Annual report to partners 2010-2011

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 web sites which is built collaboratively by the National Library, 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 2010-2011 financial year.

1.1 PANDORA partner news and training

In May and June 2011 Russell Latham, the web archiving operational supervisor, visited the state libraries of Victoria, Queensland and New South Wales and the Northern Territory Library. The visit was an initiative to provide partners with some practical operational and training updates for curator staff; and to update staff and managers on proposed future directions for web archiving at the National Library in the context of the Library's Digital Library Replacement Project (DLIR).

1.2 Consultation mechanisms

The National Library continued to inform other PANDORA participants about the operation of PANDORA through the two email discussion lists: *pandoraconsult-l* and *pandora-l*.

1.3 Reports

Each month, a report on the growth of the Archive, usage statistics, and a summary of responses to the online PANDORA user survey forms is sent to both email discussion lists. 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 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 annual report of progress and activities to the Chief Executive Officers of partner agencies is also provided. These reports are also 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 It includes the gathering of a monograph that has been archived once only, the first gathering of a serial title or integrating title (for example, a web site that changes over time), and all subsequent gatherings.

1.4 National & State Libraries Australasia (NSLA) project

In November 2010 the Library proposed to a meeting of NSLA representatives that there was an emerging imperative to review the current approach and methods of collaborative web archiving. This is driven by the increasing size of the web and its changing nature to a more dynamic and socially interactive publishing medium; and by increasing stresses on aging infrastructure and limited resources.

The outcome of this initial proposal by the Library was a NSLA workshop in Sydney in February 2011, at which the Library presented possible future models for web archiving for discussion. This workshop delivered recommendations for a proposed model that extended the collaborative curation model for web archiving and included more thematic harvesting; and for the establishment of a NSLA group to manage the collaborative opportunities of the changes to the way web archiving may be undertaken. These recommendations were endorsed at the March 2011 NSLA meeting.

Since the proposed model retains a centralised infrastructure established and maintained at the National Library, the move toward a new model for collaborative web archiving is dependent upon the direction and outcomes of the Library's Digital Library Infrastructure Replacement (DLIR) Project. The proposed NLSA group will work within an endorsed project framework to maintain collaborative input to the implementation of the new web archiving model and methods.

2. *Growth of the Archive*

2.1 Size and annual growth of the Archive

The Archive continued to show steady growth in 2010-2011, although the percentage growth rate for Titles and Instances and Usage (page views) was less than for the previous financial year. The growth rate for the data size (terabytes) was steady with the previous year at around 30%.

	30 June 2010	30 June 2011	Growth 2009-10
Titles	25,549	28,298	2,749 (10.8 %)
Instances	55,919	65,923	10,004 (17.9 %)
Terabytes[2]	4.01	5.24	1.23 (30 %)
Usage (page views)	4,985,676	5,919,337	933,661 (18.7%)

Government publications remain a substantial component of the collecting focus and comprise approximately 55 % of the titles in the Archive.

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

2.2 Select analysis of archival content

This year's analysis of the contribution to the growth of the Archive is focused on the four principal measures:

1. Number of titles contributed;
2. Number of archived instances contributed;
3. Number of files contributed; and,
4. Data size of the contribution.

These measures are analysed to look at the contributions of each partner agency and, in order to reveal trends, report over the 2008-09, 2009-10 and 2010-11 financial years. The National Gallery of Australia has not been included in this analysis because they joined the PANDORA collaboration half-way through the period analysed in early 2010.

The statistical data is presented to show, for each of the four measures, the actual amount of each partner's contribution (2.2.1), the relative percentage of each partner's contribution to the Archive (2.2.2) and the variation of each partner's contribution over the past three financial years in percentage values (2.2.3).

2.2.1 Actual contribution of titles, instances, files and data

Figure 1: Number of *Titles* contributed to the Archive by PANDORA Partners over the past three financial years

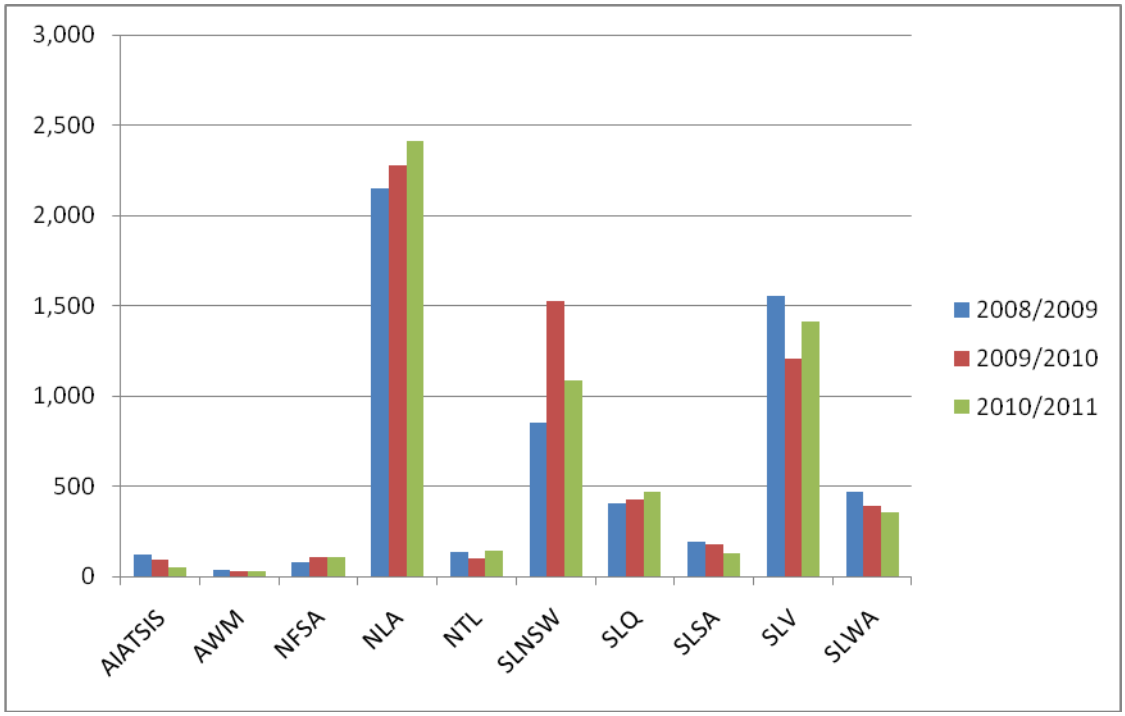


Figure 2: Number of *Instances* contributed to the Archive by PANDORA Partners over the past three financial years

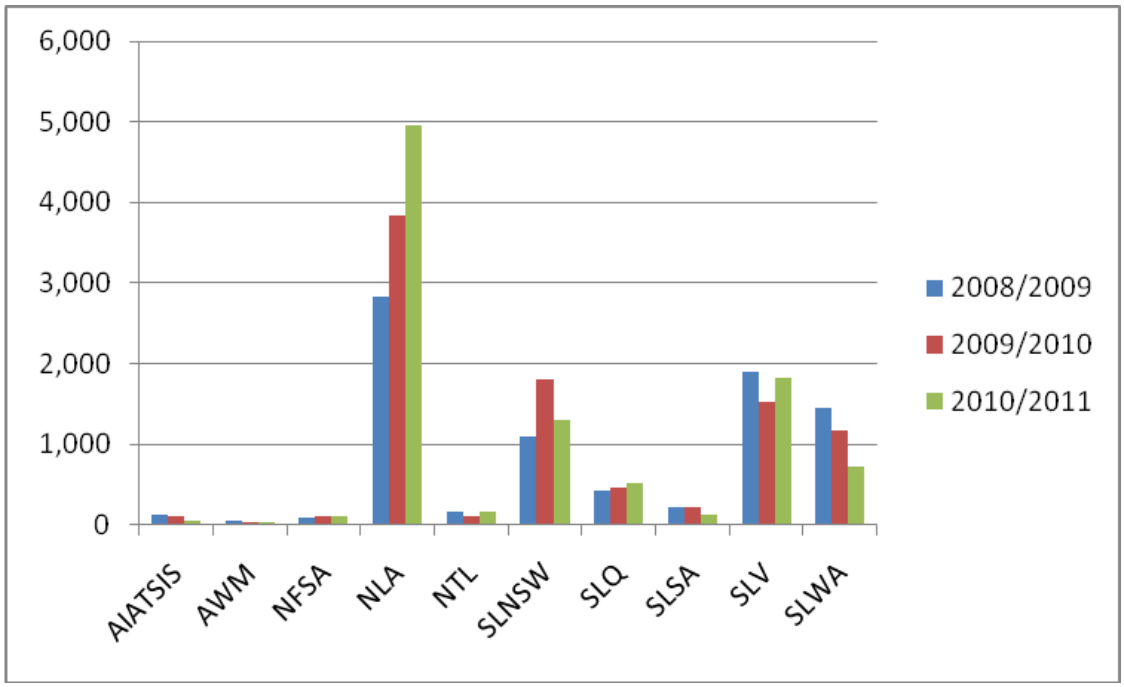


Figure 3: Number of *Files* contributed to the Archive by PANDORA Partners over the past three financial years

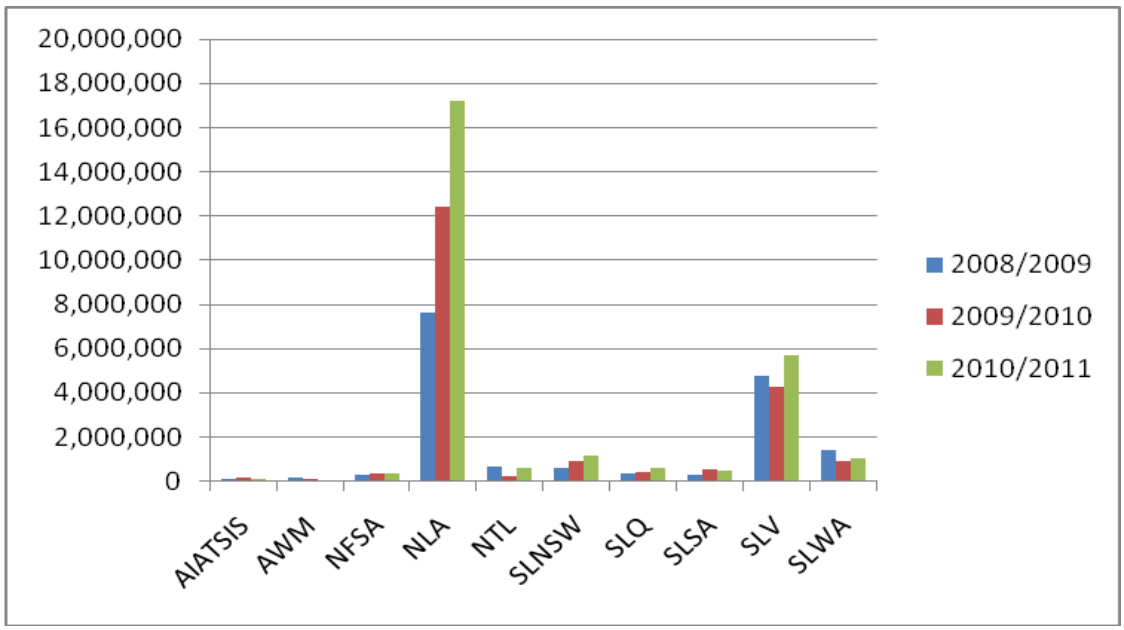
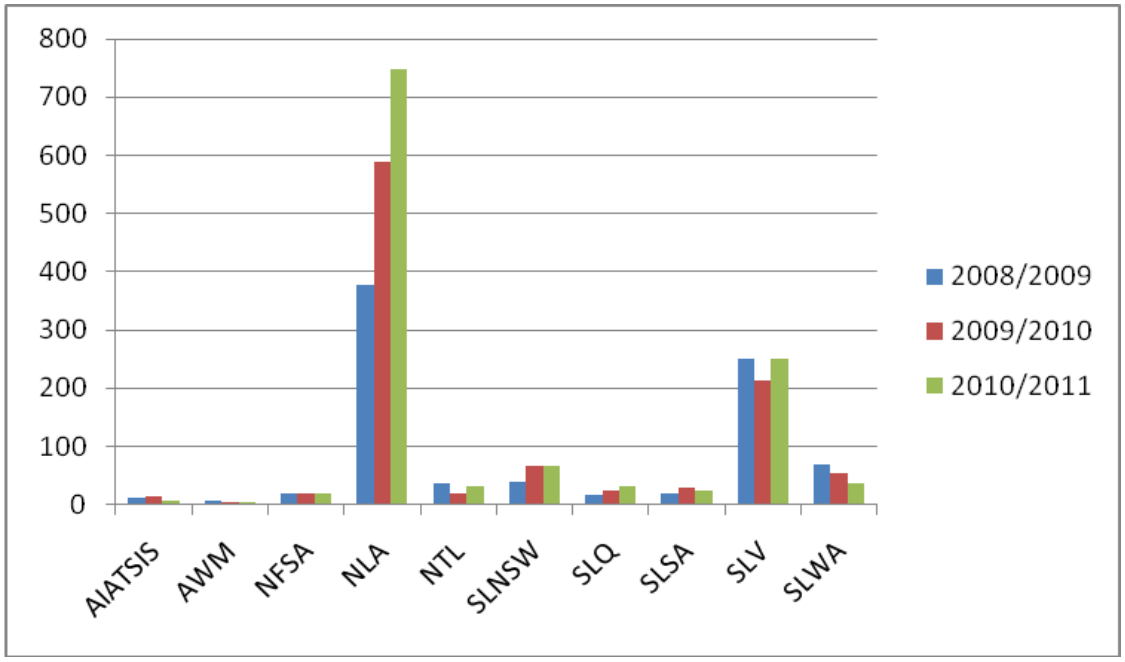


Figure 4: Amount of *gigabytes* contributed to the Archive by PANDORA Partners over the past three financial years



2.2.2 Percentage contribution of titles, instances, files and data

Figure 5: Percentage of *Titles* contributed by PANDORA Partners over the past three financial years

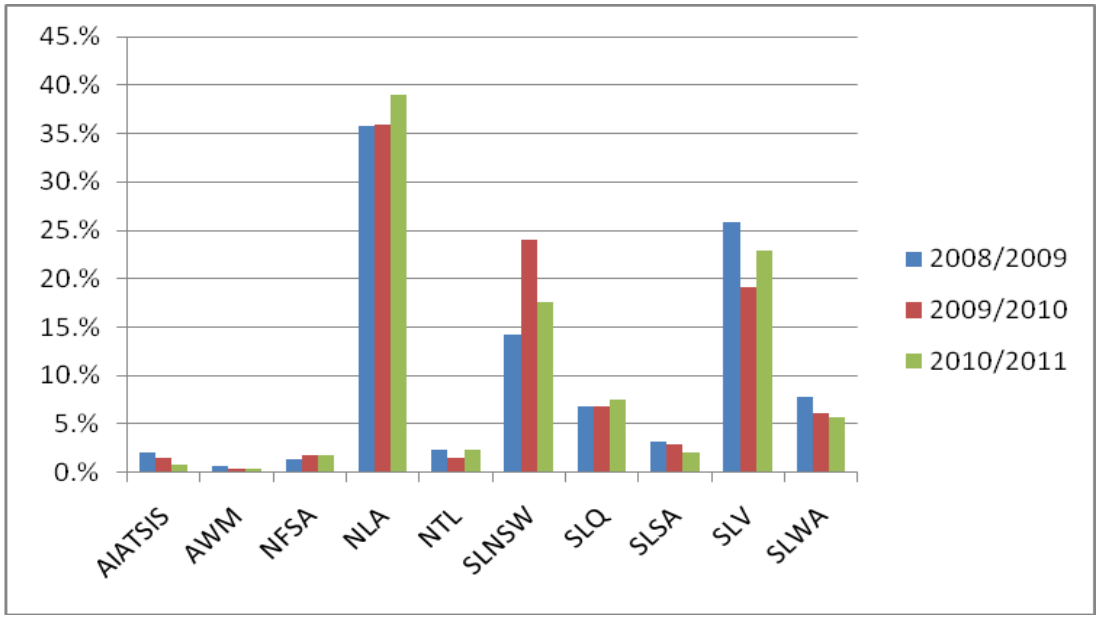


Figure 6: Percentage of *Instances* contributed by PANDORA Partners over the past three financial years

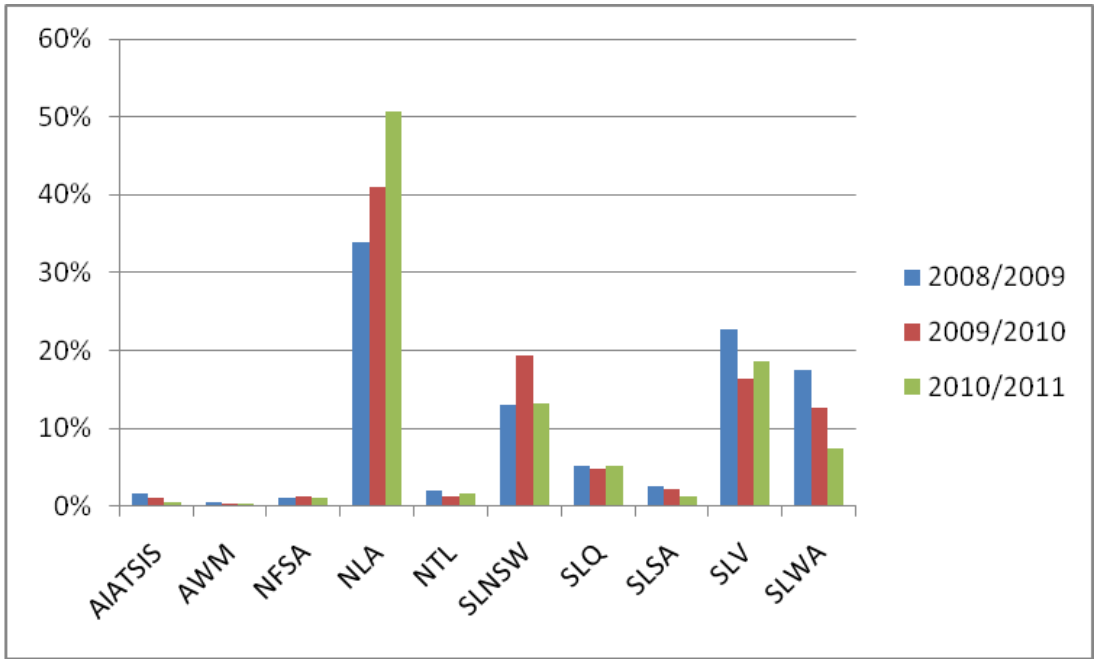


Figure 7: Percentage of *files* contributed by PANDORA Partners over the past three financial years

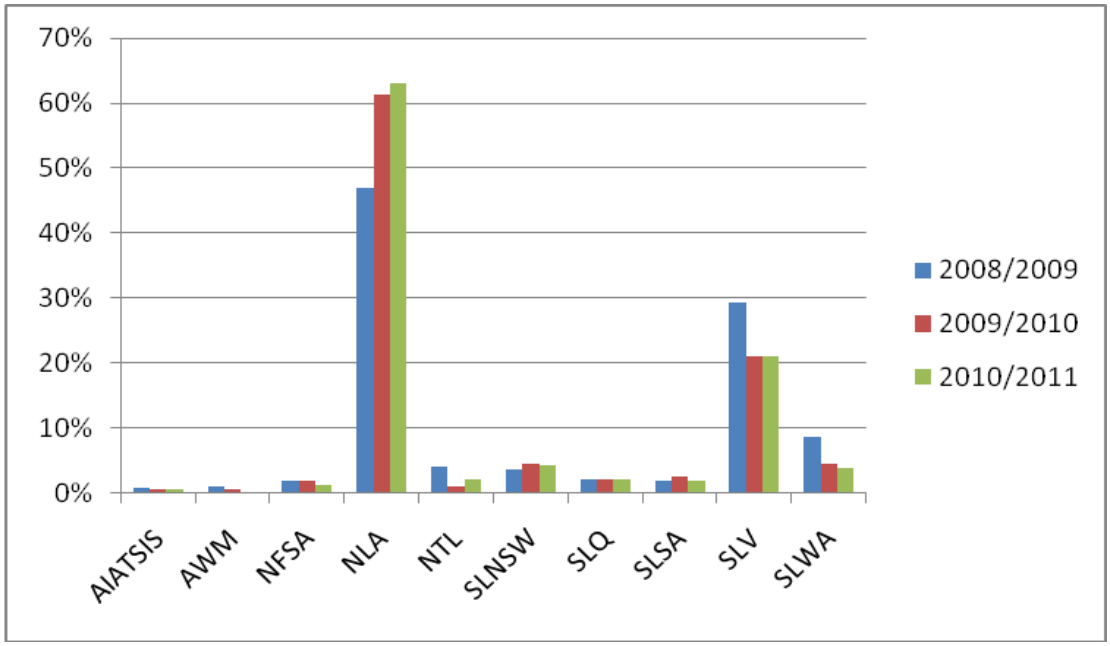
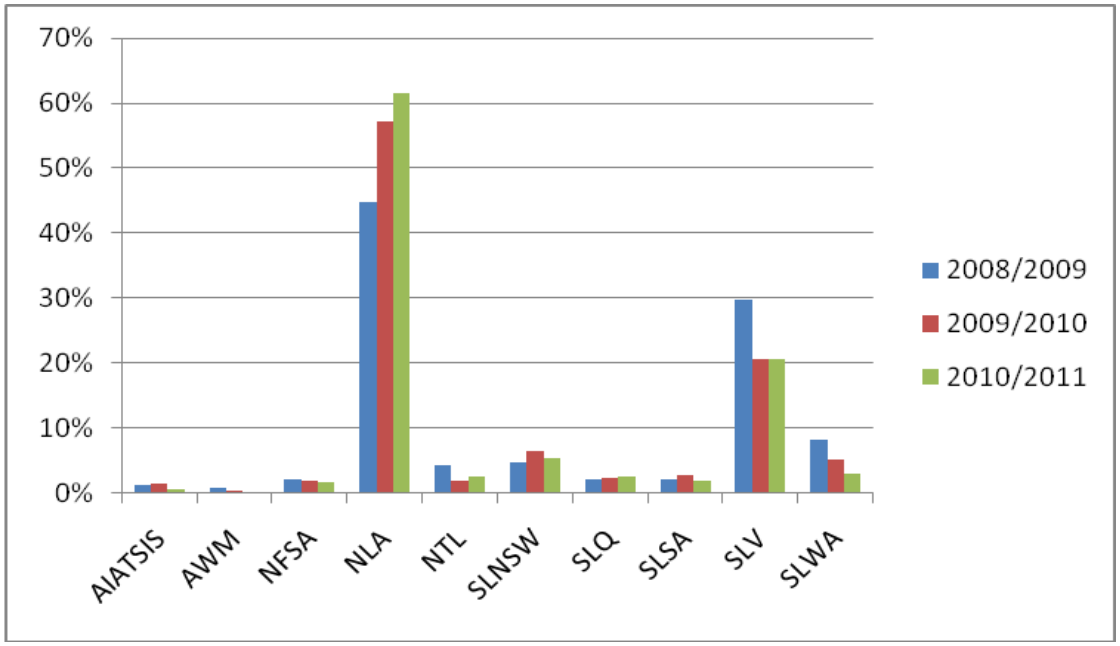


Figure 8: Percentage of *gigabytes of data* contributed by PANDORA Partners over the past three financial years



2.2.3 Percentage variation in contribution over previous financial years

Figure 9: Percentage variation from previous financial year in the contribution of *Titles* to the Archive by PANDORA Partner for the financial years 2009-2010 and 2010-2011

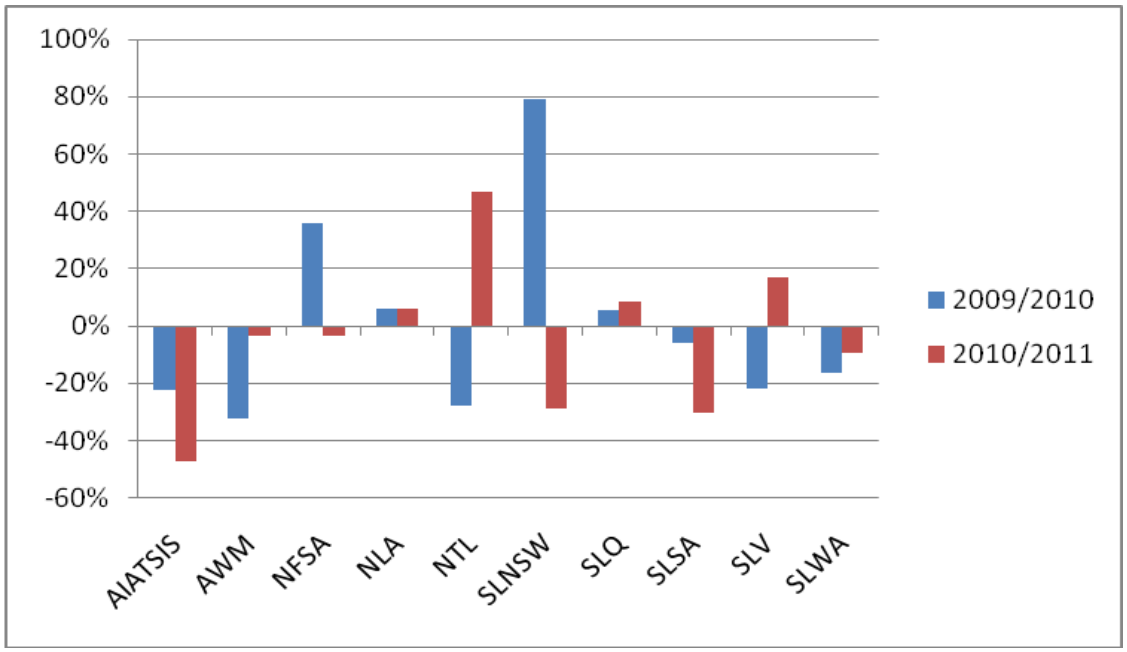


Figure 10: Percentage variation from previous financial year in the contribution of *Instances* to the Archive by PANDORA Partner for the financial years 2009-2010 and 2010-2011

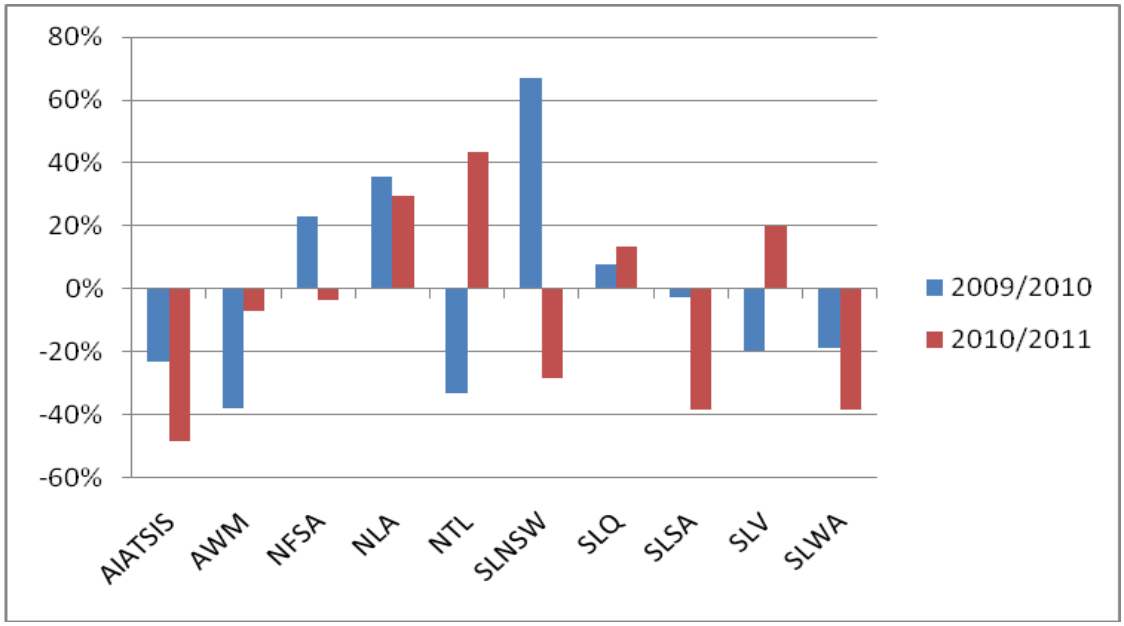


Figure 11: Percentage variation from previous financial year in the contribution of *files* to the Archive by PANDORA Partner for the financial years 2009-2010 and 2010-2011

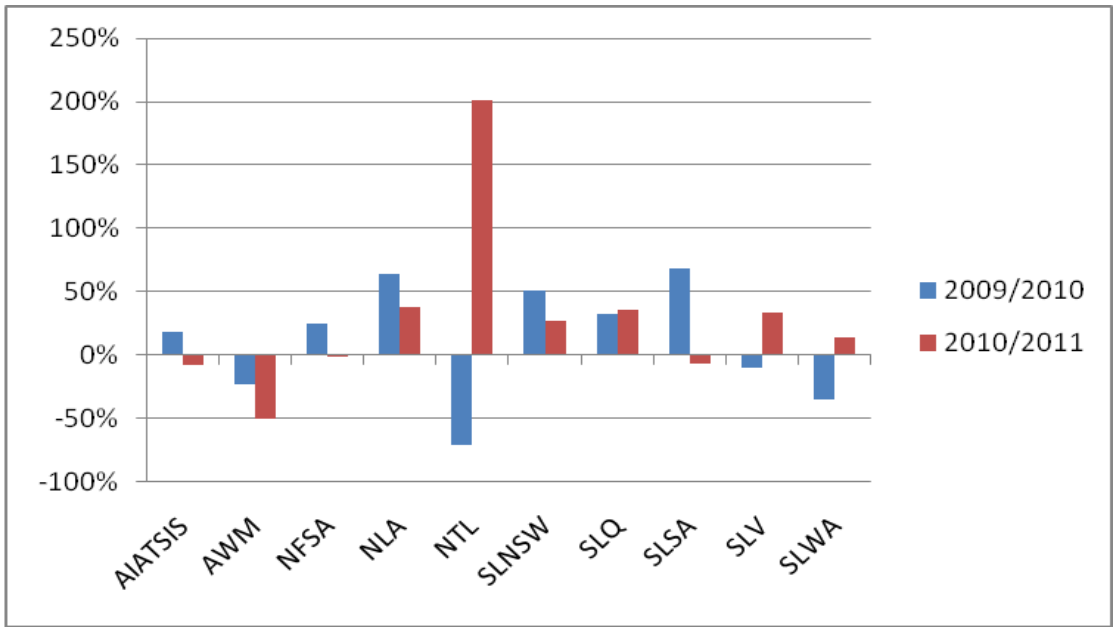
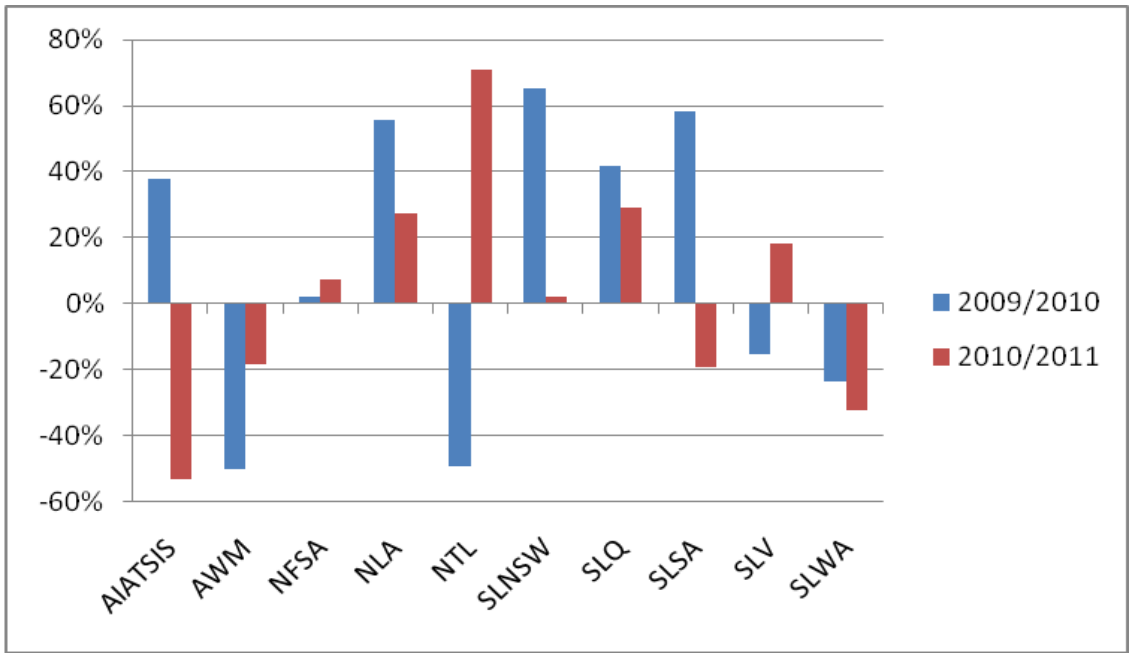


Figure 12: Percentage variation from previous financial year in the contribution of *data* (gigabytes) to the Archive by PANDORA Partner for the financial years 2009-2010 and 2010-2011



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 web sites to PANDORA. This does not include cataloguing, which is carried out in separate local systems.

No major development on PANDAS was undertaken in 2010-2011.

3.2 Australian web domain harvest

In the first quarter of 2011 the Library conducted the sixth 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 February and March 2011 and around 660 million unique documents were captured, amounting to 30.71 terabytes of data. Following this harvest the combined total for all six Australian domain harvests has now reached 3.7 billion files amounting to around 134 terabytes of data.

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

Domain Harvest	2005	2006	2007	2008	2009	2011
Unique files	185 m	596 m	516 m	1 billion	765 m	660 m
Hosts	811,523	1,046,038	1,247,614	3,038,658	1,074,645	1,346,549
Size (Tb)	6.69	19.04	18.47	34.55	24.28	30.71

In the absence of legal deposit provisions for online publications and web sites at the Commonwealth level, the access that the 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 Whole-of-Government arrangements for Commonwealth 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 Accountability and Management (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 addition, in March 2011 a harvest of around 800 government URL domains was completed in conjunction with the 2011 domain harvest (see section 3.2). This harvest collected 7.4 million files amounting to 538 gigabytes of data. It is expected that this collection will be made available some time in 2012.

4. *Focus on users*

As for previous annual reports, an analysis of usage of the Archive over the last three financial years was undertaken.

4.1 User page views of the Archive

The analysis showed a steady increase of 18.7% in usage (based on page views) during the 2010-2011 financial year over the previous year and an increase of nearly 1 million page views over the previous financial year.

Usage in 2010 - 2011

Total page views	Average per month	Month of highest use	Month of lowest use
5,919,337	493,278	August 2010 531,316	June 2011 382,377

Usage in 2009 - 2010

Total page views	Average per month	Month of highest use	Month of lowest use
4,985,676	415,473	July 2009 729,131	August 2009 285,932

Usage in 2008 - 2009

Total page views	Average per month	Month of highest use	Month of lowest use
3,861,089	321,757	September 2008 516,286	December 2008 249,755

Detailed 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 2010-2011.

Archived Title	Partner Responsible	Live site
First Families 2001	SLV	No
Sydney Centre for Studies in Caodaism	NLA	Yes
GamesInfo	NLA	No[1]
Life on the goldfields	SLV	No
ARIA report	SLNSW	Yes
Prime Minister of Australia, John Howard	NLA	No
Online currents	NLA	No
Masthead : arts, culture and politics	NLA	Yes
Centenary of Federation	NLA	No
Antipodean SF	NLA	Yes[2]

1. *GamesInfo* has a live 'splash' web page which automatically re-directs to the PANDORA Archive
2. *Antipodean SF* only retains the current monthly issue on the live site so the PANDORA Archive provides the only access to archival issues for the publication.

5. *Preservation*

Preservation activities particularly relevant to PANDORA during 2010-2011 include:

- Monitoring the range of file formats entering the PANDORA Archive. This is being achieved through ongoing technical processes such as the characterisation and profiling of the results of format identification tools which specify format types and versions. This process also includes the testing of these tools against benchmark results and over unidentified formats in the Archive. Results of this work are being used as requirements for the new workflow in the NLA DLIR project.

- The Digital Preservation and Web Archiving sections working together to articulate preservation intent for various files in the PANDORA Archive based on the function, role and format class. The preservation intent has been expressed as both a statement of intent and as an understanding of the limitations of how the content is harvested for the Archive.

- Experimental and conceptual work describing how a web object can be intellectually expressed as a complex object. This work helps the NLA understand what process and preservation actions will be required to maintain access to a web page and specific components of a web page over time.

- Continued participation in the IIPC Preservation Working Group activities.

6. International relations and representation

During 2009-2010 the National Library continued its active participation in the International Internet Preservation Consortium (IIPC)[3] particularly in the work of the Preservation Working Group.

Paul Koerbin (Manager, Web Archiving) participated in the IIPC General Assembly in The Hague in May 2011, re-establishing the Library's interest in the Access Working Group.

7. Promoting the Archive

7.1 PANDORA Fact Sheet

The 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

7.2 Publications and public presentations

There were no publications or public presentations of note during the 2010-2011 financial year.

7.3 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 the following organisations:

- National Library of Indonesia (October 2010)
- Northern Territory Library (October 2010)
- Swedish National Library (March 2011)

8. Concluding summary

Some of the highlights of 2010-2011 include:

- Continuing steady growth of the Archive content (section 2).
- Completion of the sixth large scale harvest of the Australian web domain (section 3.2).
- Completion of the first harvest of Commonwealth Government web content under whole-of-government permission arrangements (section 3.3).
- Delivery of update training and support to a number of partner libraries (section 1.1).

3 Information about the IIPC is available from its web site at http://netpreserve.org/about/index.php