Annual report to partners 2012-2013

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 National Film and Sound Archive, 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 partners on activities and developments in the 2012-2013 financial year.

1.1 Consultation mechanisms

The National Library continued to inform other PANDORA participants about the operation of PANDORA through the two email discussion lists and the 'PANDORA Wiki'.

In June 2013 the Manager Web Archiving and Senior Librarian Web Archiving started a weekly newsletter sent out mid-week via the PANDORA mail list. This brief newsletter is intended to keep partner agency staff informed of important news, useful reading materials, mentions of PANDORA in social media and the news, as well as providing a space to highlight matters relating to operational procedures (e.g. problems, tips, best practices). The newsletters can also be accessed on the PANDORA 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 partner 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 partners 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 partner agencies is prepared annually and is also made available on the PANDORA website Partners page http://PANDORA.nla.gov.au/partners.html.

[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 It includes the gathering of a monograph that has been archived once only, the first gathering of a serial title or integrating title (for example, a web site that changes over time), and all subsequent gatherings.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 Archive*

2.1 Size and annual growth of the Archive

The Archive maintained steady growth in 2012-2013, with the percentage growth rate for Titles and Instances and data size of a similar magnitude to the previous 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 2012	30 June 2013	Growth 2012-13
Titles	31,421	34,694	3,273 (10.4 %)
Instances	76,439	86,977	10,538 (13.8 %)
Terabytes[2]	6.79	8.74	1.95 (28.7 %)

Government publications remain a substantial component of the collecting focus and comprise approximately 56 % of the titles in the Archive.

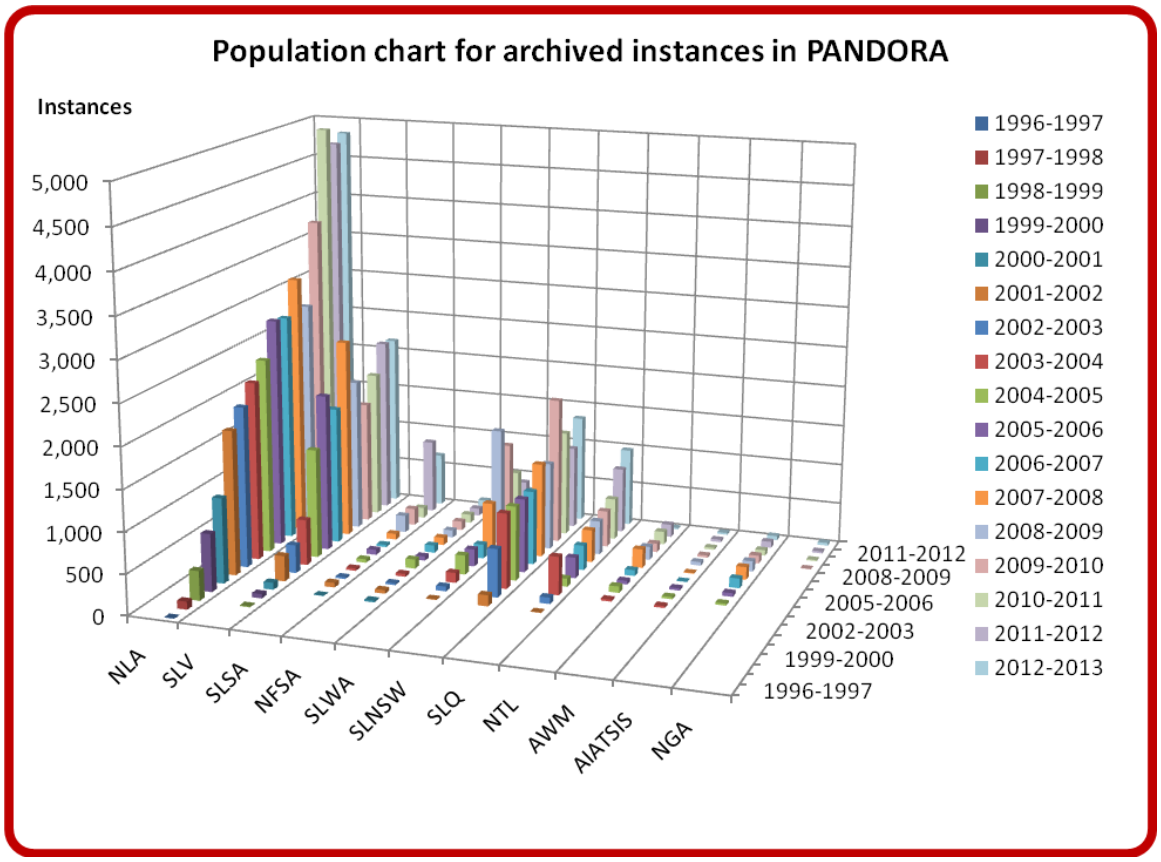
2.2 Select analysis of archival content

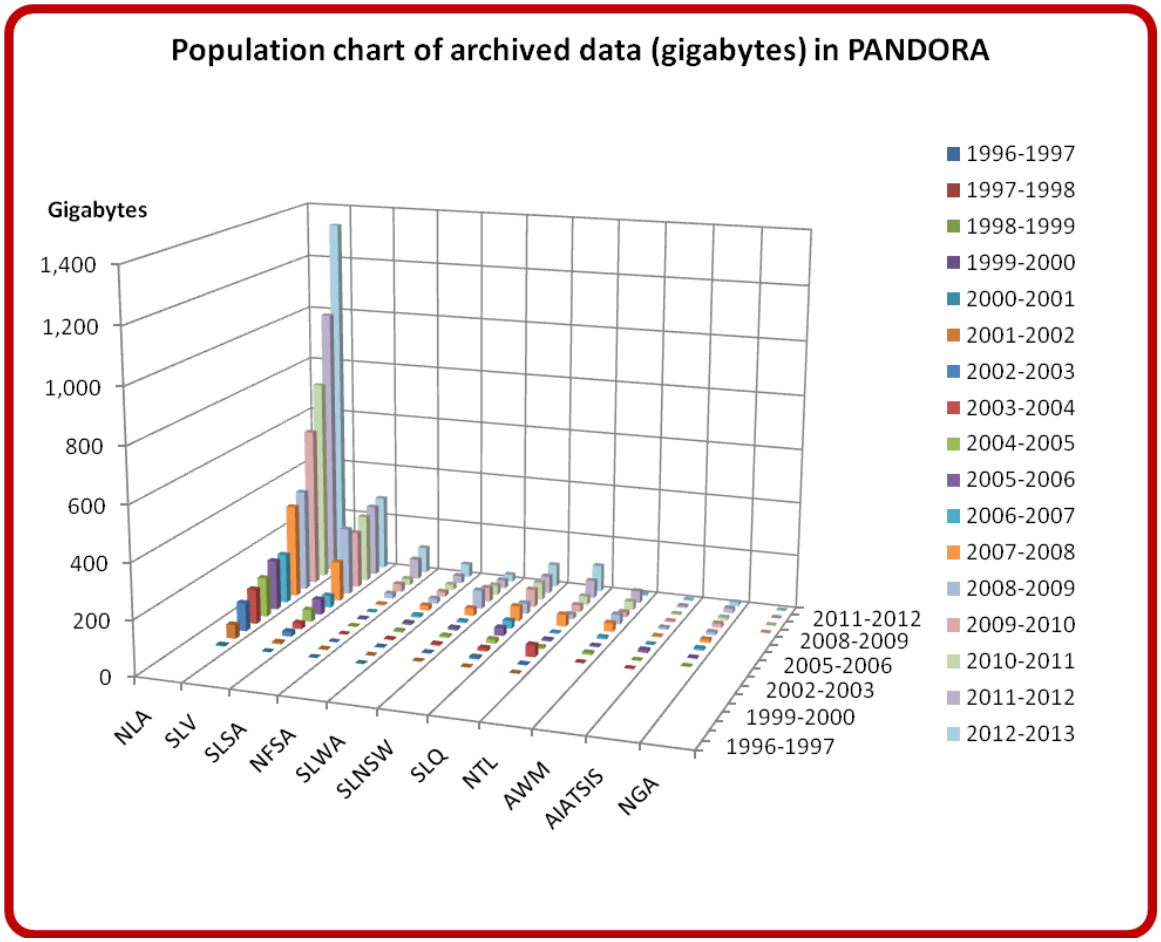
This year an analysis was undertaken of trends in respect to the scale of content collected for the archive by PANDORA partner agencies. The analysis looks at the year-by-year contributions of each partner agency showing the general trend in their archiving activity and the relative magnitude of titles, instances and data (gigabytes) collected. Longitudinal trend graphs are also provided showing the percentage of the contribution of each partner agency to each year's total PANDORA archiving activity.

[2] This figure does not include the preservation and other master and back up copies.

2.2.1 All content contributed over the life of the archive

The first two charts present a visualisation of how the archive has been populated with content since its inception in the 1996-1997 financial year by all contributing agencies. The first chart shows the contribution of archived instances by each partner agency. The second chart shows the amount of data (in gigabytes) contributed by each agency.

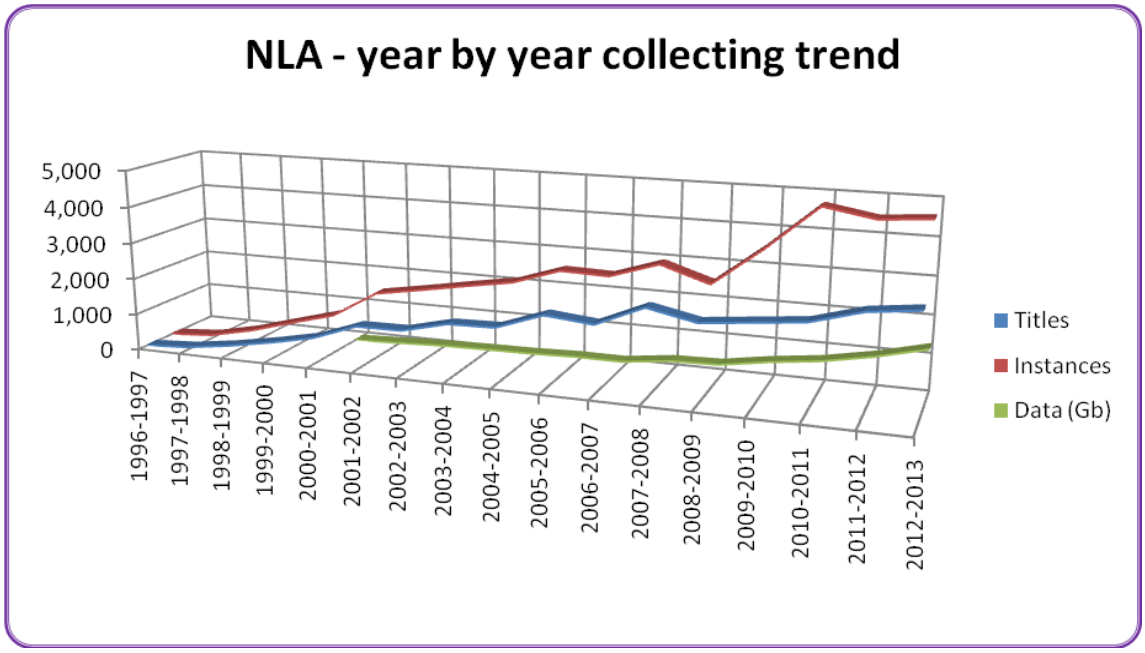


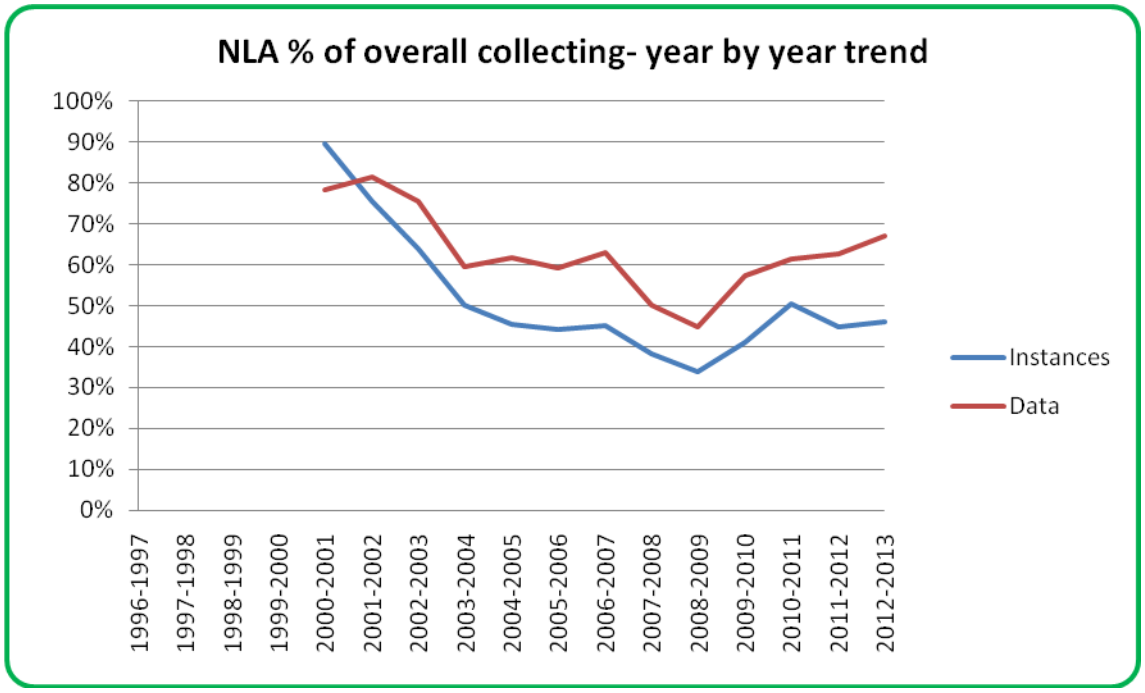


2.2.2 Individual partner contribution trends over the life of the archive

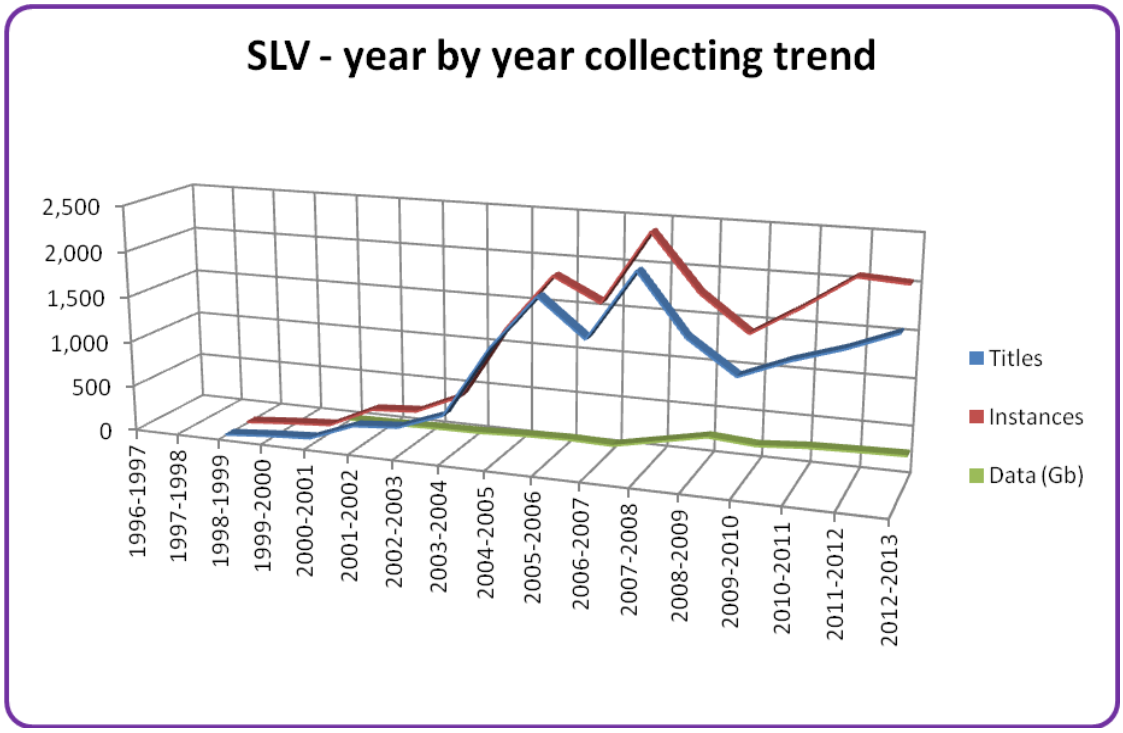
The following series of charts show the year-by-year collecting trends in respect to each participating PANDORA agency. The first chart for each agency shows the collecting trend measured by archived titles, instances and gigabytes. The second chart for each agency shows the year-by-year trend in respect 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.

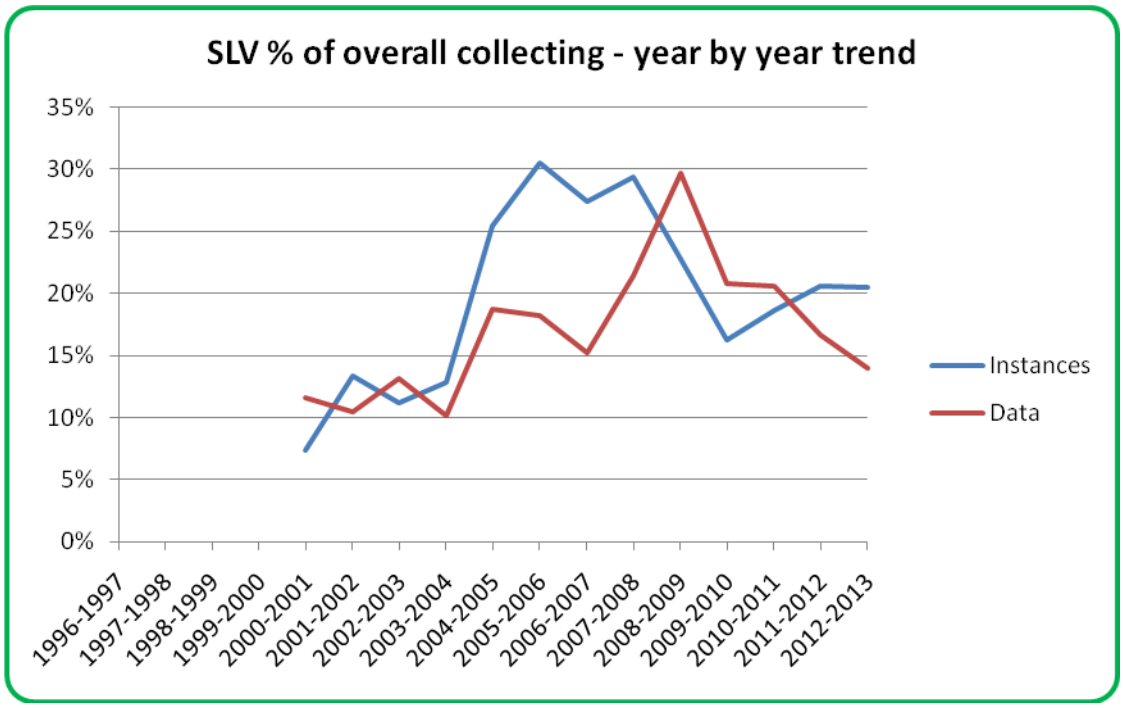
National Library of Australia (1996+)



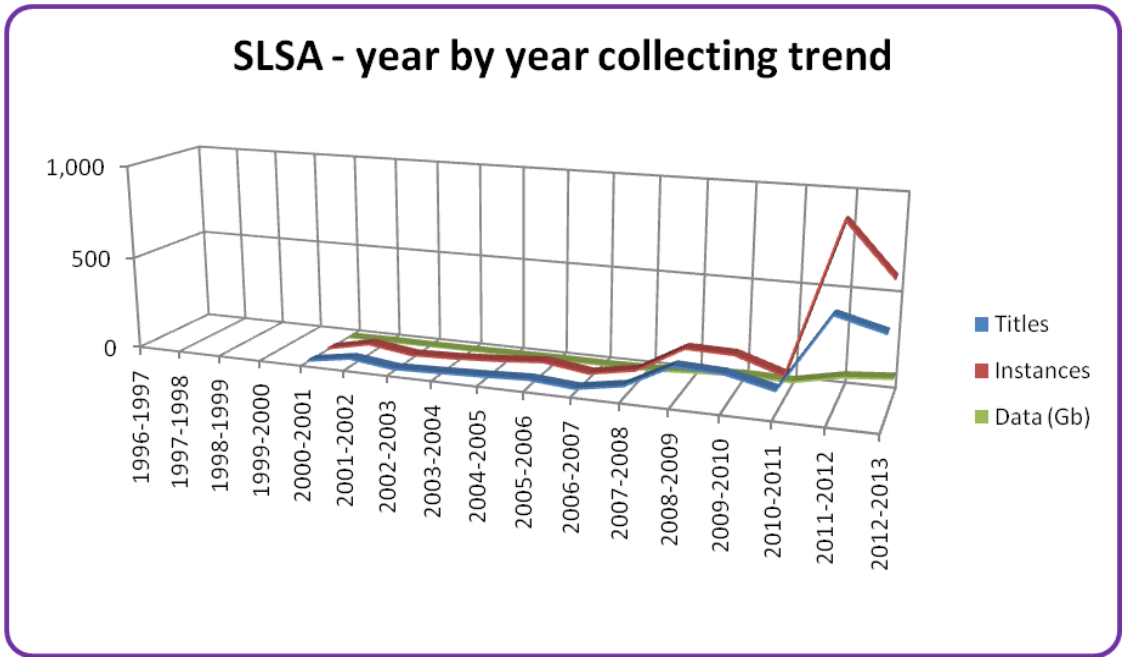


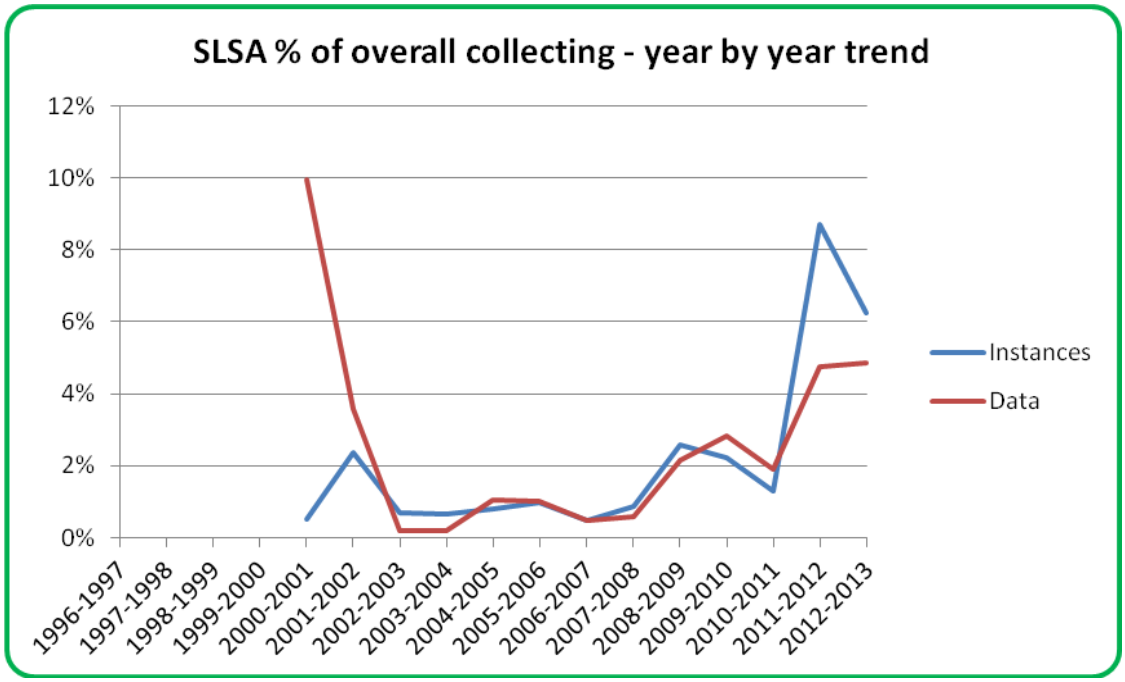
State Library of Victoria (1998+)



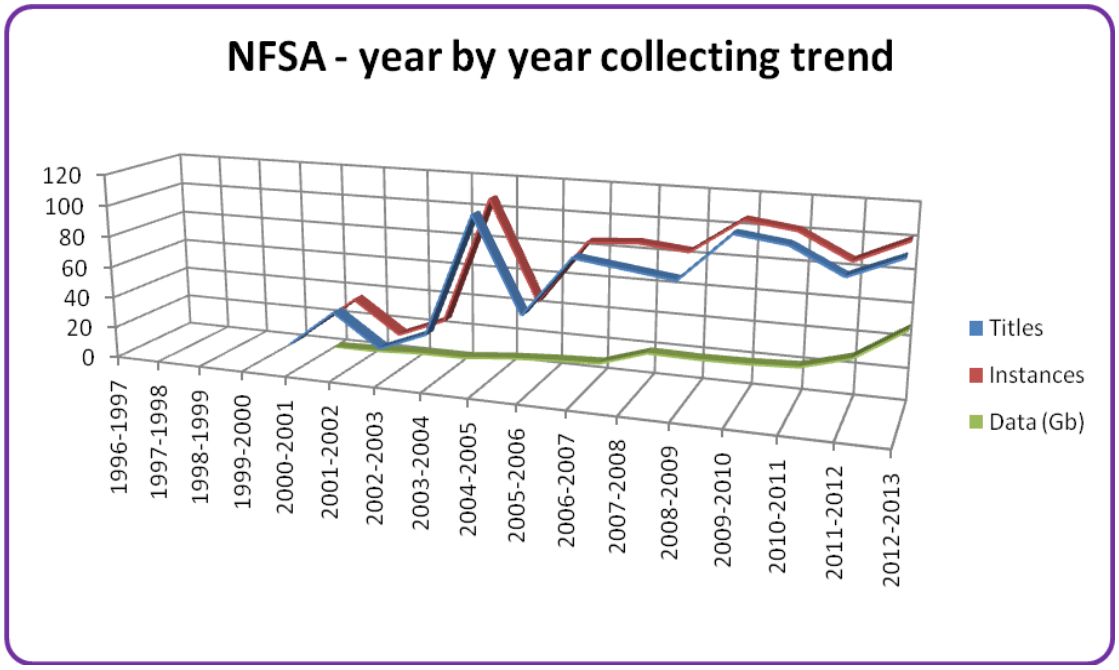


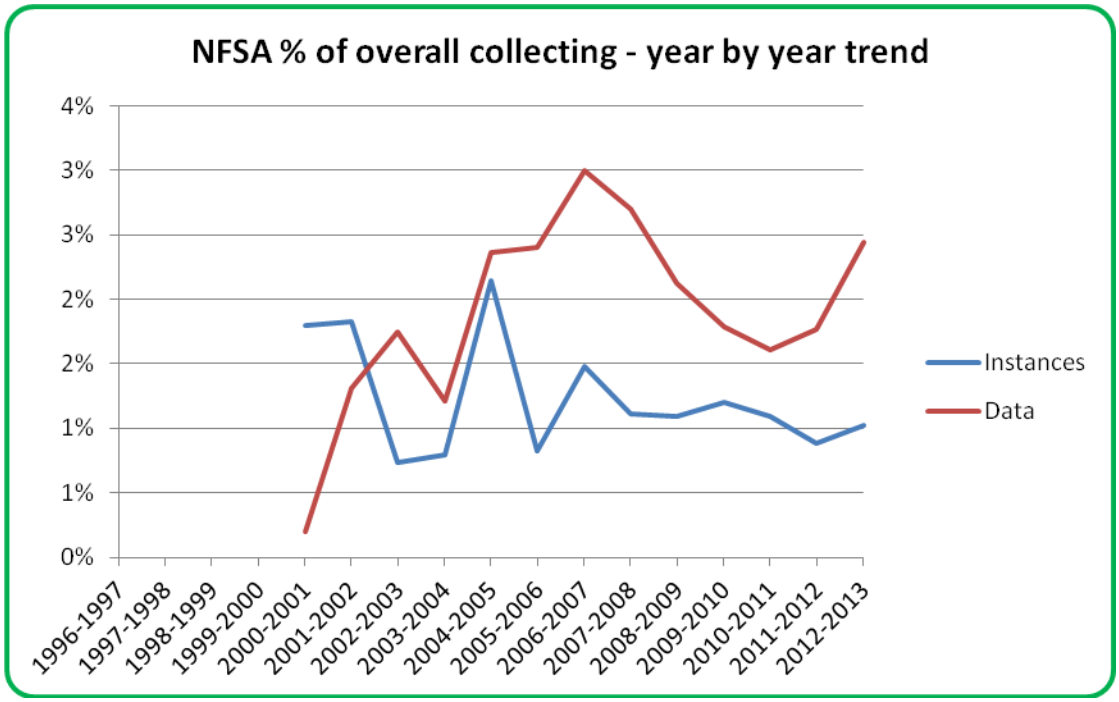
State Library of South Australia (2000+)



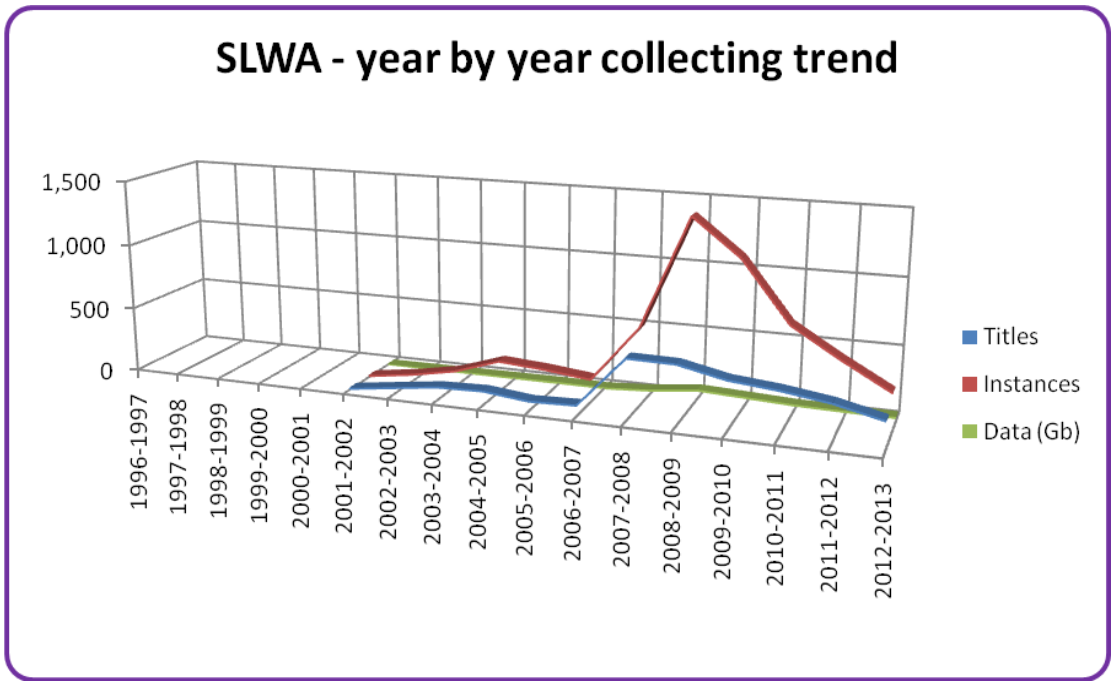


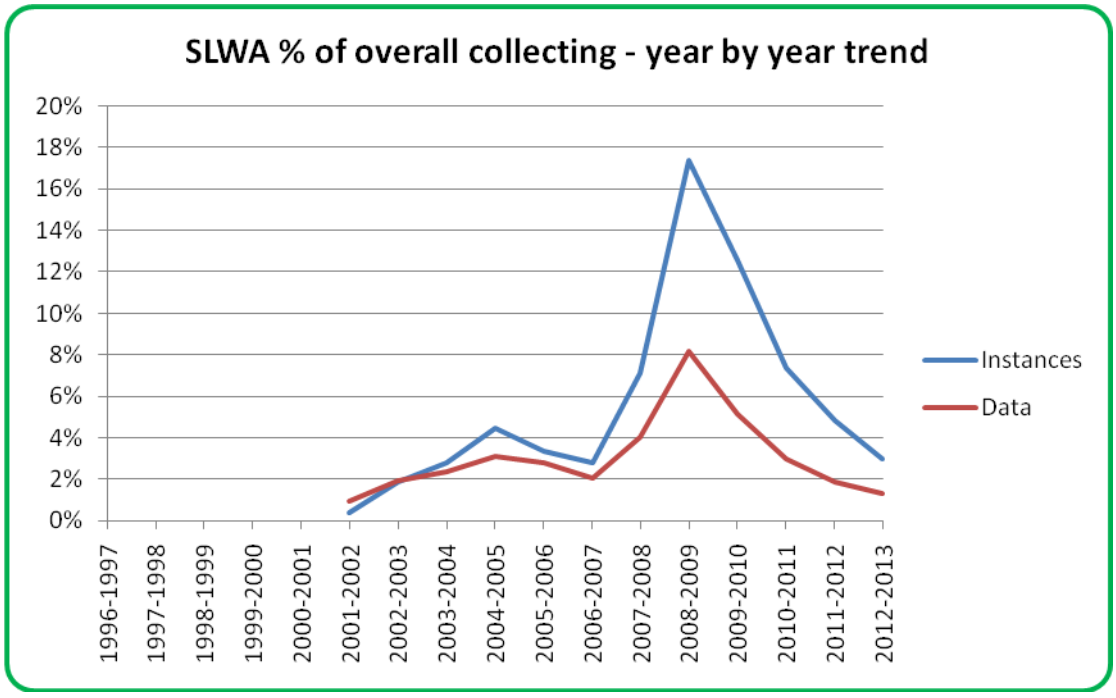
National Film and Sound Archive (2000+)



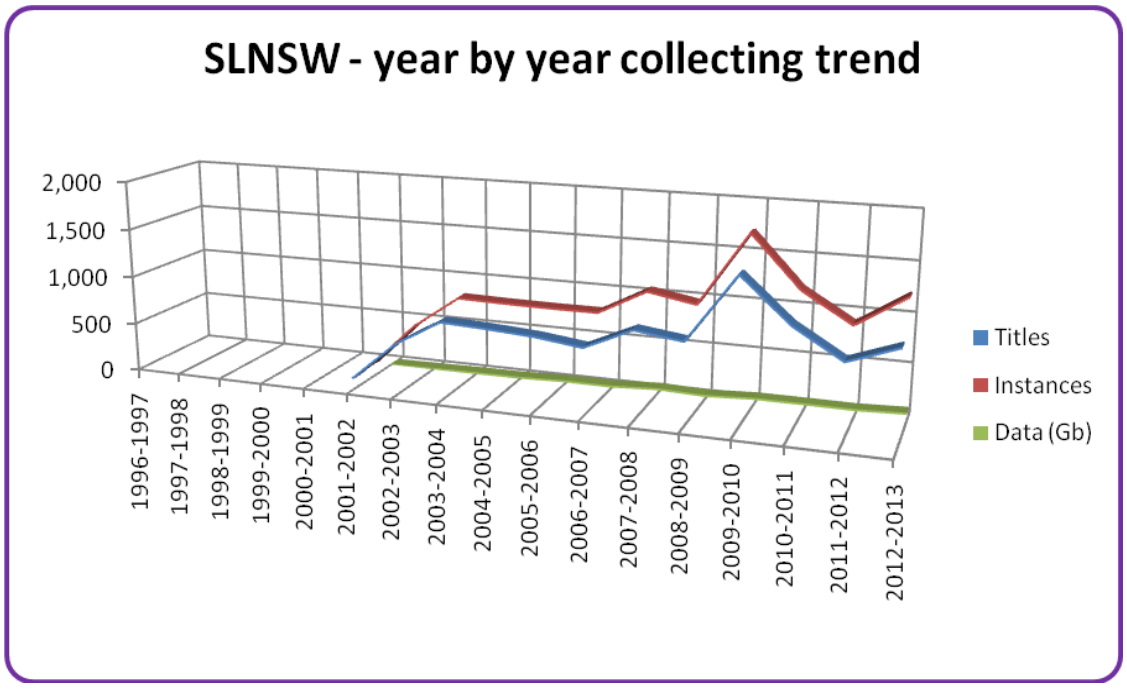


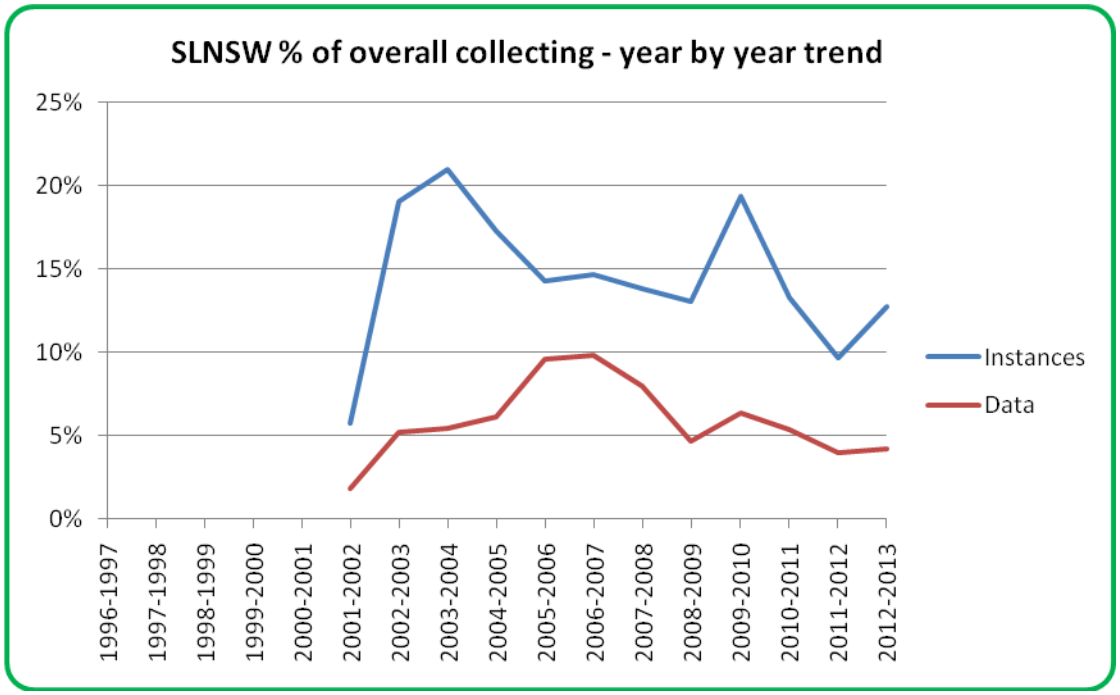
State Library of Western Australia (2001+)



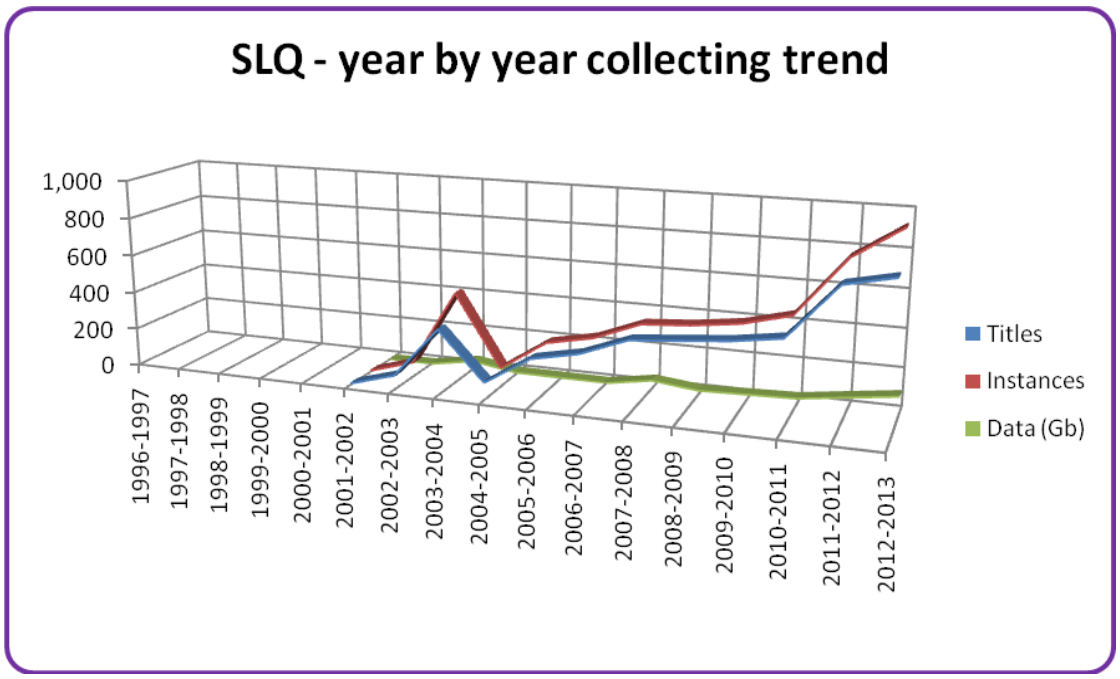


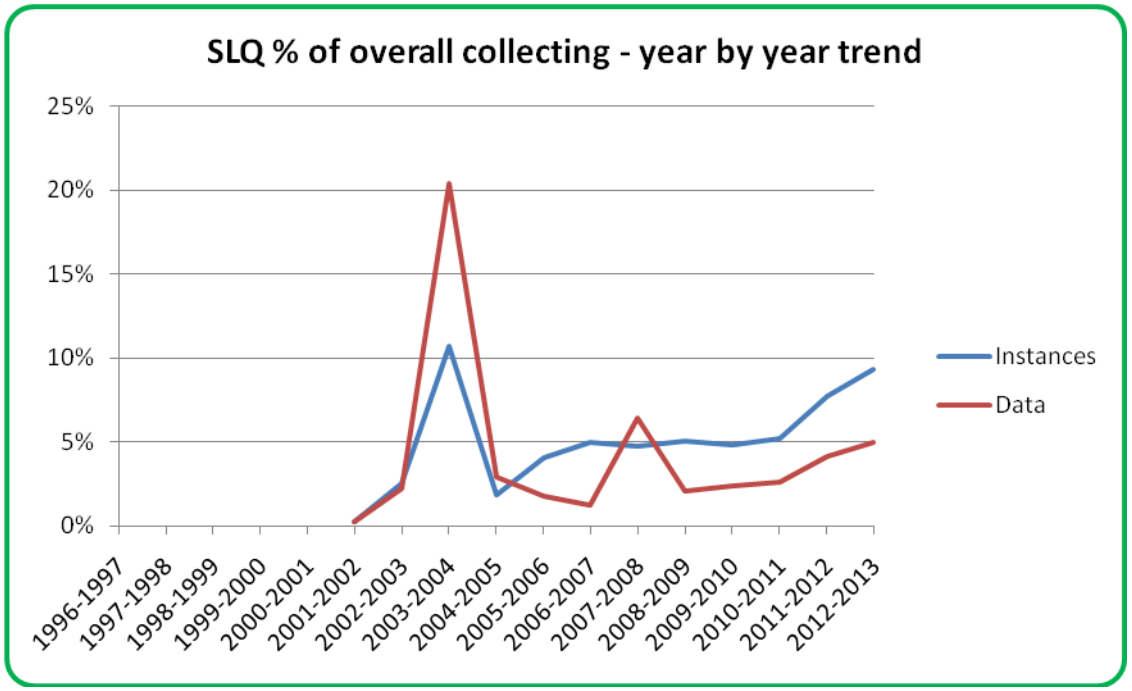
State Library of New South Wales (2001+)



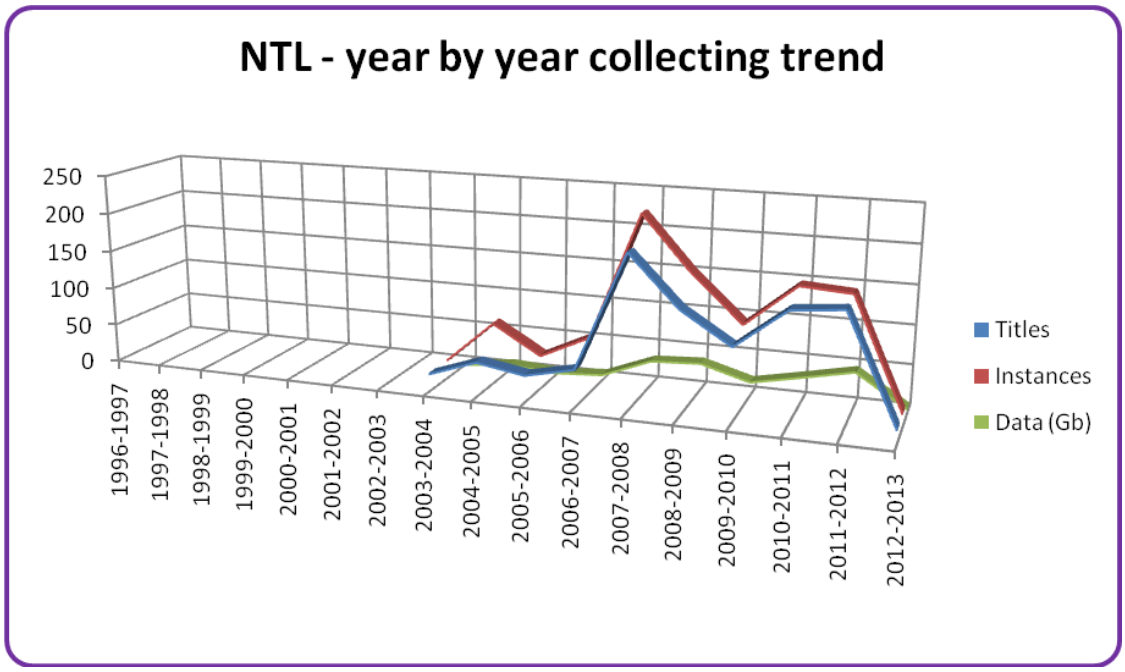


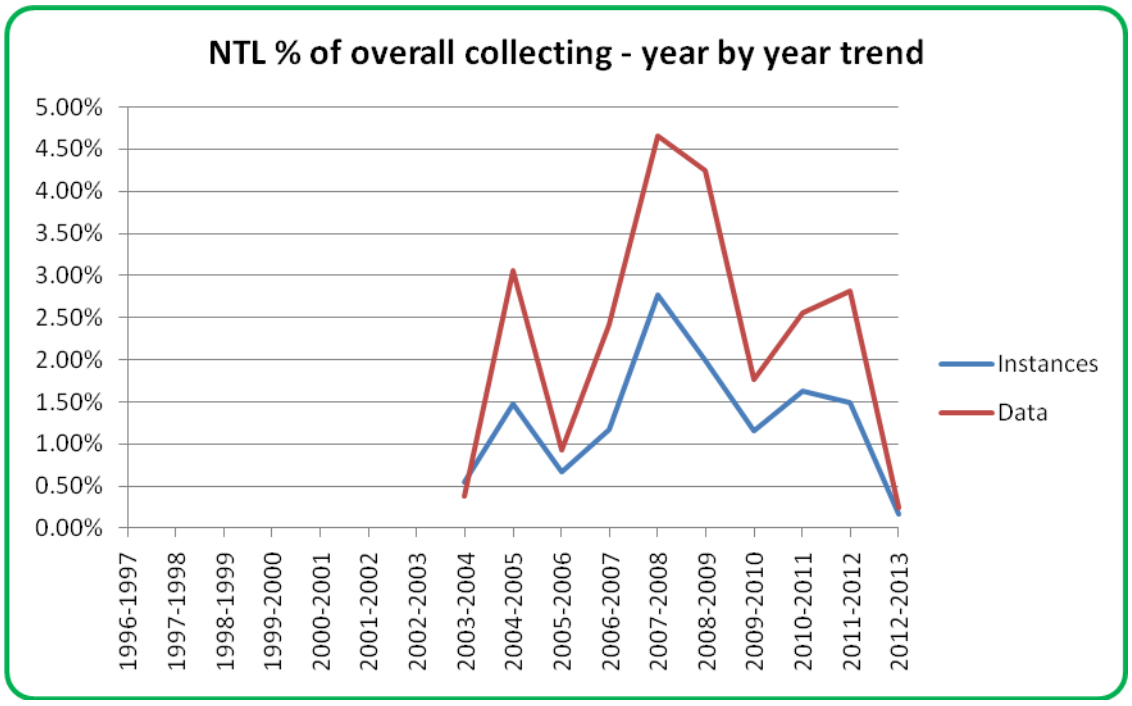
State Library of Queensland (2002+)



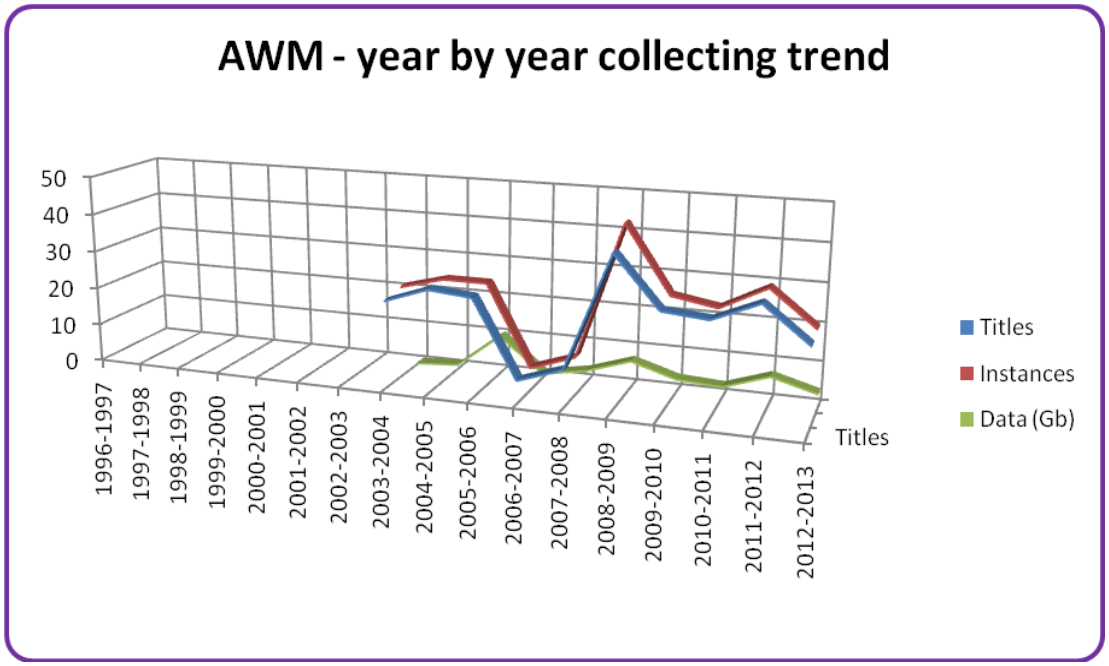


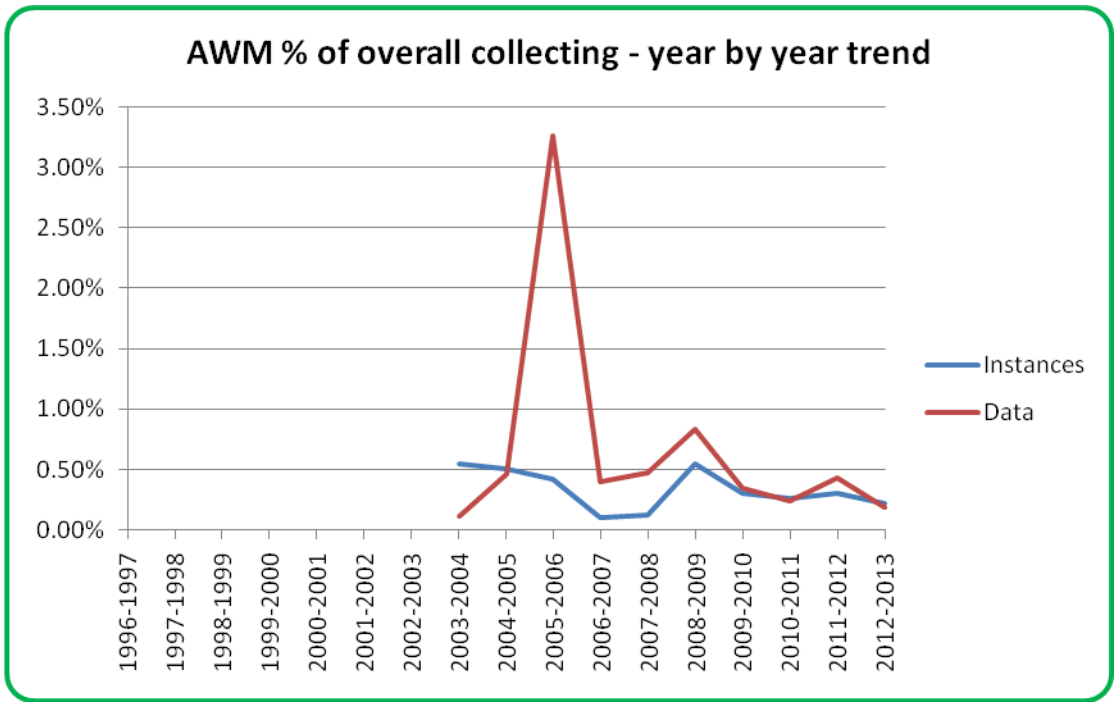
Northern Territory Library (2002+)



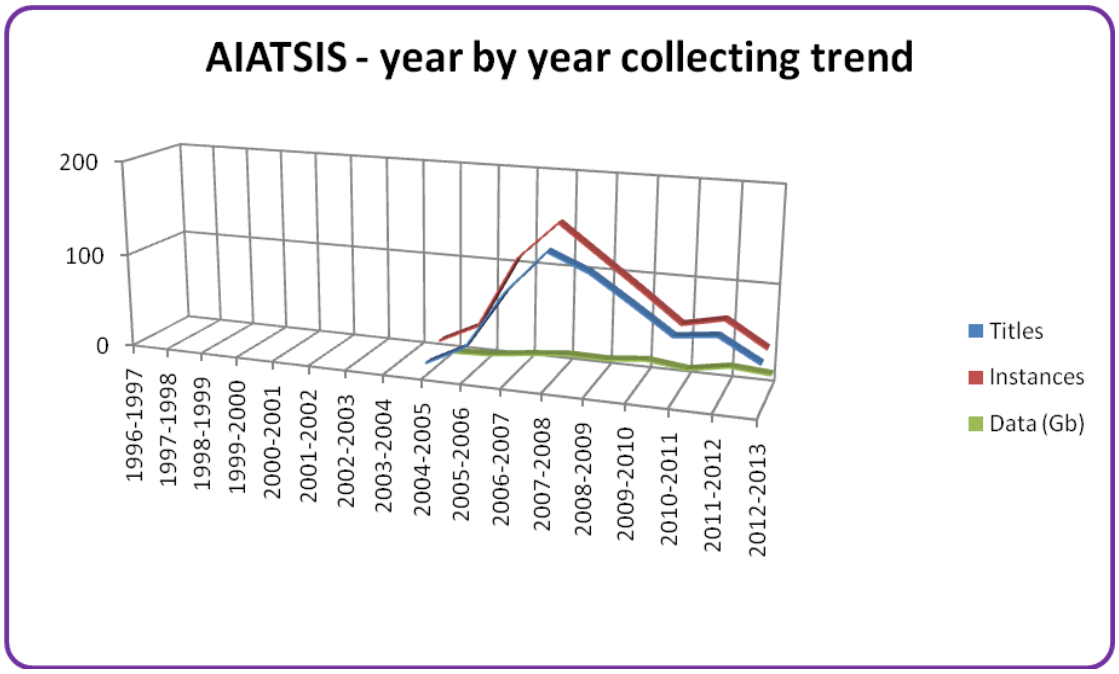


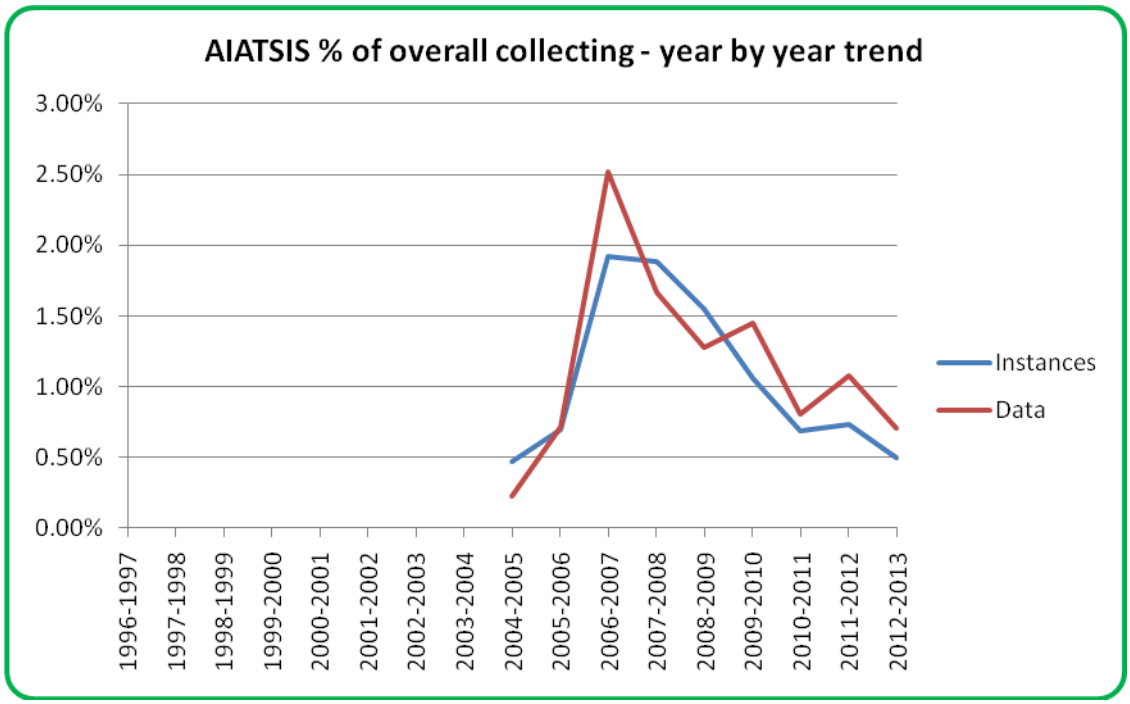
Australian War Memorial (2003+)



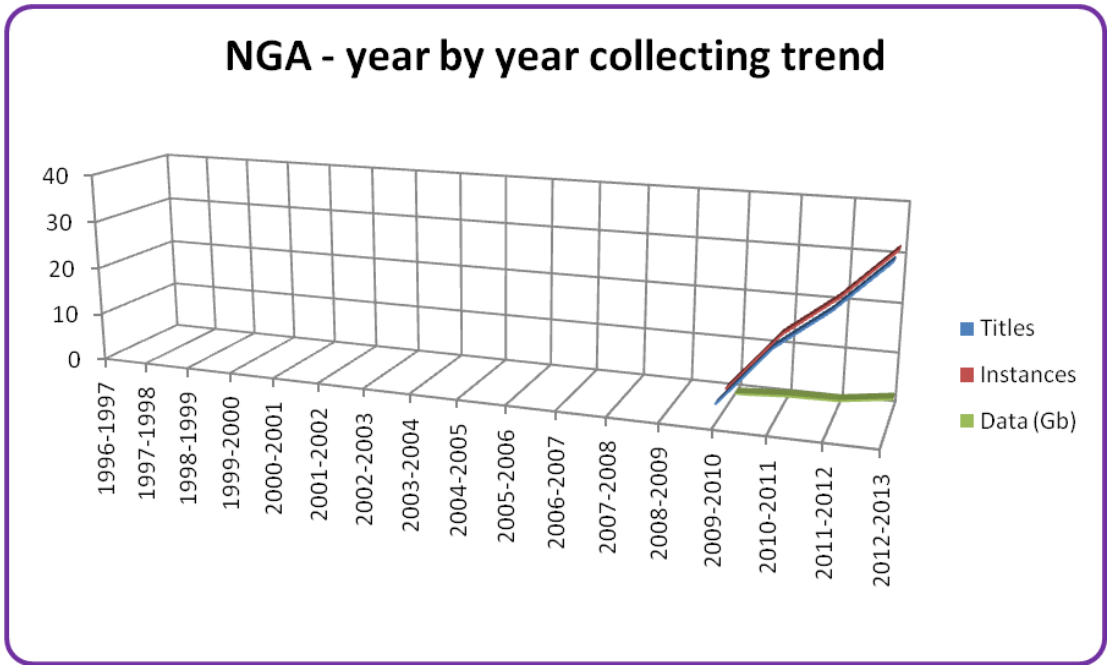


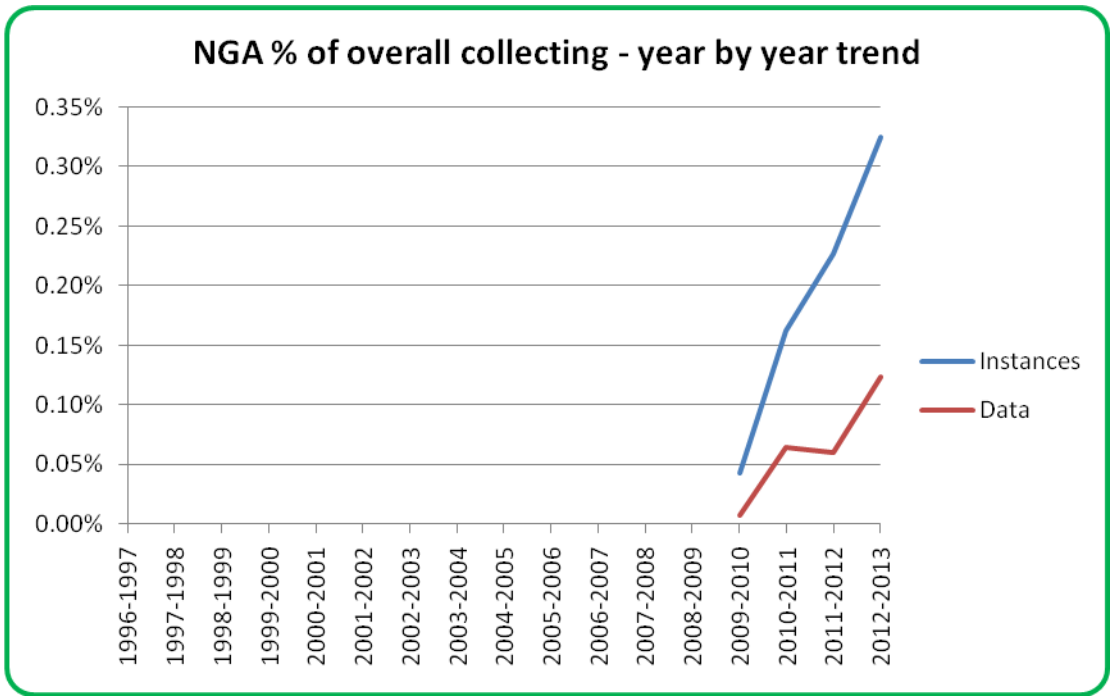
Australian Institute of Aboriginal and Torres Strait Islander Studies (2004+)





National Gallery of Australia (2009+)





3. *Development of the 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 which support the collection of Australian web resources.

3.1 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.

No major development on PANDAS was undertaken in 2012-2013. The Library has embarked on its Digital Library Infrastructure Replacement (DLIR) program which will involve development of new web collecting infrastructure over the coming two to three years. The PANDAS workflow system as it currently exists is unlikely to be further redeveloped.

3.2 Australian web domain harvest

In the first quarter of 2013 the Library conducted the eighth large scale harvest of the Australian web domain.

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 March and April 2013 and around 660 million unique documents were captured, amounting to 29.17 terabytes of data from nearly 1.7 million hosts. Following this harvest the combined total for all eight Australian domain harvests has now reached 5.3 billion files amounting to around 205 terabytes of data.

The following table shows the amount of content collected for each of the eight domain harvests conducted to date.

Domain Harvest	Unique files	Hosts	Size (TB)
2005	185 m	811,523	6.69
2006	596 m	1,046,038	19.04
2007	516 m	1,247,614	18.47
2008	1 billion	3,038,658	34.55
2009	756 m	1,074,645	24.28
2011	660 m	1,346,549	30.71
2012	1 billion	1,467,158	41.88
2013	660 m	1,690,232	29.17

In the absence of legal deposit provisions for online publications and websites at the Commonwealth level, the access that the National Library can provide to the whole domain harvest remains limited and they are not currently available to the general public. Unlike the selective Archive, we have not been able to negotiate prior permission individually with publishers to provide access to the collected content.

3.3 Collecting Commonwealth Government online publications

In May 2010 the Commonwealth Secretaries' ICT Governance Board (SIGB) endorsed whole-of-government arrangements proposed by the National Library to simplify the administrative procedures for obtaining permission to collect and preserve Commonwealth Government online publications. The arrangements allow the Library to collect publicly available Commonwealth Government online content without the need to seek prior individual permissions. The arrangements apply to Commonwealth agencies subject to the *Financial Management and Accountability (FMA) Act, 1997*. On the basis of this new arrangement, procedures were established for determining if selected government web content was covered by this general permission and for the recording of these permissions against government agencies in the PANDAS management system.

In March 2013 a third Government harvest was undertaken (the first being undertaken in 2011 and second in 2012). This harvest collected 7 million files and 736 gigabytes of data

In March 2012 a temporary specialist web archiving engineer position was established in the Web Archiving and Digital Preservation Branch with the appointment of Dr Mark Pearson. The initial focus of Dr Pearson's work has been to develop an infrastructure and interface to provide searchable public access to the Commonwealth Government web collections. It is expected that the 2011, 2012 and 2013 government collections will be made available to the public in late 2013.

4. Focus on users

4.1 User page views of the Archive

Owing to problems with the reporting of web statistics, meaningful figures have not been able to be produced for user page views over the past financial year. However, the web usage statistics for PANDORA are available from the Library's website at: http://stats.nla.gov.au/cgi-bin/report_index.cgi?report=PANDORA

4.2 Most viewed titles (websites) in the Archive

Around 6 % 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 10 sites accessed in 2012-2013.

Archived Title	Partner Responsible	Live site	Page views
Sydney Morning Herald	NLA	Yes	1,840,141
First families 2001	SLV	No	572,675
Digger history	AWM	No	379,496
Sydney Centre for Studies in Caodaism	NLA	Yes	378,831
Cultureandrecreation.gov.au	NLA	No	292,465
Life on the goldfields	SLV	No	199,446

Nova : science in the news	NLA	Yes	191,090
GamesInfo	NLA	No[1]	175,821
Centenary of Federation	NLA	No	156,119
Federal Minister for Fisheries, Forestry	NLA	No	147,977

1. *GamesInfo* has a live 'splash' web page which automatically re-directs to the PANDORA Archive

5. Preservation

Preservation activities particularly relevant to PANDORA during 2012-2013 include:

- The Web Archiving and Digital Preservation (WADiP) sections at the National Library have continued to work together to articulate preservation intent for various files in the PANDORA Archive based on the function, role and format class. Preservation intent statements have been completed for the selective web harvesting (PANDORA Archive), whole domain harvests and the Australian Government Web Archive collections. These statements are available online at: http://www.nla.gov.au/content/statements-of-preservation-intent

- Continued participation in the International Internet Preservation Consortium (IIPC) Preservation Working Group activities. See Section 6 below for more details.

6. International relations and representation

During 2012-2013 the National Library continued its participation in the International Internet Preservation Consortium (IIPC)[3] particularly in the work of the Preservation Working Group.

Paul Koerbin was invited to give a presentation at National Conference on eResources in Malaysia in Penang in December 2012 (see 7.2 for details).

7. Promoting the Archive

7.1 PANDORA Fact Sheet

The National Library has continued to update the PANDORA Fact Sheet and statistics page on a monthly basis and to distribute these to participants for publicity purposes. The fact sheet summarises key information about the Archive and supplements the printed PANDORA Brochure. The PANDORA Fact Sheet is made available online for the benefit of partners and other interested parties. See http://PANDORA.nla.gov.au/overview.html#factsheet

[3] Information about the IIPC is available from its web site at http://www.netpreserve.org/

7.2 Publications and public presentations

Presentations given and papers published by National Library Web Archiving staff during the 2012-2013 financial year include the following:

- *PANDORA past, present and future – national web archiving in Australia*. A presentation delivered at the Seminar Kebangsaan Sumber Electronik di Malaysia 2012 (National Conference on eResources in Malaysia) in Penang on 6 December 2012. Presentation and transcript of the talk is available online at: http://www.nla.gov.au/content/pandora-past-present-and-future-national-web-archiving-in-australia
- *'Oh, you wanted us to preserve that?!' Statements of preservation intent for the National Library of Australia's digital collection*. A paper by Colin Webb, David Pearson and Paul Koerbin published in D-Lib Magazine vol. 19, no. ½ (January/February 2013). http://www.dlib.org/dlib/january13/webb/01webb.html
- Presentation to the Australian Society of Archivists (Canberra Branch) at the National Archives of Australia on 16 October 2012 by Russell Latham.

7.3 'Australia's Web Archives – Curating Australia's Online Heritage' blog

The National Library launched a blog devoted to web archiving and the PANDORA archival collection in August 2012. The blog is open to contributions from all PANDORA partners. It includes posts highlighting interesting or topical content, archiving case studies as well as posts exploring content themes, issues and challenges. During the 2012-2013 financial year the following posts were published:

- *PANDORA Archive reveals earliest collection of Olympic Games websites* (August 2012, NLA)
- *Robert Hughes (1938-2012) – 'Our shared heritage' republic speech* (August 2012, NLA)
- *Looking for jeff.com.au* (August 2012, NLA)
- *Out of service? Cablog's last fare …* (September 2012, NLA)
- *Keating redesigned – and the length to which web archive curators (sometimes) go!* (September 2012, NLA)
- *Taking the web archive off the virtual shelf – presenting archived websites in a library exhibition* (October 2012, SLQ)
- *Bali bombing 10 years on* (October 2012, NLA)
- *Welcome back to the web jeff.com.au – reviving a 'lost'* website (November 2012, NLA)
- *Archiving the history of history on-line* (December 2012, NLA)
- *Archiving the protest site* (January 2013, NLA)
- *What is the oldest website? And will an artefact do?* (May 2013, NLA)
- *How fast is the web archiving?* (May 2013, NLA)
- *Queensland regional festivals – how we are celebrating* (May 2013, SLQ)

7.4 Presentations to visitors to the National Library

The National Library regularly hosts visitors from other libraries and organisations. Presentations on PANDORA, web archiving and PANDAS were provided to visitors to the Library from Malaysia and Japan.

8. Concluding summary

Some of the highlights of 2012-2013 include:

- A new weekly newsletter to partners delivered through email discussion list (section 1.1)
- Continuing steady growth of the Archive content (section 2).
- Completion of the 2013 large scale harvest of the Australian web domain (section 3.2).
- Completion of the 2013 harvest of Commonwealth Government web content under whole-of-government permission arrangements (section 3.3).
- Launch and ongoing posting to the new 'Australia's web archives' blog (section 7.3)