PANDORA - past, present, and future
National web archiving in Australia

Dr Paul Koerbin
Manager Web Archiving
National Library of Australia

National Conference on eResources in Malaysia
Penang, Malaysia, December 2012
National web archiving in Australia

1. PANDORA Web Archive – a brief history
2. Web archiving at the National Library of Australia (NLA) today
3. Issues for future web archiving at the NLA
4. Experiences and lessons learned
5. The importance of web archiving
1. PANDORA - context and history

1. What is web archiving?
2. Why it is important to do it?
3. How did the NLA approach it from the start?
4. Why the NLA has approached web archiving in the manner we have?
   • Timeline of major milestones from 1996 to now
What is web archiving?

• Web archiving involves:
  – Selecting or scoping what to collect
  – Collecting content from the web
  – Preserving what we collect
    • Strategies, metadata, maintaining bitstream, actions
  – Providing access to the collection
    • Long term and current
  • Creating heritage artefacts
    • Creating the time dimension for the web
What are we collecting?

• Web sites (and all they contain)
  – Complex objects
    • Text, images, media, style elements, client side scripts
    • No control over formats, systems, creation of content

• Includes sites with embedded media
  – lots of formats (mpeg, flv, QT, wmv, rm, Shockwave)

• Content is harvested with crawl robots
  – A browser view not underlying database
  – Dynamic content becomes static HTML
Why do web archiving?

• Statutory responsibility
  – National Library Act (1960)
    – Maintain, develop comprehensive collection relating to Australia and Australian people
    – Make national collection available in the national interest
• Just another publishing medium
  – Prior experience with non-print formats and ephemera
• National leadership
• Vision – recognising importance of new way information is communicated
How did we approach it?

• Selective approach
• Proof of concept approach
• Workflows and infrastructure developed and implemented in an ongoing way
• Initiated collaboration with other collecting institutions
Why did we approach it this way?

• Scalable to available resources
• Staged approach – do what we can
• Work within the legal constraints
• Able to realise objective of current access
• Heuristic – this was pioneering times and options were limited
Timeline

• April 1996: ‘Electronic Unit’ established
  – Part of ‘Acquisitions Branch’
  – 3 staff, 6 months to develop selection (scope) guidelines and identify resources

• September 1996: ‘Australian Serials and Electronic Unit’ established
  – Technical services restructure, multi-tasking, matrix management
  – October 1996 first titles harvested

• November 1996: ‘PANDORA’ born as ‘proof of concept project’
  – Preserving and Accessing Networked Documentary Resources of Australia
  – As at June 1997, 30 titles harvested

• May 1998: public access to PANDORA titles
Timeline (continued)

• July 1998: first PANDORA ‘partner’ began participation
  – 11th participant joined in 2010
• June 2001: PANDAS v.1 released
  – Web archiving workflow system developed by NLA
• 2002: Digital Archiving Branch
  – Our own identity at last!
  – Began first trial of ‘mainstreaming’ web archiving in Serials and Govt Deposit sections
• August 2002: PANDAS v.2 released
• July 2003: joined IIPC
  – International Internet Preservation Consortium
Timeline (continued)

- 2004: PANDORA added to UNESCO Australian Memory of the World Register
- July 2005: first .au domain harvest
- July 2007: PANDAS v.3 released
- 2010: PANDORA search moved to Trove
- May 2010: Whole-of-government (federal) permission
- March 2011: first gov.au bulk seed list collection
2. Web archiving at NLA today

- Organisation
- Staffing
- Collaboration participants
- NLA web archive collections
- Workflows
- PANDAS
Staffing – skills and expertise

• Collection development
  – Selection expertise in ‘new media’
  – Corporate objectives, priorities, resources

• Collection management
  – Cataloguing: MARC, LCSH, Dewey
  – PANDORA subjects

• Technical skills
  • Scoping gather filters and settings
  – Harvesting and code problem analysis and resolution (HTML, JavaScript, stylesheets)
  – Understanding web technologies

• Experience and self-learning
  – New technologies, Web 2.0, timely collecting, always new challenges
PANDORA participants

• 11 participants including the NLA
• State and territory libraries (not Tasmania and ACT)
• Major heritage institutions
  – National Film and Sound Archive
  – Australian War Memorial
  – Australian Institute of Aboriginal and Torres Strait Islander Studies
  – National Gallery of Australia
PANDORA participants

• Memorandum of Understanding
  – Respective obligations (NLA and Agencies)
  – Adherence to policy and procedures

• Curatorial and collection management (operational staff)
  – Selection (participants have their own guidelines)
  – Permissions
  – Harvesting – scoping and quality checking
  – Cataloguing
  – Publishing – access through PANDORA
Web archive collections

• Three collecting approaches and collections
  • ‘PANDORA Archive’ collection
    – Selective web archiving since 1996
  • ‘Australian Web Domain’ collection
    – Large scale, outsourced (IA), annual collection, since 2005
  • Australian Government Web Archive’ collection
    – Bulk seed list harvesting, outsourced (IA), annual collection, since 2011

*Plus* Asia-Pacific web collections using the *Archive-It* collecting and hosting service (since 2007)
Selective
‘targets’, ‘titles’

Themed
Curated seed lists or ‘titles’

2nd L Domain
e.g. gov.au

TL Domain
i.e. .au

Whole Web
Internet Archive

Small scale

Moderate scale

Moderate to large scale

Large scale

Reactive
Timely
Scheduled

Scheduled
Timely

Scheduled (moderate control)

Scheduled (low control)

Ongoing
Unscheduled

High curation

High curation

Moderate curation

Low curation

No curation control

High access Controlled

High access Controlled

Moderate access Moderate risk

Low/mod access
High risk

No access control

gov.au 2011-2012

AusCrawl 2005-2012

PANDORA

2
Domain
e.g. .au

Domain
i.e.

Whole
Web

Internet Archive

Large scale

Large scale

Large scale

Large scale
# Statistics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Files</td>
<td>167 million</td>
<td>4.72 billion</td>
<td>21 million</td>
<td>4.9 billion</td>
</tr>
<tr>
<td>Data</td>
<td>7.52 TB</td>
<td>175 TB</td>
<td>1.45 TB</td>
<td>184 TB</td>
</tr>
</tbody>
</table>
Statistics - all web collections

Files archived

- 21 million
- 167 million

Data archived

- 1.45 TB
- 7.52 TB

Legend:
- PANDORA
- Whole Domain Archive
- Aust Govt Web Archive
Web archiving workflow
<table>
<thead>
<tr>
<th>Worktrays</th>
<th>National Library of Australia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transferred</strong></td>
<td></td>
</tr>
<tr>
<td>View Unacknowledged Titles Transferred to Me [1]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Selection</strong></td>
<td></td>
</tr>
<tr>
<td>View Nominated Titles [32]</td>
<td>Refresh</td>
</tr>
<tr>
<td>View Monitored Titles [52]</td>
<td>Refresh</td>
</tr>
<tr>
<td>View Selected Titles [50]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Permission</strong></td>
<td></td>
</tr>
<tr>
<td>View Permission Requested Titles [153]</td>
<td>Refresh</td>
</tr>
<tr>
<td>View Permission Granted Titles [32]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Gather</strong></td>
<td></td>
</tr>
<tr>
<td>View Scheduled Titles [213]</td>
<td>Refresh</td>
</tr>
<tr>
<td>View Gathering Titles [2]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Preserve</strong></td>
<td></td>
</tr>
<tr>
<td>View Instances For Upload [12]</td>
<td>Refresh</td>
</tr>
<tr>
<td>View Instances in QA [17]</td>
<td>Refresh</td>
</tr>
<tr>
<td>View Instances Referred to IT [11]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Publish</strong></td>
<td></td>
</tr>
<tr>
<td>View Archived Titles [16]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Catalogue</strong></td>
<td></td>
</tr>
<tr>
<td>View Titles Requiring Cataloguing [119]</td>
<td>Refresh</td>
</tr>
<tr>
<td><strong>Reports</strong></td>
<td></td>
</tr>
<tr>
<td>View Generated Reports [330]</td>
<td>Refresh</td>
</tr>
</tbody>
</table>
PANDAS system workflow diagram

1. Identify Resource
   - Management/policy decision

2. Create Metadata
   - Create a file in the system
   - Associate publisher, agency owner, indexer
   - Apply restrictions at title level
   - Add copyright notes
   - Add gathering information

3. Gather Resource
   - Apply filters for the gathering engine
   - Gather resource to the work area or storage
   - Make a preservation copy of the resource files
   - Make metadata files of the preservation copy

4. QA on the Resource
   - Access resource files in the work area
   - Conduct quality assurance checks on the files
   - Create a report for IT if there are any problems

5. Archive Resource
   - Archive the resource
   - Make a display preservation copy

6. Display Resource
   - Add resource to the list (collections)
   - View the Title Entry Page (TEP) generated at run time from the metadata
   - Add groups and issues for the resource on the TEP
   - Associate metadata in Dublin Core format with the TEP
   - Provide resource discovery mechanisms & searching options
3. Issues for the future

• The environment for web archiving
• Organisation, infrastructure and workflows
• Access, search, discovery and promotions
Future web archiving at the NLA

• Issues – the environment
  – Ever increasing scale of information online
  – Dynamic delivery of content
  – Complexity of content
    • Technical and intellectual
  – Changing technology
  – Changing understanding of publishing
  – Legal deposit (collecting) and access rights
Future web archiving at the NLA

• Issues – the organisation
  – Infrastructure
  – DLIR
  – Emphasis on digital collecting
  – New approaches to collaboration
  – New workflows and systems
Future web archiving at the NLA

• Access, search, discovery, promotion
  – Full-text search
  – URL search
  – Browse paths and collections
  – Trove – one search service
  – More innovative discovery and visualisation
  – Research (big) data
4. Experience suggests ...

1) Important to take actions as soon as practicable
   – Web content is ephemeral like no other
   – The timing of collecting content is critical
   – Things will change but not get easier or simpler to deal with
   – Do what you can – can’t solve all problems up front
   – Gain experience by doing
   – Be as agile in your approach as possible so as to deal with the dynamic nature of the task
Experience suggests ...

2) Make it sustainable

– Understand it is commitment over the long term
  • The preservation objective
– Consider the implications of your technology and infrastructure choices
– Chose the right approach for your business
  • in-house or out-sourced service?
– Engage and support skilled, interested, self-motivated staff
Experience suggests ...

3) Focus on purpose and outcomes
   – Add value through a focus on access and discovery
   – Demonstrate the value of collecting web materials
     • Embrace the responsibility of creating the historic cultural artefacts out of the ephemeral
     • Promote the activity – papers, conferences, social media, website
   – Engage stakeholders through collaboration and opportunities to contribute
     • Provided this genuinely advances the objectives
Importance of web archiving

• Major task is to advocate the importance of web archiving
  – Hard to demonstrate when content remains on live site
  – Value of content is variable
  – So much a part of everyday life – easy to overlook value
Importance of web archiving

• Only evidence of our cultural expression from the web is in the archives
• Vulnerable medium – ‘unpublishing’
• Democratic medium – value in understanding our society
• Grey literature – moving to online only
• Avoid the digital ‘black hole’
Thank you.

Dr Paul Koerbin
Manager Web Archiving
National Library of Australia
pkoerbin@nla.gov.au