Zack Lischer-Katz
10/24/10
Digital Preservation

Multi-Institutional Repository Project: PANDORA

PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia), a project established by the National Library of Australia in 1996, was initially developed to address concerns about increasing levels of information being made available solely on the internet (websites, eJournals, etc.). Edgar Crook explains that many libraries initially thought that access could be ensured through creating and maintaining catalogs of the online material, but it quickly became clear that "many online publishers could not guarantee the long-term availability of their own publications, particularly given the unreliability of many Internet Service Providers at that time."[1] Thus, the National Library of Australia realized that it needed to maintain its own copies of digital materials if it hoped to keep them accessible to the Australian people.

PANDORA, as the project is now known, started out fairly slow. Of the hundreds of titles selected for the first attempt at web archiving, only 30 were successfully archived by the end of the first year (1996-1997) and by the second year, the number of archived sites had only increased to 90.[2] Now, after 14 years of work, the PANDORA project is archiving at a far quicker pace, with 181 titles added in just the last month, bringing the collection up to 26,183 titles, according to statistics on the PANDORA website.[3]

Unlike other web archiving projects like the Internet Archive that attempt domain-wide crawls, PANDORA and its partners "do not attempt to collect all Australian online publications

[1] Edgar Crook, "The Work of Pandora," *National Library of Australia Gateways,* No. 82 (August 2006), accessed 10/19/10 at http://www.nla.gov.au/pub/gateways/issues/82/story01.html
[2] ibid.
[3] http://pandora.nla.gov.au/statistics.html

and web sites, but select those that they consider are of significance and to have long-term research value."[4] The PANDORA website states that it only archives websites that are

> About Australia, [...] by an Australian author, on a subject of social, political, cultural, religious, scientific or economic significance and relevance to Australia, or are by an Australian author of recognized authority and make a contribution to international knowledge.[5]

Each partner institution creates a document that lists its selection criteria, posts it on the PANDORA site: http://pandora.nla.gov.au/selectionguidelinesallpartners.html.

As a rule, state libraries are charged with selecting and collecting regional and state-related materials, while the National Library collects material of a national interest. Institutions with specific subject areas will collect within the areas that are most appropriate to their mission statements. For example, the Australian War Memorial is charged with collecting sites "relating to Australian military history," while the National Film and Sound Archive collects moving images and sound-related web pages.[6] Currently, PANDORA's list of partners consists of

> The Australian Institute of Aboriginal and Torres Strait Islander Studies; Australian War Memorial; National Library of Australia; Northern Territory Library; National Film and Sound Archive; State Library of New South Wales; State Library of Queensland; State Library of South Australia; State Library of Victoria; and the State Library of Western Australia.[7]

Besides acquiring web content through these partners, PANDORA also acquires a small number of websites through deposit requests from content producers. An electronic form is provided on the website that allows producers to submit their sites to be considered for deposit.

It is worth noting that, regardless of its stated collection policies, PANDORA *has* conducted several domain-wide crawls. As Fellows et al. discuss in "Separating the Wheat from

---

[4] "PANDORA Selection Guidelines," *National Library of Australia*, updated 9/10/10, accessed 10/20/10: http://pandora.nla.gov.au/selectionguidelinesallpartners.html
[5] ibid.
[6] ibid.
[7] "PANDORA Partners," *National Library of Australia*, accessed 10/24/10: http://pandora.nla.gov.au/partners.html

the Chaff: Identifying Key Elements in the NLA .AU Domain Harvest," PANDORA conducted crawls of the entire .au domain in 2005 and 2006,[8] along with its usual selective, subject or region-related archiving.

PANDORA's collection of web sites is accessible through its homepage, http://pandora.nla.gov.au/. Access is open to anyone in the world with an Internet connection and a modern web browser. However, PANDORA admits "there is a very small proportion of titles that are restricted for commercial reasons, or because their content is sensitive," but this content can be accessed onsite at the National Library.[9] Users of the website may browse the content of PANDORA via one of 18 subject headings, the first letter of the website title, or a long, alphabetized list of all titles in the collection. In addition, the National Library's new online database/catalog, *Trove,* is designed to search within PANDORA's collection of websites. Trove's search engine allows for fairly precise searching using Keyword, Title, Creator, Subject, ISBN, ISSN, and Public Tag (keyword tags that users can update), and content can be limited to only Australian-produced content, or by Format, Availability, and Language. PANDORA's general cataloging system uses a complex amalgam of

> *AACR2*, 2nd ed., 2002 revision, Ch. 9 'Electronic Resources'; OCLC's *Cataloguing Internet resources: a manual and practical guide*; *Cataloguing electronic resources: OCLC-MARC coding guidelines*; the Library of Congress Guidelines for the use of field 856; Guidelines for coding electronic resources in Leader/06; and *MARC 21 Format for bibliographic Data Field List*.[10]

PANDORA uses a workflow and metadata system called PANDAS (PANDORA Digital Archiving System), which is currently in version 3.0. The system allows the National Library

---

[8]Geoff Fellows, et al., "Separating the Wheat from the Chaff: Identifying Key Elements in the NLA .au Domain Harvest." *Australian Academic & Research Libraries* v. 39 no. 3 (September 2008) p. 137-48.

[9] "Frequently Asked Questions About Pandora," *National Library of Australia*, updated February 8, 2010, accessed Oct. 24, 2010. http://pandora.nla.gov.au/panfaqs.html#whoaccess

[10] "Original Cataloging of Electronic Resources," *National Library of Australia,* accessed 10/24/10, http://pandora.nla.gov.au/manual/kincat.html

and its partners to set seed URLs for crawls, add metadata, keep track of projects, and record permissions information, store correspondence with content creators, and generate request emails. [11] These last few functions of the system are integral to ensuring that websites are archived legally. The National Library cannot take advantage of national legal deposit requirements for acquisition, so PANDORA and its partners must request permission from each website producer before a site can be archived.

PANDORA also employs a system of "persistent identifiers," which "through a combination of managed URLs and a resolver system" ensures that a digital object will always be accessible using the same URL, even if the object is moved to a different location in the archive.[12] The "resolver system" works by accepting "web requests which refer to a digital collection item by its persistent identifier and redirects them to the current storage location of that item."[13]

While PANDAS manages PANDORA's digital objects and workflow, the National Library's Digital Object Storage System (DOSS) archives the files.[14] "The system consists of a Sun E450 server, a CLARiiON FC4700 disk array, [and] a storageTek Tape Library connected via a SAN switching infrastructure" with a total capacity, as of 2005, of 14 TB of disk space and 60 TB of offline tape storage.[15] This is a managed repository, with strategies in place for future migration and emulation. In addition, PANDORA archives three copies of each digital object: "a preservation master, which is 'as-acquired', a display master which has been post-acquisition QA

---

[11] "PANDAS Version 3 Release Notes," *National Library of Australia*, Accessed 10/24/10, http://pandora.nla.gov.au/pandas3notes.html
[12] "Activities," *National Library of Australia*, Accessed 10/24/10, http://www.nla.gov.au/initiatives/persistence.html
[13] "Digital Services Project: Overview," *National Library of Australia*, accessed 10/24/10, **http://www.nla.gov.au/dsp/#doss**
[14] ibid.
[15] ibid.

processed, and a metadata master, which contains information collected during pre-archival."[16] Along with this "Preservation Archive," PANDORA employs two other digital "archives," one to manage files and prepare them for ingestion into the Preservation Archive and create access-ready derivatives (the "Working Archive"), and the other houses derivative files (the "Display Archive").[17] Copies made from files in the Display Archive are hosted on the web server, where they are made accessible to the public.

To ensure the long-term sustainability of its collection, PANDORA plans to use a combination of strategies, including migration, technology preservation, emulation, and refreshing.[18] To facilitate this multiplicity of strategies, preservation metadata is captured for each digital object and stored within the preservation archive.[19] PANDORA's project web migration test project, conducted 2001, is a illustrative example of how its management team approaches the problem of long-term sustainability, and why documentation about file formats and standards is so important. In this project, an initial analysis of the archive's collections showed a variety of compatibility problems, including "127 tags dead in HTML 4.0; 7 million tags due to be made nonstandard in later versions of HTML; and 14 million tags with deprecated attributes."[20] Given these preservation risks, the preservation strategy that the test project adopted for older html-based pages was to systematically "migrate these obsolete and deprecated tags to equivalent tags that are valid in the latest version of HTML."[21]

---

[16] Steven McPhillips, "PANDORA: Technical Details," *National Library of Australia,* updated 08/23/04, accessed 10/20/10, *http://pandora.nla.gov.au/pandoratech.html*
[17] ibid.
[18] Warwick Cathro, et al., "Archiving the Web: The PANDORA Archive at the National Library of Australia," presented at the National Library of Australia at the *Preserving the Present for the Future Web Archiving Conference*, Copenhagen, 18-19 June 2001 http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1314/1600
[19] ibid.
[20] ibid.
[21] ibid.

As of October 2010, PANDORA's future looks bright. Not only has the project developed a preservation repository from the ground up, and has created strong partnerships with the major libraries and collecting institutions within Australia, but its collection is now more visible and accessible, through its integration with the National Library's online database *Trove*, (in April of 2010). Online statistics[22] show that in September 2010, PANDORA's online collection received over 117,000 visitors, each viewing an average of 4.3 pages per visit (although they don't specify if these were unique visitors). In that time PANDORA archived an additional 143GB of data, adding 3,229,535 files to its collection, or 181 unique titles (or 937 archive instances).[23] Moving forward, PANDORA could benefit greatly by changes in the legal deposit legislation at the national level. According to PANDORA's website, Tasmania is the only state in Australia that has clear and unambiguous legal deposit legislation that covers online content. This has enabled its state archives to preserve online content "without the need to seek permission from publishers,"[24] which yielded two important websites: *Our Digital Island: Preserved Tasmanian Web Sites* and *STORS: Long-term storage of Tasmanian electronic documents*. In addition to efforts to reform Australia's legal deposit laws, PANDORA is also eager to collaborate with other institutions and organizations, like the International Internet Preservation Consortium, and establish standards for "acquisitions techniques, preservation formats and the like"[25] (McPhillips, 2004). With its many years of experimentation and innovation, the leaders of the PANDORA project are certainly well poised to contribute significantly to the establishment of worldwide web archiving standards.

[22] http://stats.nla.gov.au/cgi-bin/report_index.cgi?report=pandora
[23] http://pandora.nla.gov.au/statistics.html
[24] "Legal Deposit," *National Library of Australia*, Accessed 10/24/10, http://pandora.nla.gov.au/statistics.html
[25] McPhillips, 2004.

Bibliography

"Activities: Persistent Identifiers." *National Library of Australia*. Accessed 10/24/10. http://www.nla.gov.au/initiatives/persistence.html

Beagrie, Neil. *National Digital Preservation Initiatives: An Overview of Development in Australia, France, the Netherlands and the United Kingdom and of Related International Activity*. Washington, D.C.: Council on Library and Information Resources and Library of Congress, 2003. http://www.clir.org/pubs/reports/pub116/pub116.pdf

Cathro, Warwick, Colin Webb and Julie Whiting. "Archiving the Web: The PANDORA Archive at the National Library of Australia" Presented at the National Library of Australia at the *Preserving the Present for the Future Web Archiving Conference*, Copenhagen, 18-19 June 2001 http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1314/1600

Crook, Edgar. "The Work of Pandora." *National Library of Australia Gateways,* No. 82 (August 2006). Accessed 10/19/10 at http://www.nla.gov.au/pub/gateways/issues/82/story01.html

Day, Michael. *Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust*. Bath: UKOLN, University of Bath, 2003. http://www.jisc.ac.uk/uploaded_documents archiving_feasibility.pdf

"Digital Services Project: Overview." *National Library of Australia*. Accessed 10/24/10. http://www.nla.gov.au/dsp/#doss

Fellows, G., et al., "Separating the Wheat from the Chaff: Identifying Key Elements in the NLA .au Domain Harvest." *Australian Academic & Research Libraries* v. 39 no. 3 (September 2008) p. 137-48.

"Frequently Asked Questions About Pandora." National Library of Australia. Updated Feb. 8, 2010. Accessed Oct. 24, 2010. http://pandora.nla.gov.au/panfaqs.html#whoaccess

Hodge, Gail and Evelyn Frangakis. *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice: A Report Sponsored by the International Council for Scientific and Technical Information and CENDI (*February 2004). http://cendi.dtic.mil/publications/04-3dig_preserv.html

McPhillips, Steven. "PANDORA: Technical Details." National Library of Australia. Updated 08/23/04. Accessed 10/20/10, *http://pandora.nla.gov.au/pandoratech.html*

"New Look marks Milestone for PANDORA Tour de Force." National Library of Australia *Gateways,* No. 46 (August 2000). Accessed 10/19/10

http://www.nla.gov.au/pub/gateways/archive/46/p01a01.html

"PANDAS Version 3 Release Notes." *National Library of Australia*. Updated 5/7/10. Accessed 10/20/10. *http://pandora.nla.gov.au/pandas3notes.html*

"PANDORA Partners." National Library of Australia. Accessed 10/24/10. http://pandora.nla.gov.au/partners.html

"PANDORA Overview." National Library of Australia. Updated 4/27/10. Accessed 10/21/10. http://pandora.nla.gov.au/overview.html

"PANDORA Selection Guidelines." National Library of Australia. Updated 9/10/10. Accessed 10/20/10. http://pandora.nla.gov.au/selectionguidelinesallpartners.html

Pymm, B., et al., "Dealing with Digital Collections: Interviews with the National Library and Selected State Libraries of Australia." *Australian Academic & Research Libraries* v. 38 no. 3 (September 2007) p. 167-79

Smith, Wendy. "The National Library of Australia's PANDORA project." *Libri* v. 47 no. 3 (September 1997) p. 169-79.

Smith, Wendy. "Wine on the Web: Australian Wine Information on the Web and its Prospects for Long-Term Preservation and Access." *Australian Academic & Research Libraries* v. 35 no. 2 (June 2004) p. 111-28.