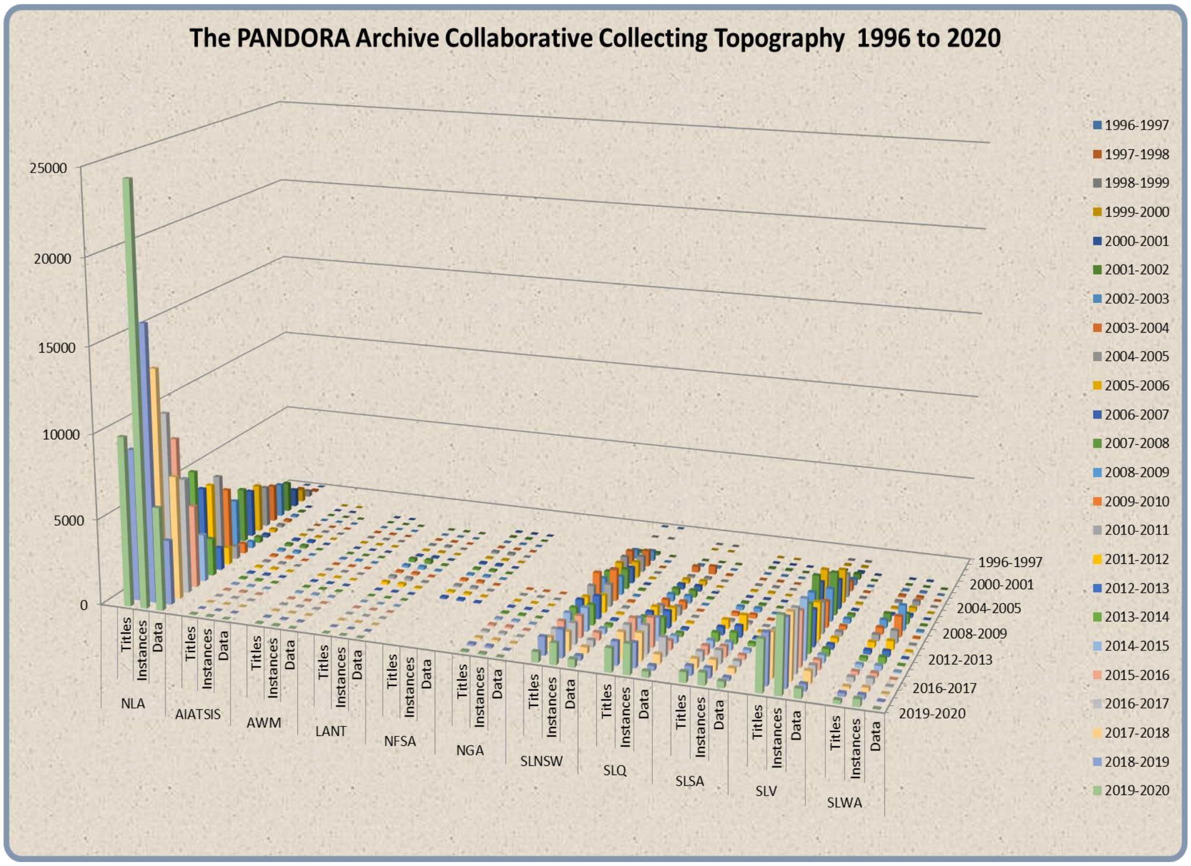


*The figures for AIATSIS and LANT are too small to provide a meaningful percentage change figure.

3. Analysis

This year we repeat an analysis last done in the 2012-2013 financial year to look at the collecting trends of each PANDORA partner agency. The analysis looks at the year-by-year contributions, of each partner agency (since commencing as a PANDORA contributor) showing both the trend in archiving activity and the percentage of the contribution of each partner agency to each year's total PANDORA archiving activity.

The first chart shows the 'topography' of all contributions by title, instance and data, for all agencies for the entire history of PANDORA collecting activity from 1996 to June 2020.

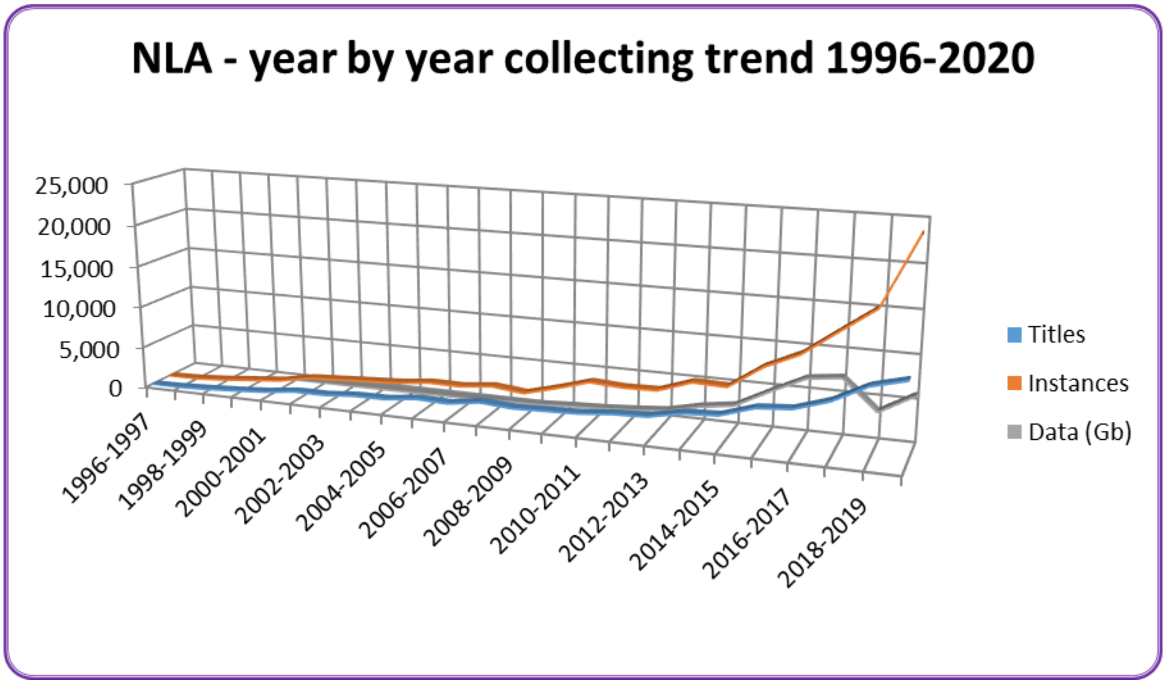


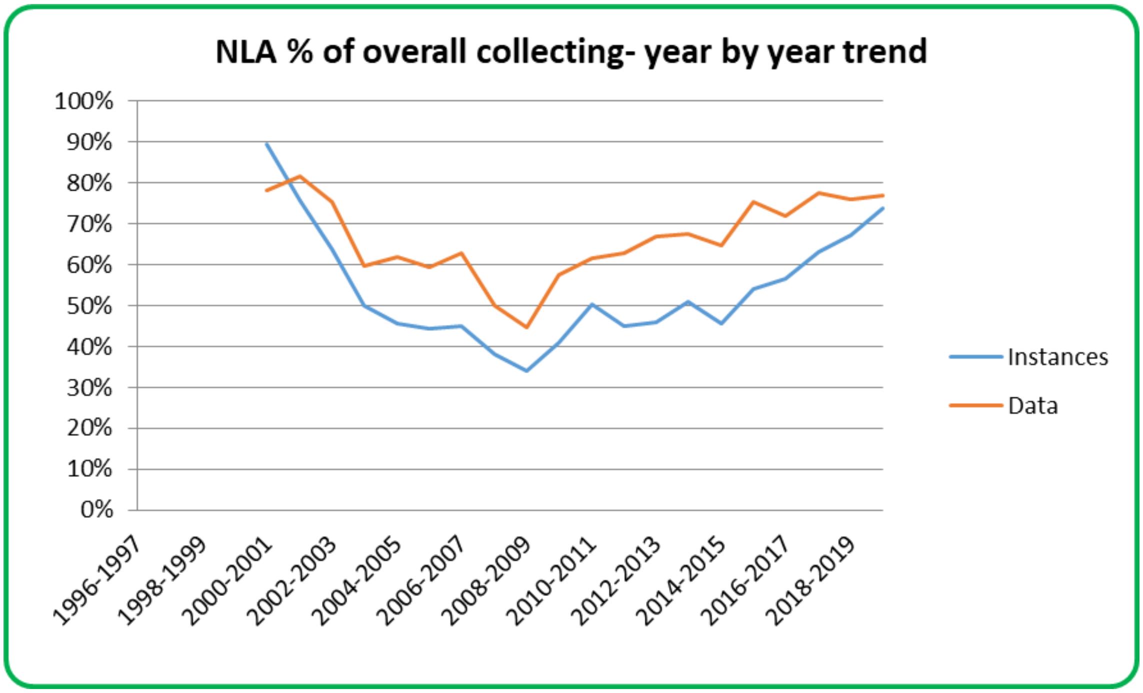
The PANDORA Archive Collaborative Collecting Topography 1996 to 2020

3.1 Individual partner contribution trends over the life of the archive

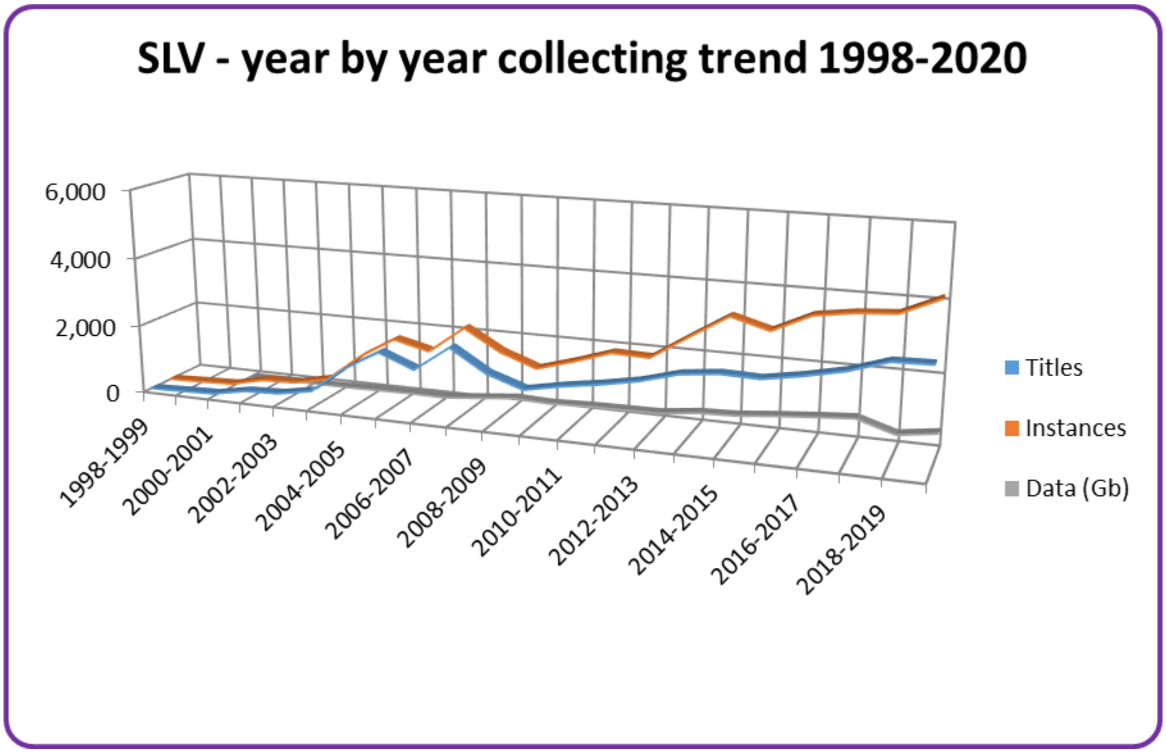
The following series of charts show the year-by-year collecting trends in respect to each participating PANDORA agency. The first chart for each agency shows the collecting trend measured by archived titles, instances and gigabytes. The second chart for each agency shows the year-by-year trend in respect to the percentage of content contributed by the partner agency to the overall collecting for PANDORA. The second chart in the series commences with the 2000-2001 financial year which is the first year that the data can be meaningfully measured. The trend data for this measure is most meaningful from around the mid-decade by which time a number of partner agencies were contributing to the Archive.

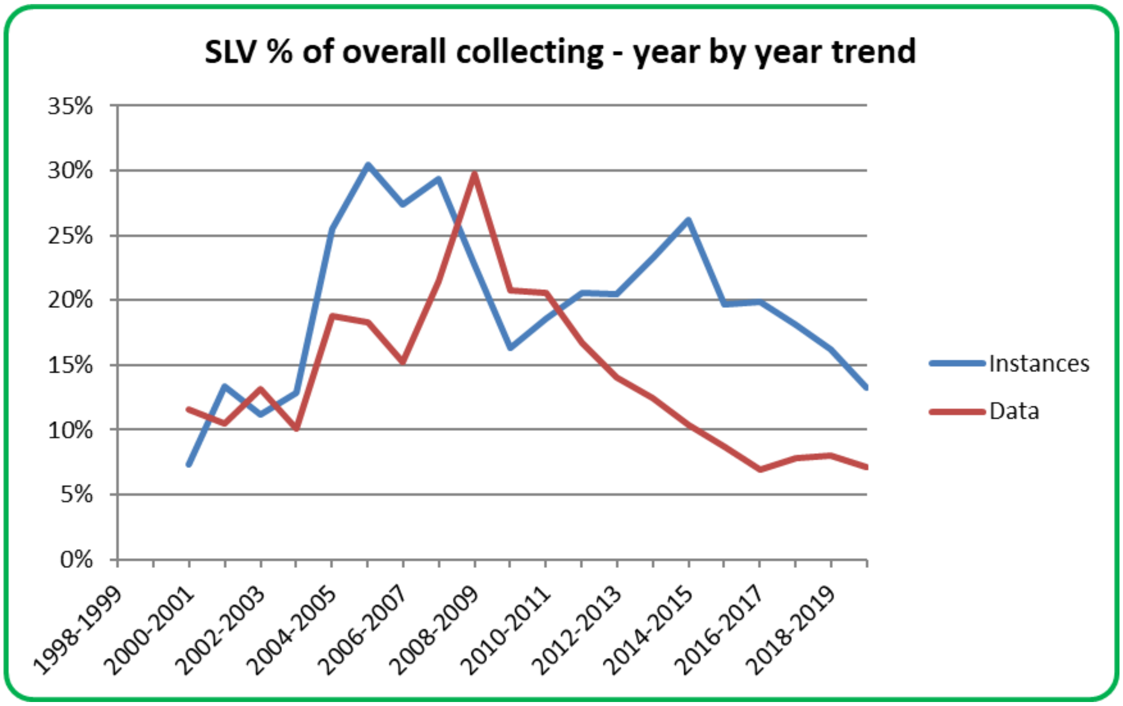
National Library of Australia, 1996-2020



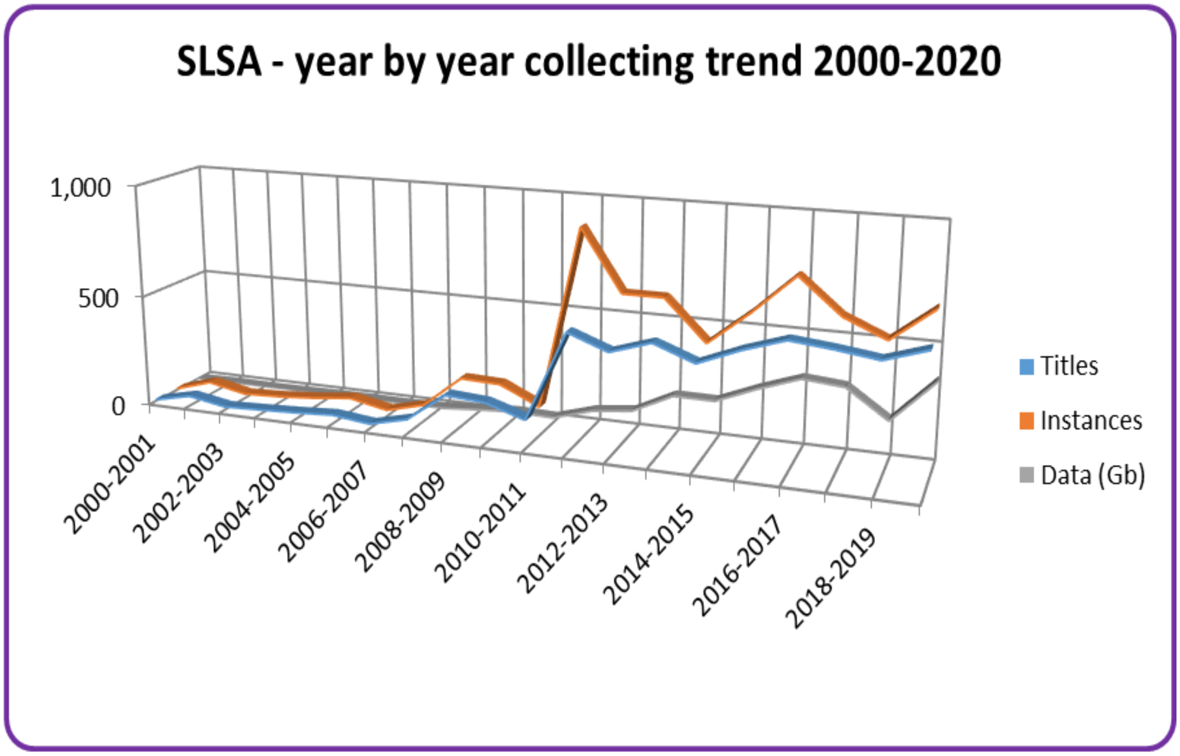


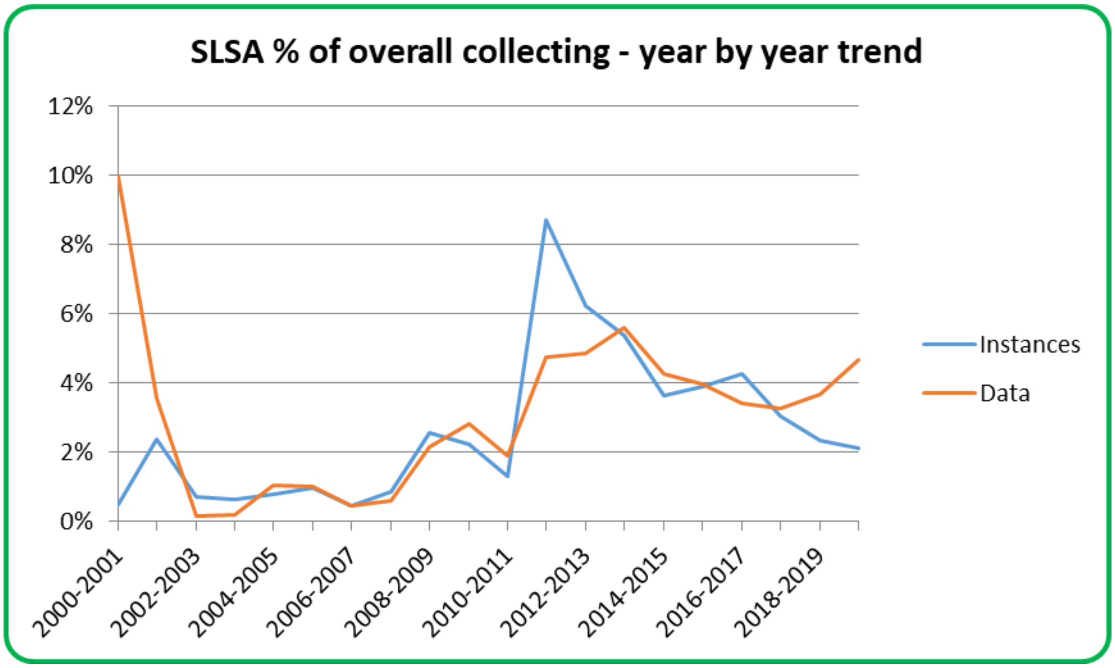
State Library of Victoria, 1998-2020



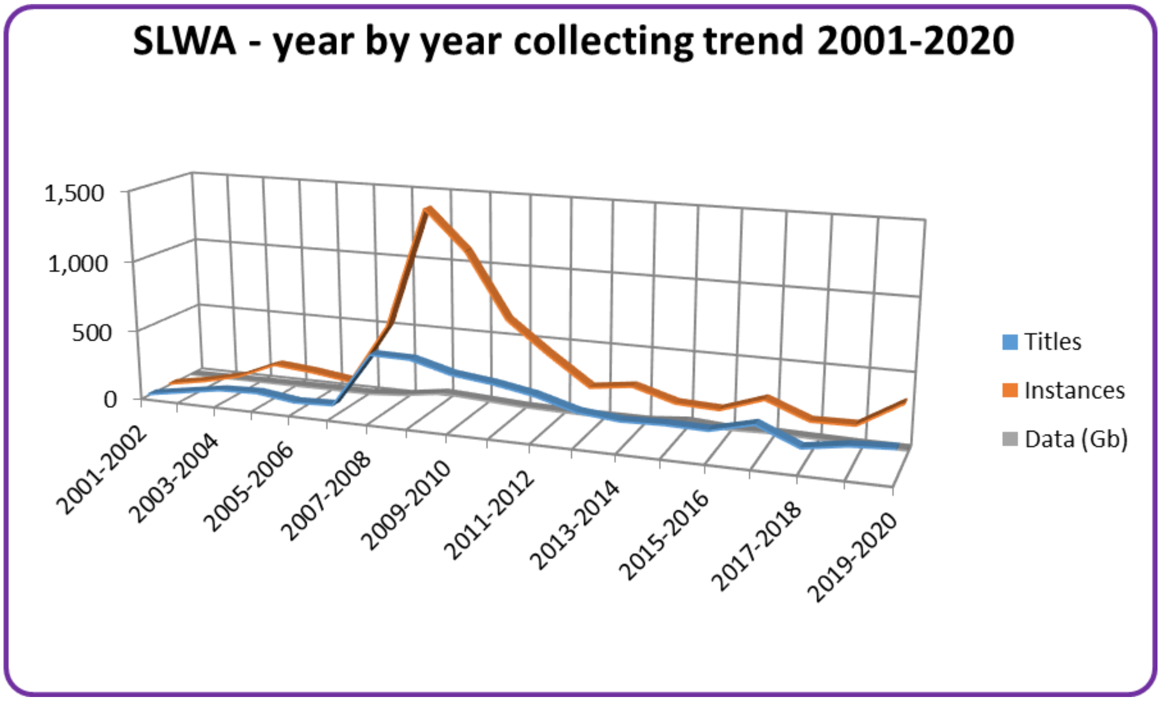


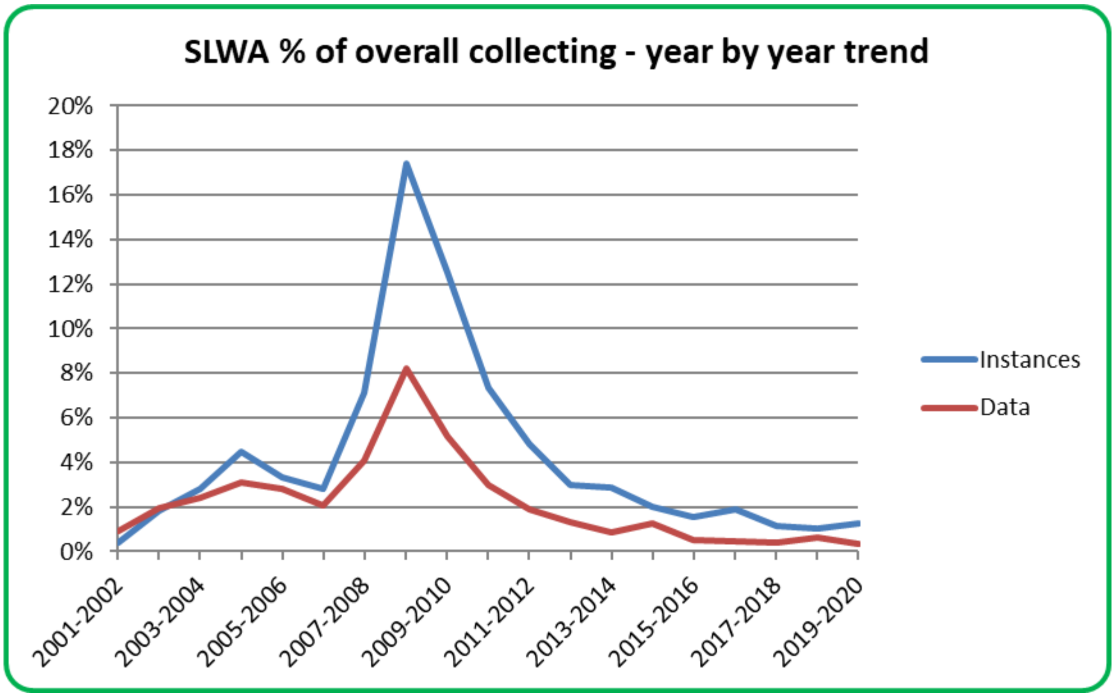
State Library of South Australia, 2000-2020



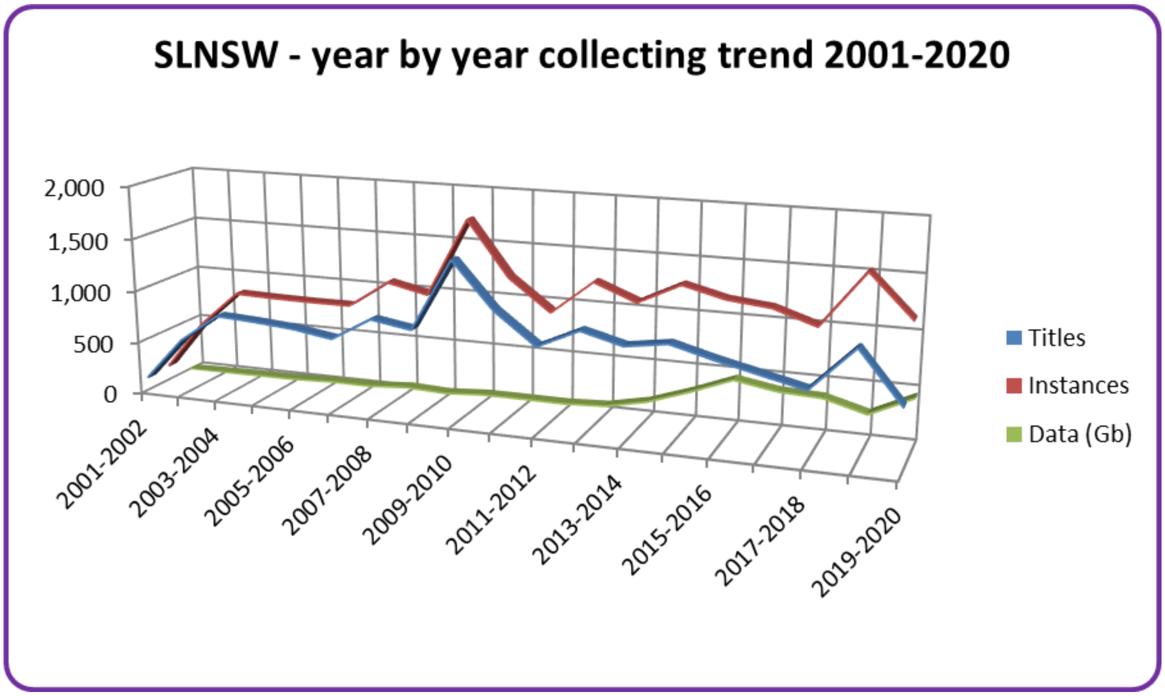


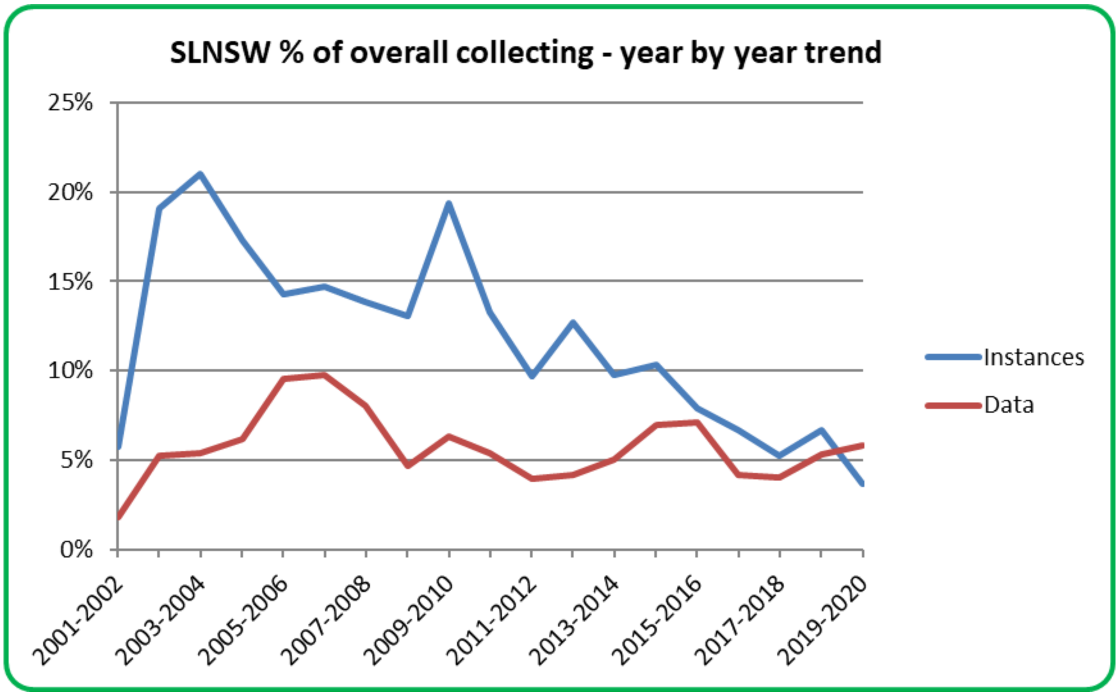
State Library of Western Australia, 2001-2020



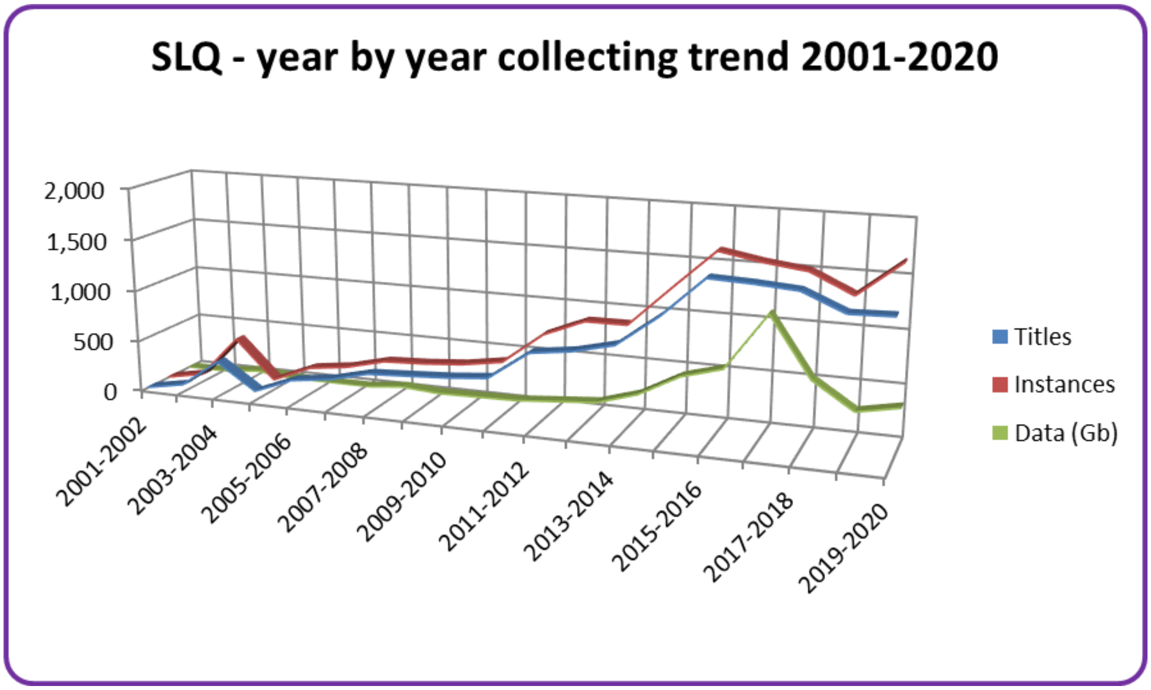


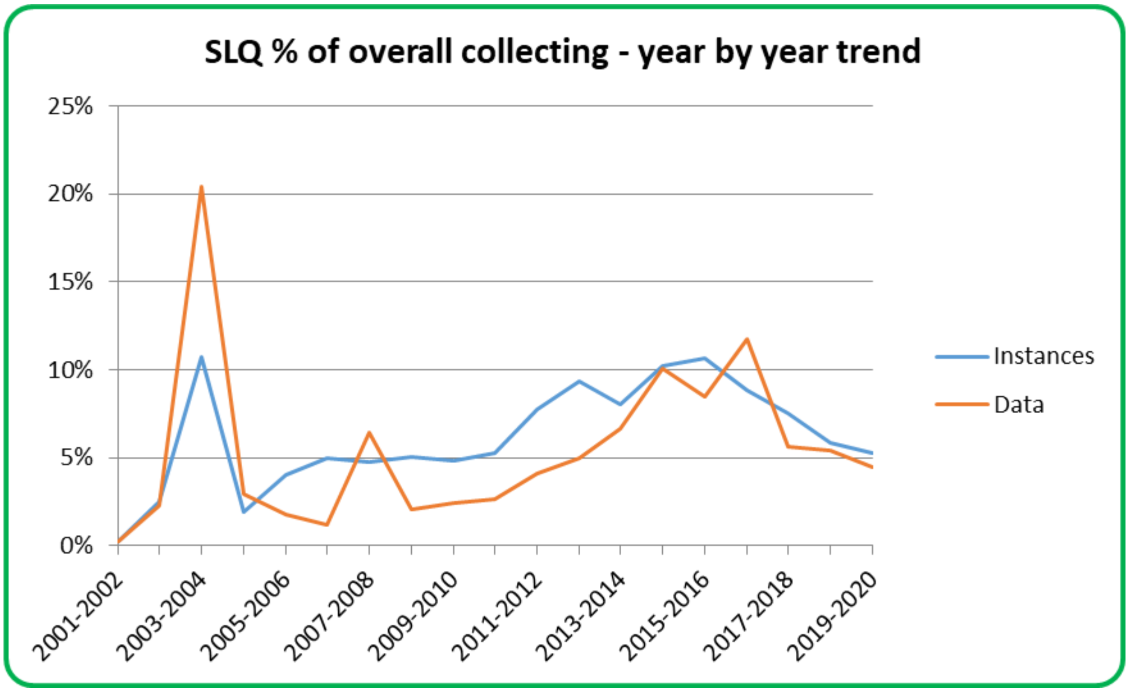
State Library of New South Wales, 2001-2020



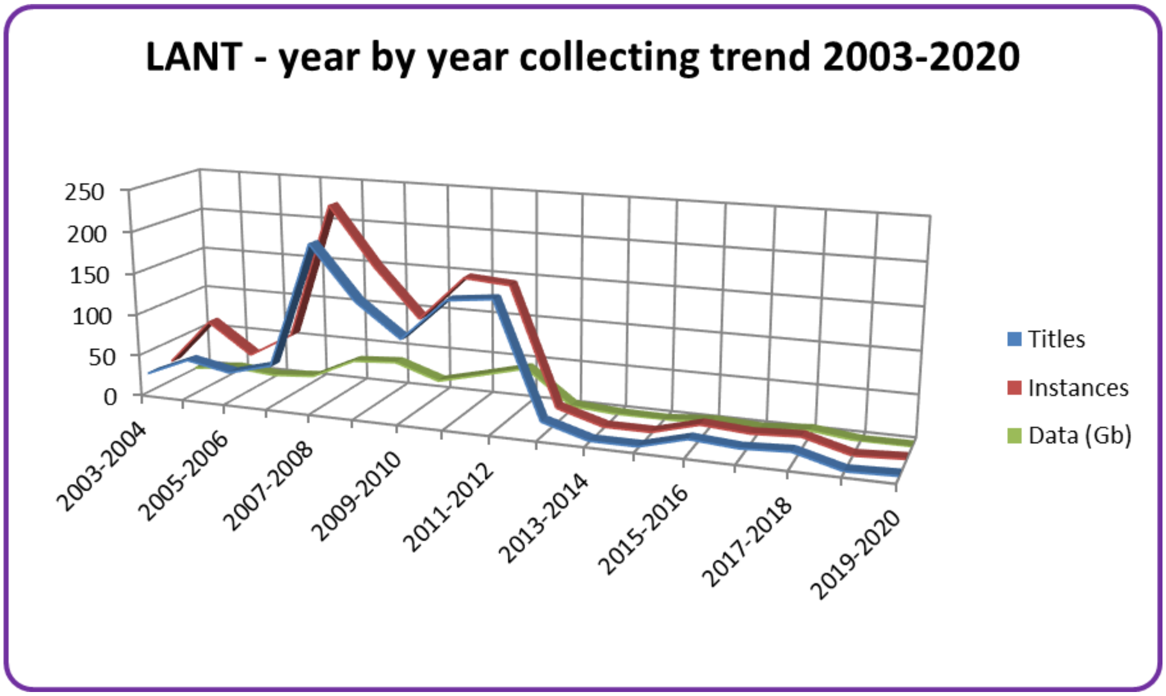


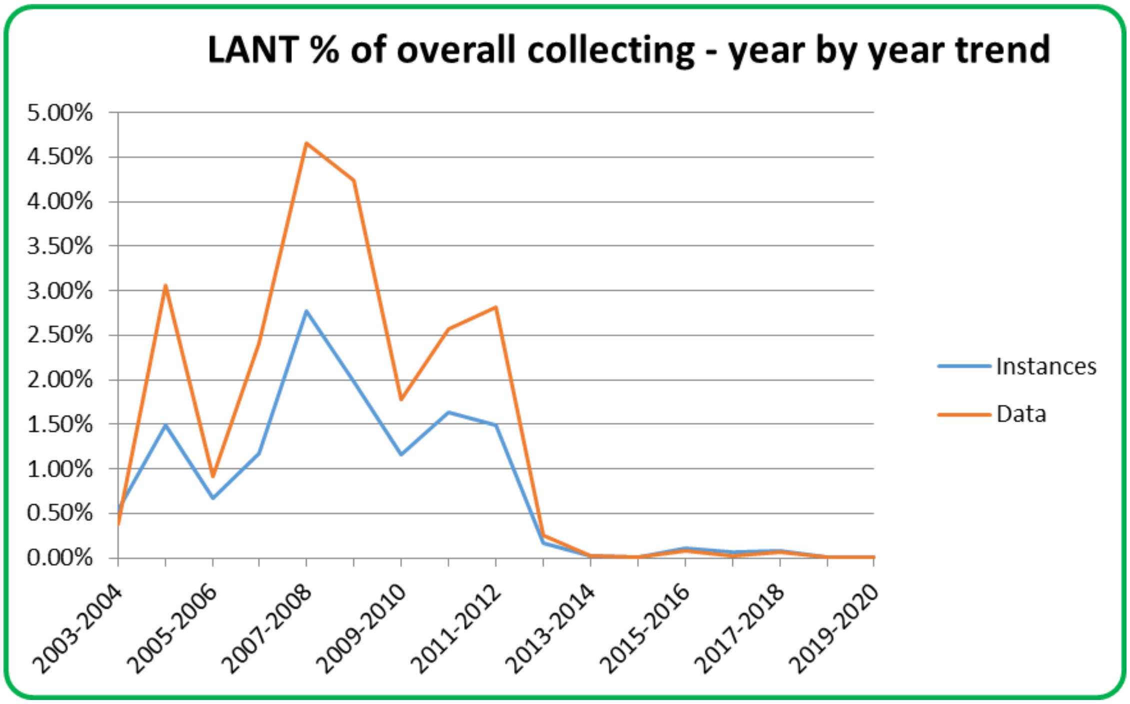
State Library of Queensland, 2001-2020



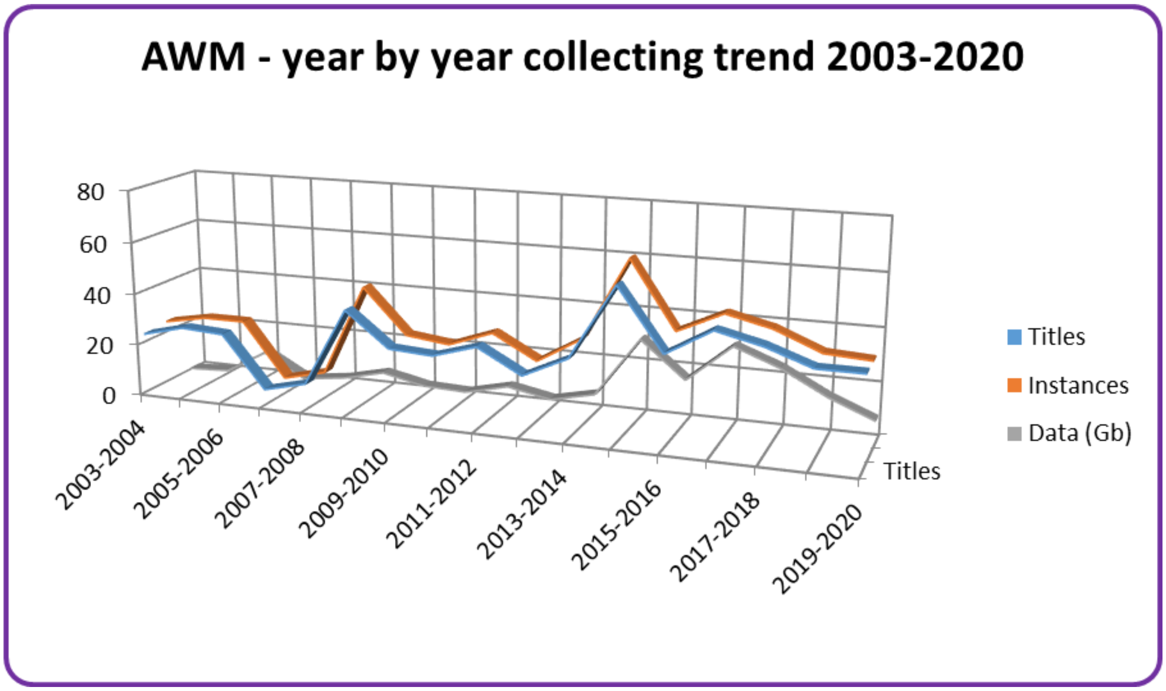


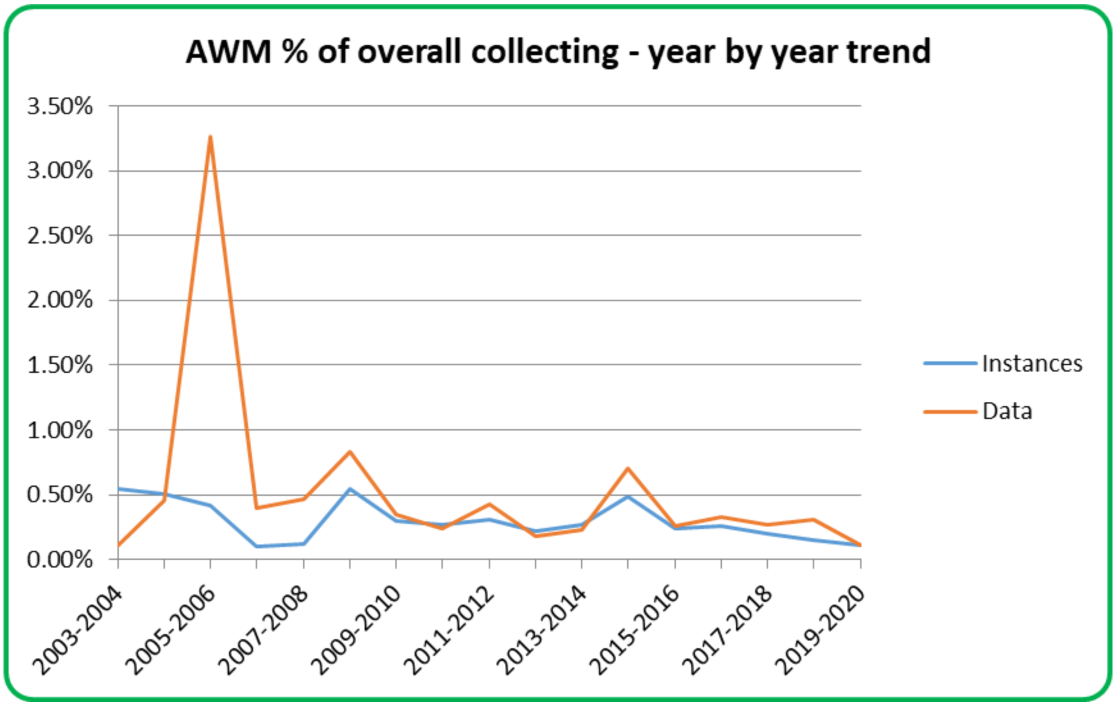
Library & Archives NT, 2003-2020



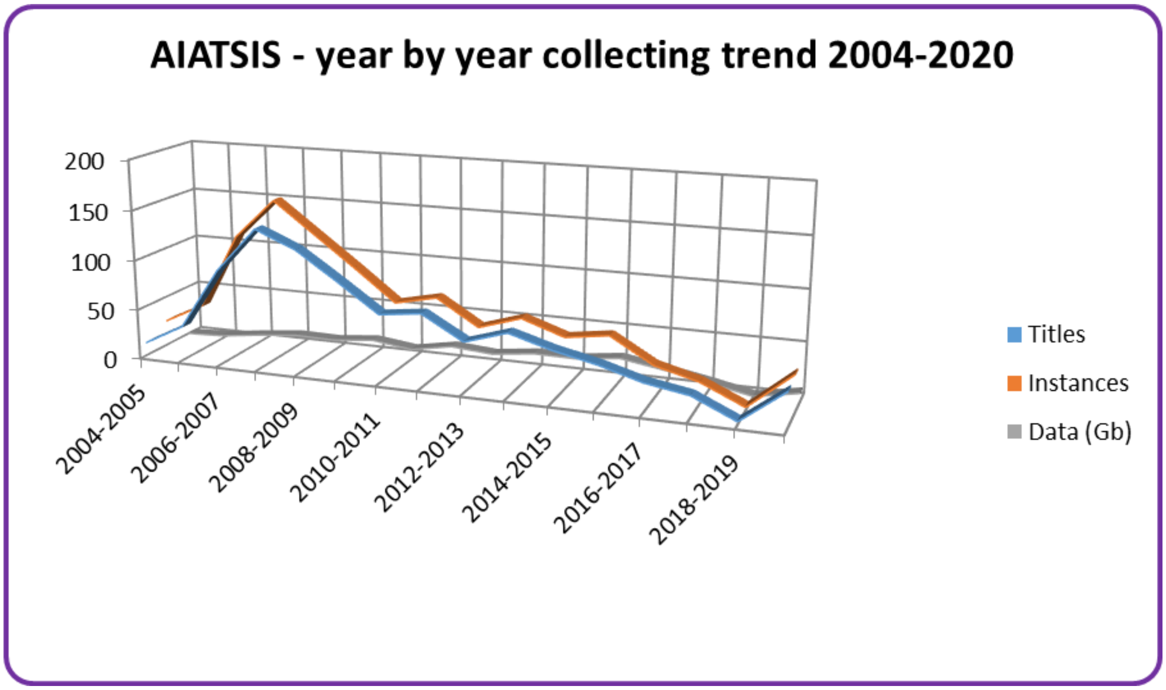


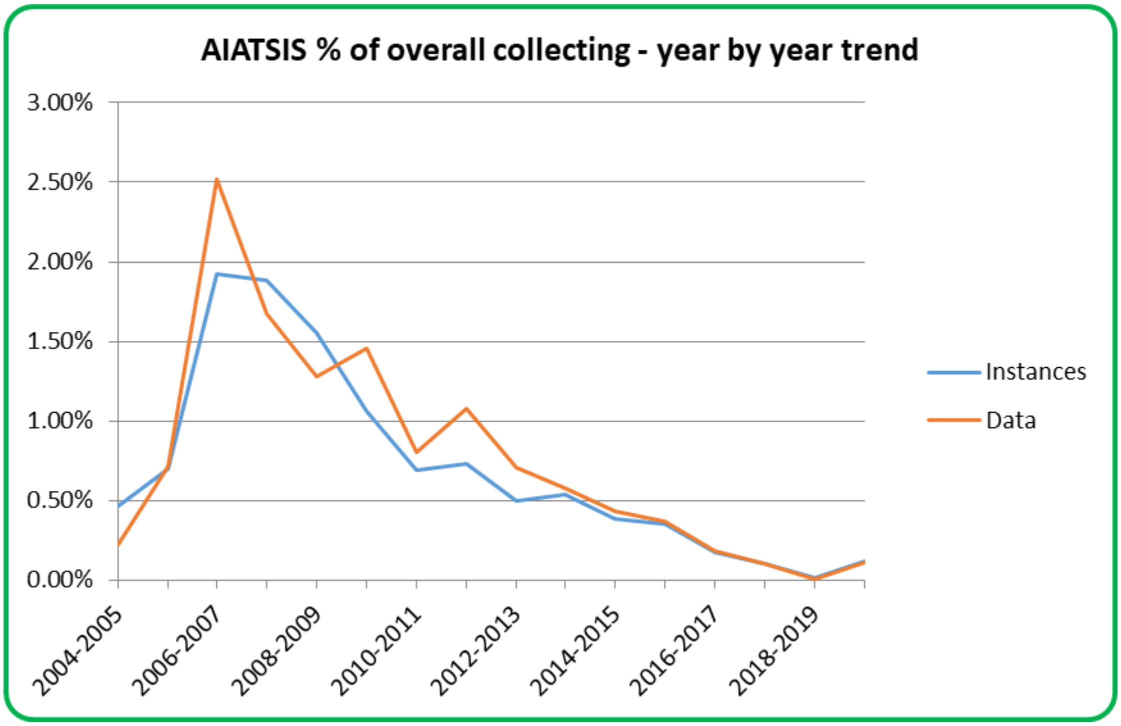
Australian War Memorial, 2003-2020



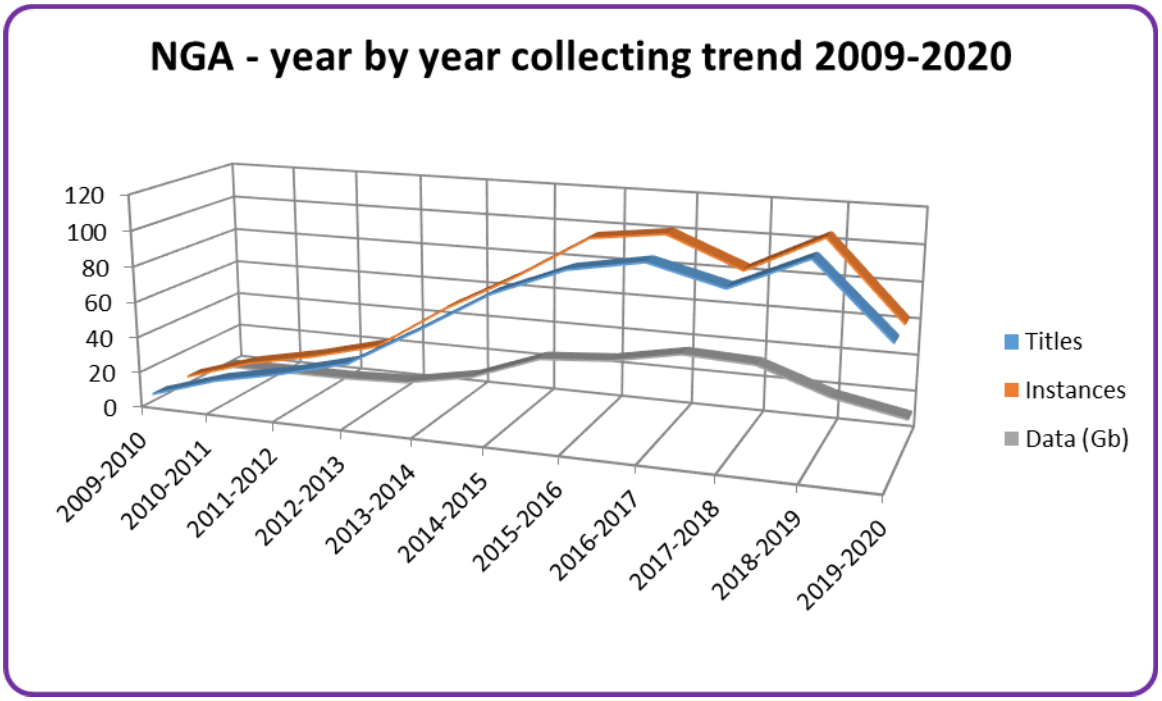


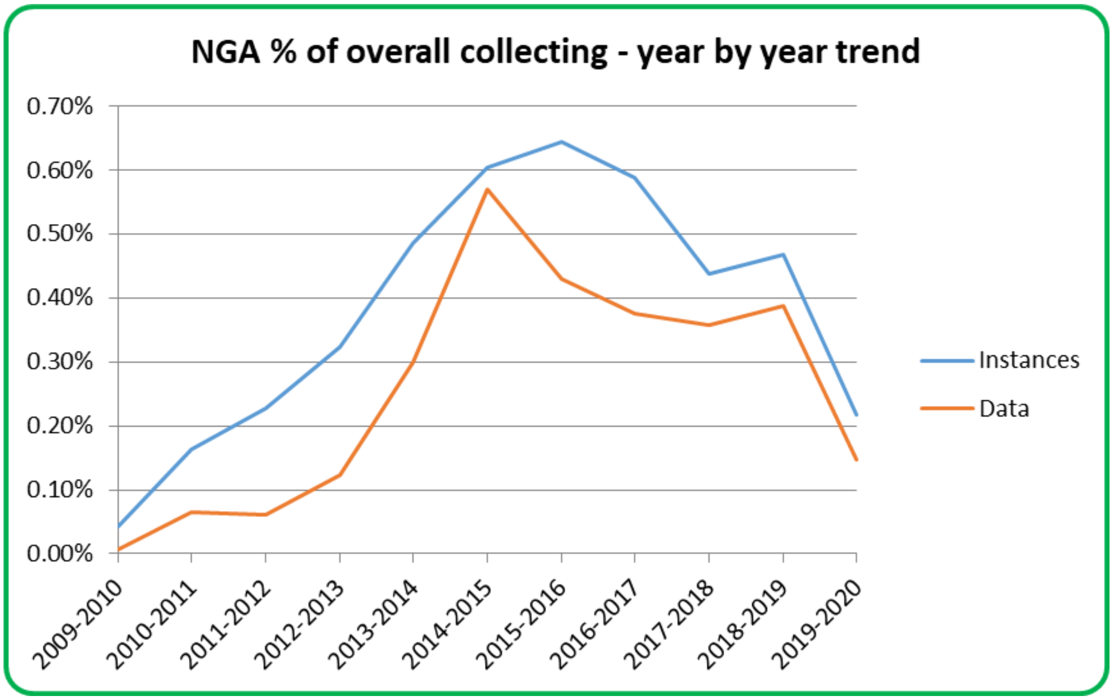
Australian Institute of Aboriginal and Torres Strait Islander Studies, 2004-2020





National Gallery of Australia, 2009-2020





4. Development of the Australian 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.

4.1 Development of PANDAS and tools supporting partners

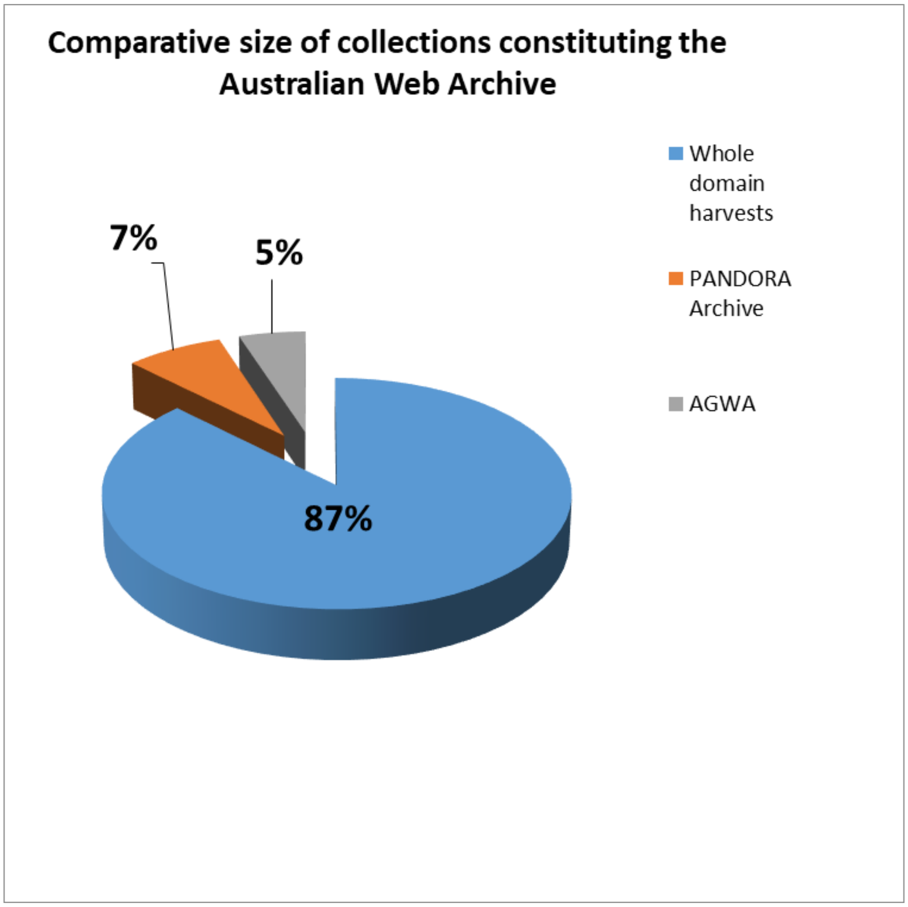
A number of developments were made by the Web Archiving Systems lead during the year that improved the workflow tools supporting partner contributions to the PANDORA web archive and improved delivery. The more significant developments included:

- Adding Heritrix as a harvester option to PANDAS.
- Providing partners with access to Webrecorder.
- Providing partners with the facility to upload WARC packages to the Australian Web Archive.
- Changes to PANDAS to facilitate partners' ability to collect content under Commonwealth Legal Deposit arrangements.
- Migrating the web archive replay tool from OpenWayback to Pywb in February 2020 which improved delivery, resolved some sandbox issues and made the delivery infrastructure supportable and sustainable.

4.2 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 Australian (.au) domain harvest collections.

The following pie chart shows the relative size (as a percentage) of the three collections that comprise the Australian Web Archive.



4.3 Australian web domain harvest

In the first quarter of 2020 the Library conducted the 15th large-scale harvest of the Australian web domain. This was the fifth 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 engineer the whole domain harvest crawl. The Internet Archive has the infrastructure, expertise and experience in this method of large scale web harvesting.

The harvest was run during the period from March to May 2020 and nearly 700 million unique documents were captured, amounting to 75 terabytes of data from more than two and a half million host domains.

Following this harvest, the combined total for all 15 annual Australian domain harvests has now reached nearly 11 billion files amounting to around over 606 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
2020	697 m	2,762,079	75.2

4.4 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 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. Bulk harvests of government websites typically include the collection of between 600,000 and 800,000 PDF documents.

5. *Focus on users*

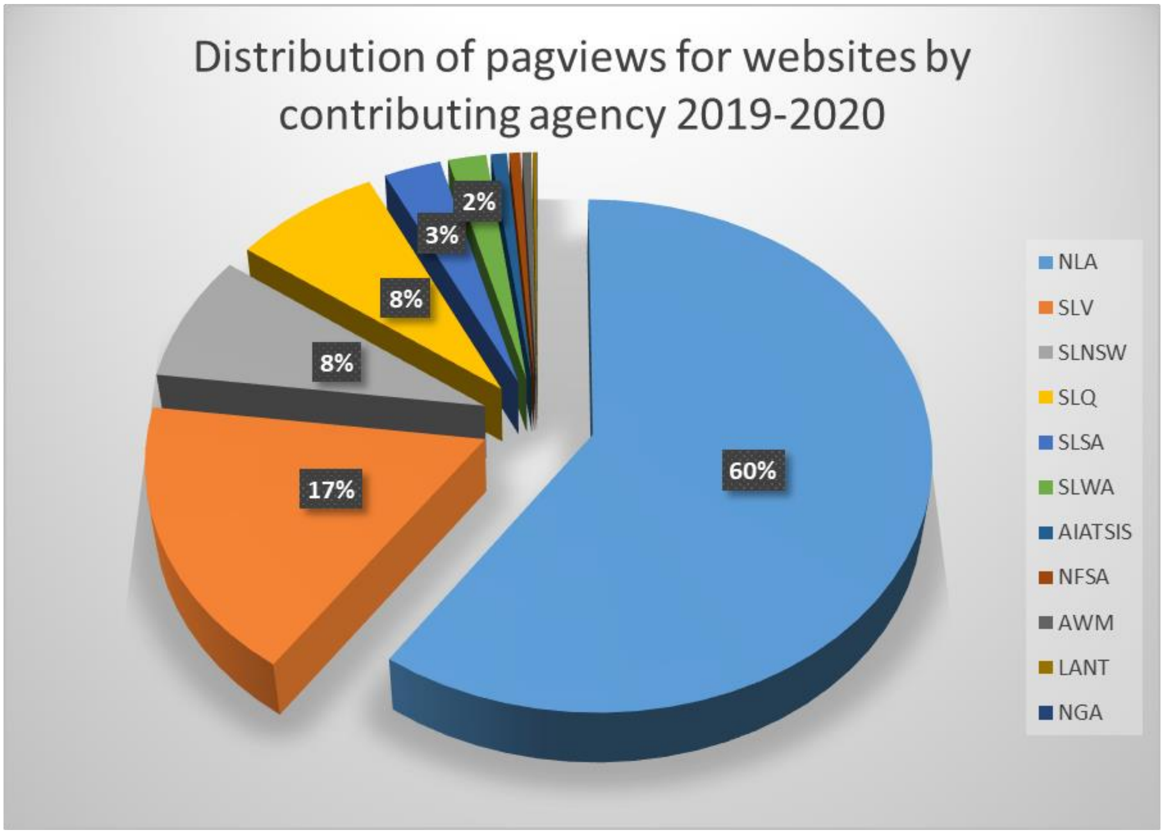
The Library uses Google Analytics reporting to record usage of the Trove Australian Web Archive content which includes both the PANDORA and the Australian Government Web Archive collections.

5.1 User views of the Trove Australian Web Archive

Usage for the period 1 July 2019 to 30 June 2020

Total page views	Number of users	Average views per month	Average pages viewed per visit
1,771,381	242,223	147,615	4.59

The following chart shows the percentage breakdown of user page views of content contributed by each partner agency.



5.2 Most viewed titles (websites) and collections 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.

The following table shows the top 20 PANDORA partner contributed websites (or titles) accessed in 2019-2020.

	Archived Title (website)	Participant Responsible
1	**First families 2001**	SLV
2	**ARIA report (Australian Record Industry Association)**	SLNSW
3	**APS Jobs Gazette**	NLA
4	**Footypedia**	NLA
5	**Legal Services Commission of South Australia**	SLSA
6	**Johnny's Pages: Old S.A.R. Shunter's Memories**	SLSA
7	**Queensland Literary Awards**	SLQ
8	**Daily weather observations (Bureau of Meteorology)**	NLA
9	**2009 Victorian Bushfires Royal Commission**	SLV
10	**Sydney Centre for Studies in Caodaism**	NLA
11	**News.com.au (November 2019)**	NLA
12	**Prime Minister of Australia – Malcolm Turnbull**	NLA
13	**Australian Sheep Industry CRC**	NLA
14	**AIATSIS Research Discussion Papers**	AIATSIS
15	**Victorian Essential Learning Standards**	SLV
16	**National Library of Australia News**	NLA
17	**Bunyips (NLA Exhibition)**	NLA
18	**The Playground (ABC)**	NLA
19	**Learn About Meteorology (Bureau of Meteorology)**	NLA
20	**Index to NSW Regulations and Selected Ordinances**	SLNSW

The following table shows the 10 most accessed PANDORA Collections during 2019-2020.

	Collection	Participant/s Responsible
1	**Olympic Games – 2000, Sydney**	NLA
2	**Coronavirus (COVID-19) Pandemic and Australia, 2020**	Collaborative
3	**Historic Gold Mining Plots**	SLV
4	**National Library of Australia Online Exhibitions**	NLA
5	**Bushfires, Australia (2019-2020)**	Collaborative
6	**Iconic Australian Brands**	Collaborative
7	**Australian Broadcasting Commission (ABC) websites**	NLA
8	**Australian Prime Ministerial websites**	NLA
9	**Election Campaigns**	Collaborative
10	**Australian Federal Government Ministers**	NLA

6. International relations

6.1 International Internet Preservation Consortium (IIPC)

- The National Library continued in its role as a member of the IIPC Steering Committee with Paul Koerbin as the Library's designated representative.
- Paul Koerbin was elected Vice-Chair of the IIPC, one of the three officer roles in the Consortium, for the 2020 calendar year.
- Alex Osborne continued his active involvement in the IIPC's Tools Portfolio leadership team.
- The IIPC General Assembly and Web Archiving Conference scheduled for April in Montreal was cancelled due to COVID-19. The WAC was not rescheduled but a much shortened virtual General Assembly was held in June.
- The National Library was a co-sponsor along with the NLNZ for a successful project proposed by the British Library for an IIPC discretionary funding project for Dr Tim Sherratt to develop a number of Jupyter Notebooks to encourage and facilitate researchers' use of web archive content. The Notebooks are available on Dr Sherratt's GLAM Workbench github space.

7. Promoting the Archive

7.1 Presentations, representations and papers

Major presentations, papers and representations during the 2019-2020 financial year included:

- Alex Osborne and Paul Koerbin gave a number of presentations at the NSLA Web Archiving meeting in September 2019 in Canberra. Presentations covered the current state of PANDORA, current initiatives, IIPC involvement, tools and infrastructure.
- Paul Koerbin contributed a chapter for a book edited by Daniel Gomes (Portuguese Web Archive) titled 'The Past Web' to be published in late 2020 by Springer. Dr Koerbin's chapter is titled 'National web archiving in Australia: representing the comprehensive'.
- Paul Koerbin used personal travel arrangements to attend the IIPC Steering Committee meeting hosted by Colombia University in New York in October 2019.

7.2 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.