Annual report to partners 2011-2012

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 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 2011-2012 financial year.

1.1 PANDORA partner news and training

1.1.1 Training support to partners

In July 2011 Mr Russell Latham, the web archiving operational supervisor at the National Library, visited the State Library of South Australia providing two training sessions and a talk to SLSA staff. In November 2011 Mr Latham visited the State Library of Western Australia to provide 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 Infrastructure Replacement project (DLIR). This visit completed a round of support visits to partner agencies located outside the ACT partner libraries completed in the previous financial year.

1.1.2 PANDORA Partners meeting

On the 29th March 2012 the National Library hosted the first PANDORA partners meeting. The meeting was attended by one or more representatives of all PANDORA partners. Representatives discussed curatorial and staffing issues such as training, standards, communication and promotion. Two specific outcomes were to establish a web archiving blog to help with promotion of PANDORA activities and to build two collaborative collections, one on "The Murray Darling and its Basin" and the other on "Australia's 21st century resources boom".

A separate session was held that covered issues with the current infrastructure and developments towards a new digital library infrastructure at the National Library and how this might impact upon web archiving operations.

A talk was also given by Mr Gordon Mohr, former lead developer at the Internet Archive, on the proposed replacement gathering tool, Heretrix.

1.2 Consultation mechanisms

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

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 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 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.

1.4 National & State Libraries Australasia (NSLA)

1.4.1 National & State Libraries Australasia (NSLA) web archiving project

In November 2011 the Manager of Web Archiving at the National Library, Dr Paul Koerbin, attended the NSLA Heritage Collections Forum at the State Library of New South Wales and gave a presentation titled 'Next steps towards a new model for collaboratively curating a national web archive'. The presentation reiterated the proposed model that was endorsed at the March 2011 NSLA meeting (some detail about this was provided in last year's Annual Report to Partners); and provided an update on issues relevant to the progress towards this model including the National Library's Digital Library Infrastructure Replacement (DLIR) project, domain harvesting activity, preservation work and progress on the extension of Legal Deposit to cover digital materials. The presentation also proposed matters for discussion and next steps including establishing a working group on collaborative curation, to consider how we build on existing and established expertise and collecting practice and to propose trial collaborative collecting themes.

1.4.2 Collaborative collecting project

At the March 2012 PANDORA partners meeting it was decided that PANDORA partners should all contribute to two collaborative collections. The two collections were, the "Murray Darling River and its Basin" and "Australia's early 21st century resources boom". The collections were scoped and developed with contributions from partner agencies through the PANDORA wiki. Each agency then contributed either items already archived or searched for new titles that could be included. The project gave all partners the opportunity to participate in a more collaborative environment that more closely resembles the new model we expect to evolve as we migrate from our current model due to the new National Library digital infrastructure. Development of the collections will continue as titles appear.

Australia's early 21st century resources boom

http://nla.gov.au/nla.arc-c11666

Murray Darling River and its Basin

http://nla.gov.au/nla.arc-c11662

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 continued to show steady growth in 2011-2012, with the percentage growth rate for Titles and Instances and data size steady from the previous financial year. The growth rate for the usage (user page views) was up by 5 percentage points from the previous financial year.

	30 June 2011	30 June 2012	Growth 2011-12
Titles	28,298	31,421	3,123 (11 %)
Instances	65,923	76,439	10,516 (16 %)
Terabytes[2]	5.24	6.79	1.55 (29 %)
Usage (page views)	5,492,693	6,799,198	1,306,505 (24%)

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's analysis is focussed on the type of content being added to the archive.

Government material makes up a large component of the material added to the PANDORA Web Archive and is often described as a priority collecting area for many partner agencies. A number of State libraries now have a digital legal deposit provision or Premiers circular that grants them permission to collect government material.

In the 2011-2012 financial year 2,997 new titles were registered and archived by all PANDORA partners. Of these new titles 1,690 were gathered from the Australian government domain. Therefore approximately 57% of all new titles are coming from the gov.au domains for the financial year. This corresponds to the overall percentage of government material held in the collection as described in 2.1.

A breakdown for each state is available below.

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

Figure1: Number of total new titles archived compared with total new government titles archived for year 2011-2012.

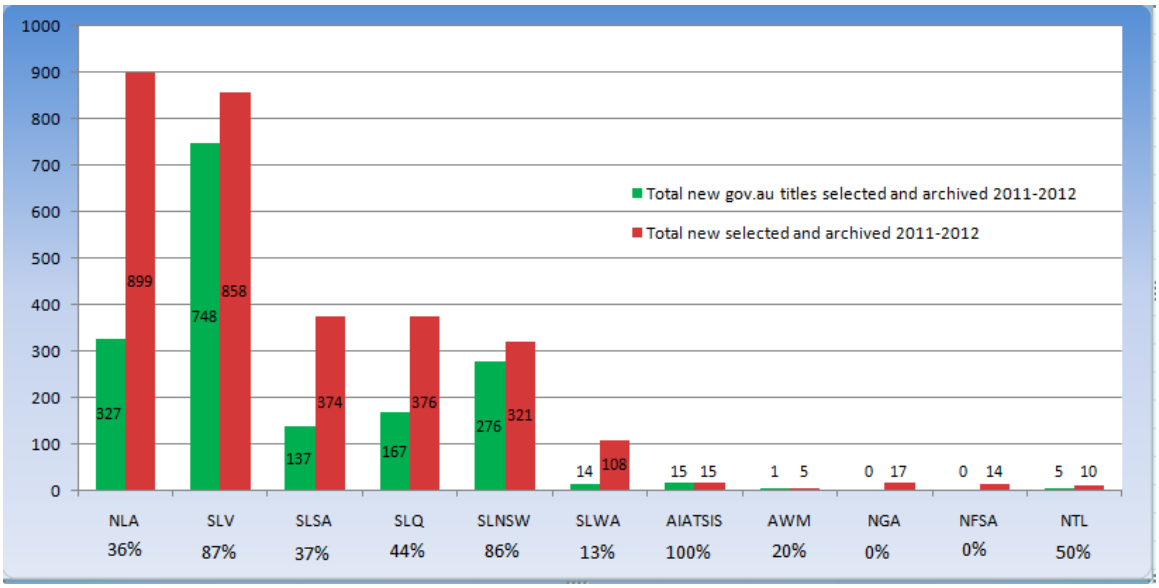
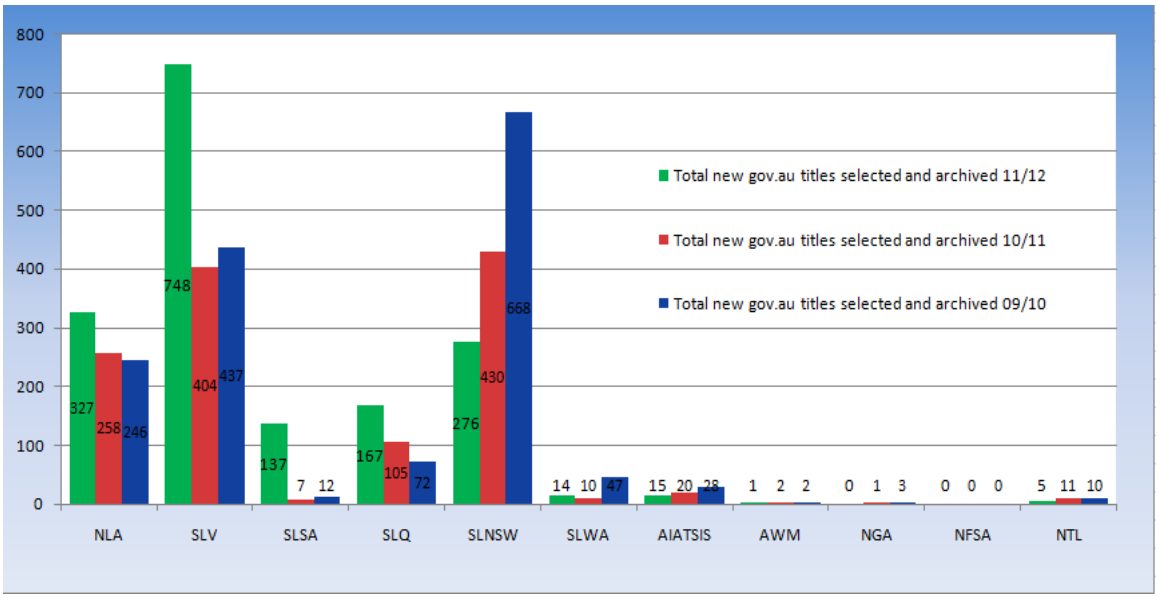


Figure 2: The number of new government titles archived each year over past three financial years.



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 2011-2012.

3.2 Australian web domain harvest

In the first quarter of 2012 the Library conducted the seventh 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 2012 and around one billion unique documents were captured, amounting to 41.88 terabytes of data. Following this harvest the combined total for all six Australian domain harvests has now reached 4.7 billion files amounting to around 176 terabytes of data.

The following table shows the amount of content collected for each of the five 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

In the absence of legal deposit provisions for online publications and web sites 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 2012 the second harvest of around 800 government URL domains was completed (the first being undertaken in March 2011). This harvest collected 13.6 million files amounting to 922 gigabytes of data.

In March 2012 a non-ongoing 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 will be to develop an infrastructure and interface to provide searchable public access to the Commonwealth Government web collections. It is expected that the 2011 and 2012 government collections will be made available to the public in mid-2013.

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 24% in usage (based on page views) during the 2011-2012 financial year over the previous year and an increase of 1.3 million page views over the previous financial year.

Usage in 2011 - 2012

Total page views	Average per month	Month of highest use	Month of lowest use
6,799,198	566,599	April 2012 732,133	July 2011 426,644

Usage in 2010 - 2011

Total page views	Average per month	Month of highest use	Month of lowest use
5,492,693	457,724	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

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 2011-2012.

Archived Title	Partner Responsible	Live site	Page views
First Families 2001	SLV	No	682,311
Sydney Centre for Studies in Caodaism	NLA	Yes	466,382
Life on the goldfields	SLV	No	248,696
GamesInfo	NLA	No[1]	225,750
Digger history	AWM	No	171,987
Antipodean SF	NLA	Yes[2]	146,582
Cultureandrecreation.gov.au	NLA	No	144,521
Centenary of Federation	NLA	No	117,668
Footypedia	NLA	No	103,632
Australian Federal Attorney-General	NLA	Yes[3]	97,147

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 previous issues for the publication.
3. The archived versions of this site include content from both current and previous Attorneys-General so some content is no longer available on the live site.

5. Preservation

Preservation activities particularly relevant to PANDORA during 2011-2012 include:

- The Digital Preservation Team have been testing software tools of potential interest for Digital Preservation activities at the National Library. These include tools that will be probably be used for the replacement of the PANDORA Archive. See *http://www.nla.gov.au/openpublish/index.php/nlasp/article/view/2452/2902*

- 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. 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. A paper on this process and the preservation intent statements was prepared by former WADiP Director Mr Colin Web, Mr David Pearson and Dr Paul Koerbin for submission to D-Lib Magazine.

- 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 2011-2012 the National Library continued its active participation in the International Internet Preservation Consortium (IIPC)[3] particularly in the work of the Preservation Working Group.

Ms Monica Omodei, Director of Web Archiving and Digital Preservation, attended the IIPC Annual General Assembly in Washington, USA, in April 2012. Ms Omodei gave a presentation at the General Assembly 'Open Day' conference titled: 'Trends in the Use of the PANDORA Archive'.

The NLA contributed to a collaborative archiving initiative of the IIPC to build a global collection of websites relating to the 2012 London Olympic Games. The National Library contributed around 20 sites to the project which was managed and hosted by the University of North Texas.

In September 2011 Mr Russell Latham, the National Library's Web Archiving Operational Supervisor, travelled to the UK and Denmark on a sponsored Binns Fellowship to visit various institutions and view their web archiving operations. Mr Latham visited the British Library and Wellcome Library to view their transition from using the PANDAS system in 2007 to the use of the web curator tool workflow system. He also visited the UK National Archives which uses outsourced archiving to collect UK government websites. Mr Latham visited two locations in Denmark that contribute to a national web archive collection. Denmark has had digital legal deposit since 2005 and makes regular whole domain snapshots using a tool they developed called *Netarchive Suite*. As well as doing domain harvests they also make thematic collections and undertake selective archiving of individual sites.

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

A number of presentations, papers and public discussions were given by National Library Web Archiving staff during the 2011-2012 financial year.

- Paul Koerbin, Manager Web Archiving, gave a presentation at the NSLA Heritage Collections Forum at the State Library of New South Wales in November 2011, titled: 'Next Steps Towards a New Model for Collaboratively Curating a National Web Archive'.

- Paul Koerbin, Manager Web Archiving, gave a paper at the Australasian Sound Recording Association (ASRA) 2011 Conference, titled: 'From Here to Perpetuity: Challenges (and a Few Confessions) in Preserving Web Based AV Content'.

- Paul Koerbin, Manager Web Archiving, was interviewed for and participated in the ABC Radio National *Future Tense* program titled *Digital Archaeology and the Temporary Nature of Technology* broadcast on 29 January 2012.

- Monica Omodei, Director of Web Archiving and Digital Preservation, gave a presentation at the Digital Humanities Australia Conference in Canberra in March 2012. The joint presentation in collaboration with Gordon Mohr of the Internet Archive was titled: 'Internet Content as Research Data: Challenges and Options for Collecting and Preserving'.

- Paul Koerbin, Manager Web Archiving, gave a presentation at the AGLIN Forum in Canberra in May 2012 titled: 'Collecting Government Web Content at the National Library of Australia'.

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 Slovenia and Korea.

8. Concluding summary

Some of the highlights of 2011-2012 include:

- Continuing steady growth of the Archive content (section 2).
- Completion of the seventh large scale harvest of the Australian web domain (section 3.2).
- Completion of the second harvest of Commonwealth Government web content under whole-of-government permission arrangements (section 3.3).
- Successful organisation of the first full PANDORA Partners face-to-face meeting in March 2012 (section 1.1.2).
- An international visit to view several national web archiving operations (section 6).
- Development of two national collections through a collaboration of PANDORA partners (section 1.4.2).