Annual report to partners 2009-2010

Contents

1. Participants working together

1.1 PANDORA partner news and training
1.2 Consultation mechanisms
1.2 Reports

2. Growth of the Archive

2.1 Size and annual growth of the Archive
2.2 Select analysis of archival content

3. Development of the Archive

3.1 Development of PANDAS
3.2 Australian web domain harvest
3.3 Whole-of-Government arrangements for Commonwealth publications

4. Focus on users

4.1 User page views of the Archive
4.2 Most viewed titles (websites) in the Archive

5. Preservation

6. International relations

7. Promoting the Archive

7.1 PANDORA Fact Sheet
7.2 Publications and presentations
7.3 Presentations to visitors to the National Library

8. Concluding summary

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 2009-2010 financial year.

1.1 PANDORA partner news and training

In early 2010 the National Gallery of Australia became the latest organisation to participate in the collaborative collecting and archiving of Australian web materials through the PANDORA Archive. Staff from the Gallery received two days initial training in May 2010.

Joanne Hocking from the State Library of South Australia visited the National Library for two days of refresher and advanced training in April 2010.

In 2010 the National Library began using the *Adobe ConnectNow* online web conferencing service to provide training support to operational staff in partner agencies. The online service is free, simple and quick to access and allows National Library staff to demonstrate and describe procedures live through a web browser in both planned and *ad hoc* contexts.

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 web

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.

site 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.

During the 2009-2010 financial year period there were some difficulties in providing these above reports each month.

An 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.

2. Growth of the Archive

2.1 Size and annual growth of the Archive

The Archive continued to show steady growth in 2009-2010, although the percentage growth rate for Titles, Instances and Terabytes was slightly less than for the previous financial year.

	30 June 2009	30 June 2010	Growth 2009-10
Titles	22,464	25,549	3,085 (13.7 %)
Instances	46,591	55,919	9,328 (20 %)
Terabytes[2]	3.02	4.01	0.99 (32.8 %)
Usage (page views)	3,861,089	4,985,676	1,124,587 (29.1%)

In the previous financial year (2008-2009) there was a decrease in the number of user page views reported due to changes in the reporting methodology. Reporting stabilised in 2009-2010 and showed a strong increase in usage with page views up by just over 29%.

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

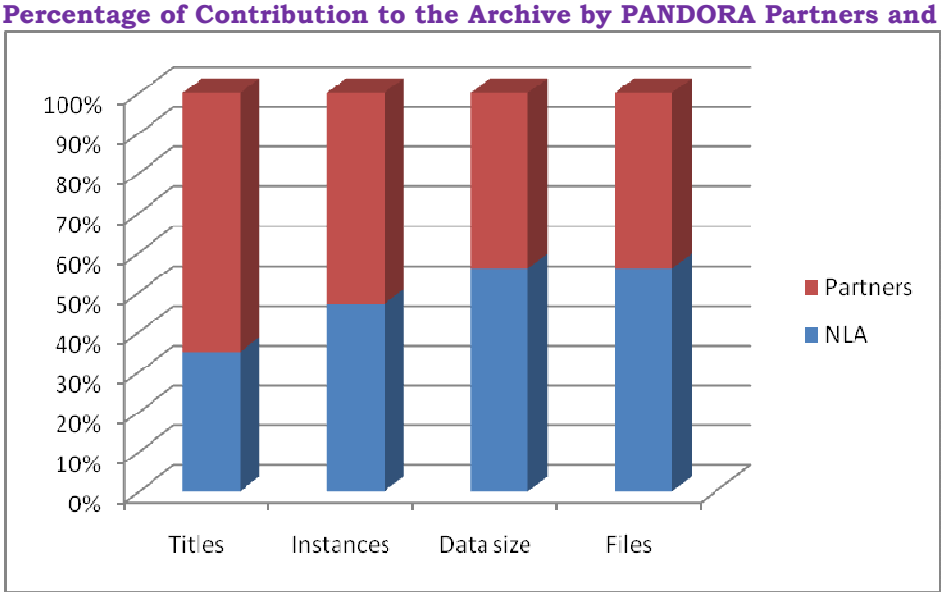
2.2 Select analysis of archival content

Balancing the selection and collection of new titles with the demands of processing scheduled re-harvested instances is one of the challenges all partner agencies face.

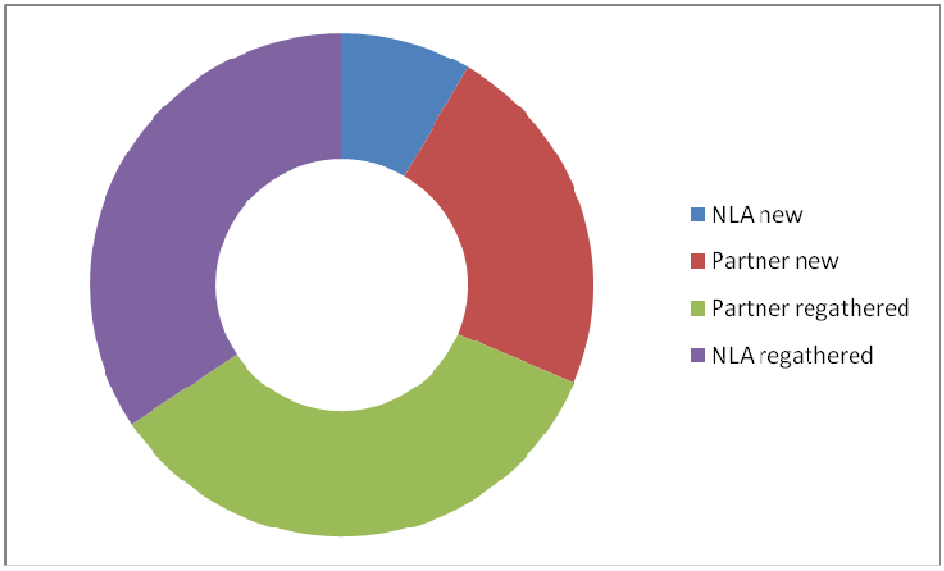
The following two charts show the relative contribution to the Archive by partner agencies (combined) and the National Library and ratios of new title collecting and scheduled re-gathering of titles.

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

Percentage of Contribution to the Archive by PANDORA Partners and the NLA

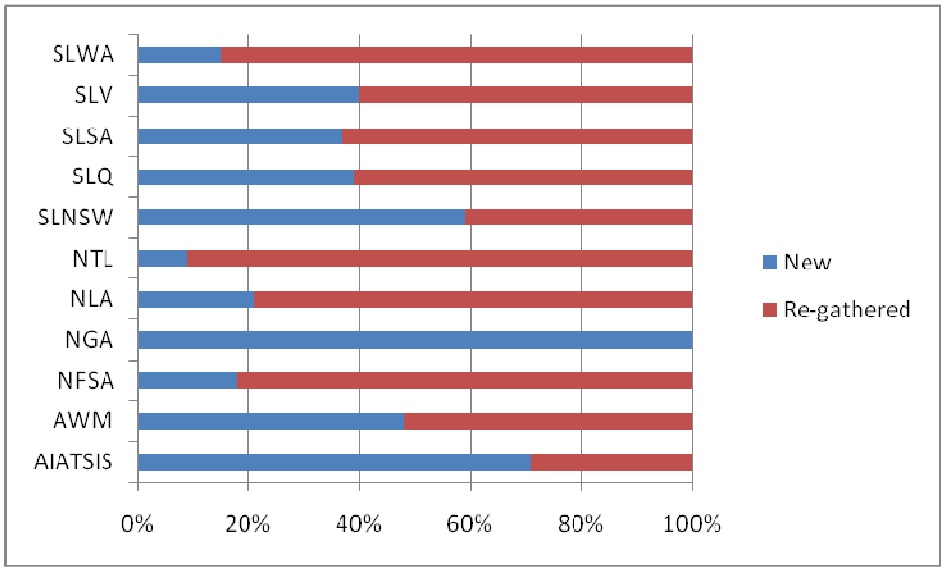


Ratio of New Titles and Re-gathered Instances Contributed by PANDORA Partners and the NLA in 2009-2010



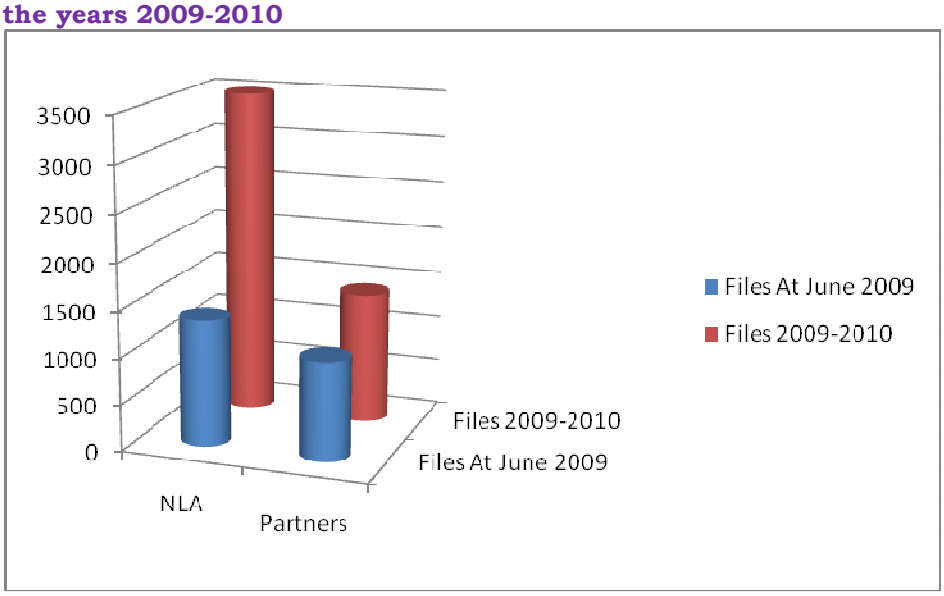
The next chart show the ratio of new titles and re-harvested (re-scheduled) instances archived for each partner agency during 2009-2010.

Ratio (Percentage) of New Titles and Re-gathered Instances Contributed to the Archive for Each PANDORA Partner in 2009-2010

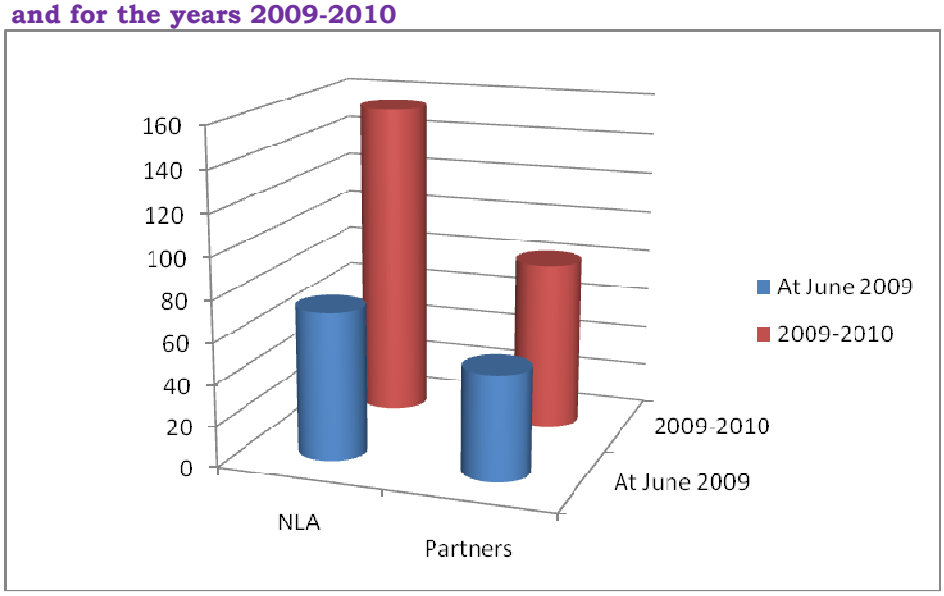


Anecdotal evidence and observation over the past year suggests that the web content being collected is more complex and the websites archived are generally larger both in the number of files and in the amount of data. This is substantiated by some basic analysis of the average instance size measured in number of files and data size. The following three charts show a comparison between the average instance size of all content collected for the Archive from its commencement up until June 2009 and for content collected in the 2009-2010 financial year. The analysis shows a significant increase in the actual average instance size and the percentage growth in size.

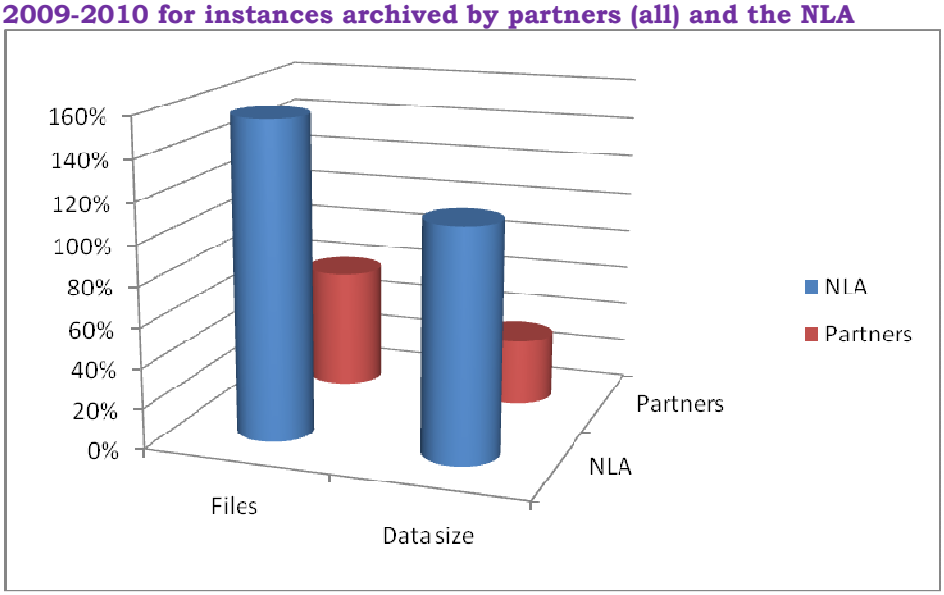
Comparison of the average number of files per instance at June 2009 and for the years 2009-2010



Comparison of the average size for an instance in megabytes as at June 2009 and for the years 2009-2010



Percentage increase in the data size and number of files per instance in 2009-2010 for instances archived by partners (all) and the NLA



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 (the 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 2009-2010. The major enhancement to the PANDORA service was the incorporation of the full-text PANDORA index into the Library's one search discovery service Trove in February 2010. This move has improved the indexing of the PANDORA Archive and allows archived web content to be discovered alongside other resources.

3.2 Australian web domain harvest

In the last quarter of 2009 the Library conducted the fifth 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 on our behalf. The Internet Archive has extensive experience in this form of web archiving. This fifth harvest had the objective of harvesting (at least) 650 million unique URLs from the .au web domain and other resources on hosts located in Australia (where these could be automatically identified as such.

The harvest was run during September and October 2009 and around 765 million unique documents were captured, amounting to 24.29 terabytes of data. Following this harvest the combined total for all five Australian domain harvests has now reached three (3) billion files amounting to around 103 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
Unique files	185 million	596 million	516 million	1 billion	765 million
Hosts	811,523	1,046,038	1,247,614	3,038,658	1,074,645
Size (Tb)	6.69	19.04	18.47	34.55	24.28

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.

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 22.5% in usage (based on page views) during the 2009-2010 financial year over the previous year. The apparent significant drop in usage between the 2007-2008 and 2008-2009 financial years was due to a change in reporting mechanisms in 2008 and does not reflect actual usage.

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

Usage in 2007 – 2008

Total page views	Average per month	Month of highest use	Month of lowest use
7,295,996	607,999	January 2008 1,084,499	July 2007 303,855

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 5 % 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 large 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 2009-2010.

Archived Title	Partner Responsible	Live site
First Families 2001	SLV	No
APEC Australia 2007	NLA	No
Sydney Centre for Studies in Caodaism	NLA	Yes
GamesInfo	NLA	No[1]
SOFWeb	SLV	No
Centenary of Federation	NLA	No
Prime Minister of Australia, John Howard	NLA	No
ARIA report	SLNSW	Yes
Antipodean SF	NLA	Yes[2]
ATSIC	NLA	No

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

The National Library continued to monitor the range of file formats entering the PANDORA archive, maintaining an ongoing profile of its technical makeup.

Preservation activities particularly relevant to PANDORA during 2009-2010 include involvement in three work packages for the IIPC Preservation Working Group (see Section 6):

1. A proposal to set up an annual technical environment scan to identify and record the current software (browsers and plug-ins) used to access web content (which may be useful information for later emulation programs).

2. Testing of popularly discussed preservation action strategies (migration and emulation) to consider their current and potential usefulness for maintaining access to archived web content.

3. Testing the time required to run format identification tools over web archive collections and analysis of the results.

6. International relations and representation

During 2009-2010 the National Library continued its active participation in the International Internet Preservation Consortium (IIPC)[3] being involved in three preservation related work packages (see section 5).

Colin Webb (Director, Web Archiving and Digital Preservation) participated in the IIPC General Assembly in Singapore in May 2010 presenting reports to the IIPC Preservation Working Group.

David Pearson (Manager, Digital Preservation) presented two papers by Digital Preservation staff at the iPRES2009 conference in San Francisco in October 2009.

7. Promoting the Archive

7.1 PANDORA Fact Sheet

The Library has continued to update the PANDORA Fact Sheet on a monthly basis and to distribute it to participants for publicity purposes. It 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

A small number of publications and presentations were completed during the year for the dual purpose of promoting our work and sharing what we have learned. These include:

- Koerbin, P. 'Issues in Business Planning for Archival Collections of Web Materials', in *Business Planning for Digital Libraries : International Approaches* edited by Mel Collier, Leuven University Press, 2010. (Book chapter).
- Koerbin, P. *PANDORA Update*. An update on PANDORA activities published in the National Library's online publication *Gateways* in December 2009. (Brief article).
- *Archiving the Music Web*. A presentation by Paul Koerbin for the Music Council of Australia's Annual Assembly, Melbourne, September 2009. (Presentation).
- 'Caught in the Web'. A highlight section on the PANDORA Archive published in the Melbourne Age (19 May 2010) as part of the cover story 'You must remember this …' by journalist Andrew Stephens based on an interview conducted in the Library with Paul Koerbin. (Newspaper feature).

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

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:

- Australian Research Council (July 2009)
- National Archives of Malaysia (November 2009)
- National Library of Indonesia (November 2009)
- Australia Council for the Arts (December 2009)
- National Library of New Zealand (February 2010)
- National Library of the Czech Republic (April 2010)
- Staff from the Library's Jakarta office (May 2010)

8. Concluding summary

Some of the highlights of 2009-2010 include:

- Continuing steady growth in the Archive (section 2).
- PANDORA search index incorporated into the National Library's one search discovery service Trove (section 3).
- Completion of the fifth large scale harvest of the Australian web domain (section 3).
- Endorsement by the Secretaries' ICT Governance Board for whole-of-government arrangements for permission to collect Commonwealth Government web publications (section 3).
- Maintaining an active role in international forums including the IIPC (section 6).
- Maintaining a commitment to finding effective ways to provide training and support of operational staff in partner agencies (section 1).
- Continuing to promote and provide information about PANDORA and web archiving activities (section 7).