PANDORA
AUSTRALIA'S WEB ARCHIVE
National Library of Australia and Partners

PANDORA, AUSTRALIA'S WEB ARCHIVE

http://pandora.nla.gov.au/
is a selective archive containing copies of significant Australian online publications and web sites issued on the Internet. The National Library of Australia and its partners are building the Archive to ensure long-term access to significant Australian documentary heritage that is published online.

PANDORA was placed on the Memory of the World Australian Register in August 2004.

PARTICIPANT AGENCIES

- Australian Institute of Aboriginal and Torres Strait Islander Studies
- Australian War Memorial
- National Film and Sound Archive
- National Library of Australia
- Northern Territory Library
- State Library of New South Wales
- State Library of Queensland
- State Library of South Australia
- State Library of Victoria
- State Library of Western Australia
- National Gallery of Australia (pending)

CONTENT

Titles in the Archive are selected according to selection guidelines developed by all partners and published on the PANDORA Web Site at http://pandora.nla.gov.au/guidelines.html With the permission of publishers, the National and State libraries archive those resources relating to the published output of their jurisdictions. The National Film and Sound Archive takes responsibility for publications and web sites relating to film and music; the Australian War Memorial archives those relating to military history; and AIATSIS archives those of our Indigenous peoples.

The Archive contains a wide range of titles. High priority is given to government publications, academic e-journals and conference proceedings. Partners also endeavour to document Australian life as it is represented on the Internet , and include sites representing cultural activity, Australia's diverse peoples, community concerns, political activity, sport, and many other topics. Many titles are re-gathered on a regular basis to capture updated content.

PANDORA is essentially a collection of computer files, which constitute copies of the publications and web sites selected by partners. A title in the Archive may consist of a single file, such as a text document in Portable Document Format (PDF), e.g., *Annual report to the NSW Environment Protection Agency* http://pandora.nla.gov.au/tep/42658, or it may be a complex web object, such as a large web site, consisting of thousands of files in a variety of formats, including text, sound, image or video, e.g., *Sydney 2000: official site of the Sydney 2000 Olympic Games*.

ACCESS

Titles in the Archive are accessible free of charge via the Internet at http://pandora.nla.gov.au/ Most titles are available to anyone, anywhere in the world, with an Internet connection. Access is restricted to a very small proportion of titles, mainly for commercial reasons, and these can be viewed on a single PC in the Library's Main Reading Room.

People can find out about titles that are in the Archive by searching partners' online catalogues or by searching the National Bibliographic Database (Libraries Australia). Access is provided via hotlinks in the catalogue record to the title in the Archive. Access is also available via subject and title lists on the PANDORA Web Site. Full-text searching is available using the Library's single search discovery service Trove. Commercial search engines, such as Google and Yahoo!, index the Archive down to the level of individual titles, but not the Archive contents.

QUALITY ASSURANCE

Significant effort is invested in ensuring the authenticity and integrity of each title archived. In copying (gathering) a publication or web site into the Archive, the policy of partners is to maintain its 'look and feel', that is, its appearance and functionality, as well as its contents, to the fullest extent possible. After gathering from the publisher's web site, each title is checked to make sure it is complete and functional.

PERSISTENT IDENTIFIERS

Each item in the Archive, from the title level down to component files, has a unique persistent identifier automatically assigned by the PANDORA Digital Archiving System (PANDAS). This enables authors to cite works and parts of works (e.g., journal articles) in the Archive using the appropriate persistent identifier. Readers can return to the cited item in the Archive again and again, confident that it will remain there persistently and that it will not change.

PANDORA DIGITAL ARCHIVING SYSTEM

To support the activities and workflows involved in contributing titles to the Archive, the National Library has developed the web-based PANDORA Digital Archiving System (PANDAS). Partners use PANDAS to:

- Register titles for inclusion in the Archive;
- Record publisher permissions;
- Set the gathering schedule – once only or regular gathering;
- Undertake quality assurance and record any actions taken or decisions made about a title;
- Consign the title to the Archive
- Create the title entry page and the list of instances archived;
- Link to publishers' copyright statements.

LEGAL DEPOSIT

Unlike other materials collected by the National Library, such as books, serials and newspapers, digital materials are not covered by the Legal Deposit provisions of the *Copyright Act 1968*. This means publishers are under no requirement to deposit web materials with the National Library. It also means that a specific licence under the Copyright Act 1968 needs to be sought from each publisher before the Library proceeds with collecting and making accessible archived websites. This process imposes a considerable limitation on the amount of content that can be collected.

PRESERVATION

The Library intends to provide perpetual access to titles archived in PANDORA. This poses a significant challenge, as software and hardware required for display changes quite quickly. *A digital preservation policy for the National Library of Australia* http://www.nla.gov.au/policy/digpres.html includes strategies for the PANDORA Archive. To preserve access to titles the Library will employ:

- Some technology preservation, including maintenance of software and some hardware;
- Negotiating with publishers to supply stable source files of some streaming or dynamic formats;
- Migration strategies for some file formats;
- Use of emulators for some file formats;
- Keeping and refreshing some files not amenable to migration or emulation in the hope that a suitable access pathway will emerge.

The Library has conducted a risk assessment which identifies in detail the risks involved in specific file types that make up the complex web objects in PANDORA.

COLLABORATION

The Library is committed to working with other libraries and cultural collecting agencies to find improved web archiving solutions. It is playing an active role in the International Internet Preservation Consortium http://netpreserve.org/about/index , contributing to the Deep Web Archiving Working Group and to the Curator Tool Project. It also coordinated the development of UNESCO's *Guidelines for the preservation of digital heritage* http://unesdoc.unesco.org/images/0013/001300/130071e.pdf which were designed to assist countries to develop policies and procedures on collecting and preserving digital heritage.

STATISTICS (as at January 2010)

Number of titles* – 24,365
Number of instances (repeat gatherings) – 52,124
Government publications – approximately 50% of total
Size of Archive (Display) – 3.59 terabytes
Usage 2008-09 – 2,458,772 page views

*"Title' is the entity selected for archiving and for which a catalogue record is created. It may be a whole or part of a web site, or a discrete publication.

TECHNICAL INFRASTRUCTURE

The architecture of PANDORA is as follows:

1. PANDORA Digital Archiving System (PANDAS), including a harvester
2. Storage system for long-term archiving and access: DOSS (Digital Object Storage System);
3. Public access/delivery system
4. Search index (Lucene) via Trove discovery service

Collection management system

PANDAS is a workflow system that enables collection managers to undertake the various tasks associated with building a selective web archive and to record information about titles and actions taken. The user interface is web-based and requires no special software to be installed on the desk top. Collection managers require a range of web browser plug-ins and associated software to view publications being archived. They system consists of :

- Workflow/management system written in Java using the WebObjects application framework;
- Metadata repository using Oracle 8i RDMS;
- Website offline browser and mirroring tool, HTTrack;
- Reporting facility based on Oracle Forms and Reports.

The workflow and metadata systems are supported on Sun Solaris servers. The gatherer uses a dedicated Linus Server. The web site analysis system runs under NT. The reporting facility is client based and runs on users windows-based desktops.

DOSS

Digital objects associated with PANDORA are stored in two ways. The preservation master and access master copies are stored on Unix file systems in a consolidated format on the Library's DOSS. These are archived and sent off-site for safe-keeping. Copies for current access are not stored on the DOSS but as individual files on UNIX files systems on the PANDORA display machine, which enables faster access.

Delivery system

The public delivery system is also built using Apache/WebObject/Java and Oracle to provide resource discovery, navigation and access control services. The actual items of digital content are delivered as static content through Apache. The service is hosted on Sun Solaris server.

Ongoing development

The Library is committed to ongoing development of PANDORA and its systems. A completely re-engineered version of PANDAS with enhanced workflows, interface and functionality was released as PANDAS 3 at the end of June 2007.

For more information about the PANDORA Archive go to the PANDORA web site http://pandora.nla.gov.au/ or email webarchive@nla.gov.au

updated 15 Jan 2010