

CASE STUDY: PANDORA - the Australian web archive

Prepared by Maria Garasa

1 November 2013

“Digital information is a vital resource in our knowledge economy, valuable for research and education, science and the humanities, creative and cultural activities, and public policy. But digital information is inherently fragile and often at risk of loss. Access to valuable digital materials tomorrow depends upon preservation actions taken today; and, over time, access depends on ongoing and efficient allocation of resources to preservation.” (Blue Ribbon Task Force 2010: 1)

As captured by the statement above, the preservation of digital material has a critical role to play in safeguarding the cultural heritage of a society, and carries an important economic value. It also comes with a significant economic cost in terms of the resources required to preserve the material and to maintain its future access. It is also fraught with numerous challenges, that are particularly technological and legal in nature, and that are inherently different from those traditionally encountered for non-digital material. The field of web archiving brings its own set of added challenges due to the dynamic, ephemeral nature of the web, and the relentless growth in the quantity and complexity of on-line materials.

As a pioneer in web archiving, the National Library of Australia (hereinafter the NLA), has pragmatically, and arguably, successfully tackled many of these challenges through its PANDORA Web Archive (hereinafter PANDORA or the Archive). From the author’s perspective, PANDORA has achieved its success by having realistic parameters of what can be achieved and the mechanisms for achieving them within a constrained budgetary environment. In this manner economic sustainability has been ensured, and in turn, the sustainability of the Archive. The NLA continues to be innovative in developing tools to facilitate more effective and efficient approaches to digital preservation, and adopted earlier this year *Preservation Intent Statements* for its digital collections, including PANDORA. This information will be used as the NLA sets in place new workflows and infrastructure that will help support



preservation management for all its digital collections. This essay will take a closer look at this latest development, namely the preservation intent information for PANDORA, and what it means for the collection. It will also look at how the NLA has fostered the Archive's economic sustainability and what opportunities could lie in the future.

In order to set the context for the analysis above, a brief overview of PANDORA's beginnings is in order. As stated, digital preservation is inherently different from traditional preservation or archival library practices, and has only come to the forefront of library and information science relatively recently. As the course material points out, the field is particularly challenging for library and information practitioners as there are no formal, and commonly accepted standards for storing and archiving digital material, although, it could be argued that a 'standard practice' in the field is surfacing due to efforts undertaken by different institutions, with PANDORA serving as one important example. (Harvey 2012: 1) In the absence of any standards, the NLA began the development of the PANDORA collection back in 1995 by first setting out scoping and selection guidelines prior to harvesting any web material. Their development was guided by the NLA's conviction that it had a responsibility to collect and preserve the published national heritage, regardless of format. Selection was also influenced by resource issues and by the NLA's goal that what it collects and preserves must be accessible. A heavy emphasis was also placed on quality assurance. (Harvey 2005) As a result, selective web harvesting was considered as being the only realistic and feasible approach.

PANDORA's initial selection guidelines were based on the NLA's existing collection development policy. The items selected were to have significant Australian content or authorship, and the work should be authoritative with long-term research value. Furthermore, in the absence of legal deposit legislation for digital publications, permission to archive from the owners/publishers needed to be sought in advance to facilitate immediate accessibility. By 1998, procedures were fully in place, and PANDORA became publicly available with a modest 36 titles. In 1998, the NLA also sought the participation of other Australian state libraries and cultural institutions. Today, collaboration is a hallmark of PANDORA where 11 organizations participate in the Archive.



The foresight to establish selection guidelines provided the NLA with a framework to answer the critical question related to digital preservation – what should be preserved? Today, selection guidelines are set by each of the participating institutions and are not standardized. In general, they set out priority areas for selection/collection; inclusions and exclusions for digital objects; file format; management of versions/editions, etc. Selection is limited to publications or digital material that exists only online. Moreover, the selection process places intellectual content before the file format of the online publication, which means that a high percentage of the web sites selected need further work in order to make them accessible. In the case of the PANDORA collection, the software used rewrites the code, resulting in the generation of archived copies that are intrinsically different from the originating live web versions. While the ‘look and feel’ of digital objects is retained, the NLA has accepted the fact that it needed to be flexible in terms of safekeeping their authenticity. Today, the PANDORA collection consists of about 75% text, 20% images, and 5% multimedia and style elements, including images. Over half of the collection is comprised of government publications, with the NLA being responsible for archiving the majority of titles. (Webb *et al.* 2013: 13).

Earlier this year, the NLA adopted *Statements of Preservation Intent* for all of its digital collections, which in effect solidly places preservation as a central piece of its collection efforts. A key impetus for adopting this approach is the NLA’s goal of ensuring accessibility to Australia’s documentary heritage. In relation to its digital collections, access to these digital objects can only be maintained if the technical/technological aspects necessary for their preservation are understood. The approach borrows from the idea of *significant properties*, which can be described as those aspects of a digital object, which must be preserved over time in order for the digital object to remain accessible and authentic. The view of the NLA is that the application of the significant properties concept facilitates the articulation of the ‘what’, ‘why’, ‘who’, and ‘for how long’ behind the preservation selection (i.e., what should be preserved; why should it be preserved; for how long; and who is responsible?), which it feels is needed before an understanding of the attributes needed for preservation can be gained (i.e., the ‘how’). It also challenges practitioners to think about what an adequate level of accessibility could look like. In practice, the approach is asking curators to consider the



implications of their collection development policies and collecting decisions in terms of the realities of providing and maintaining access. (Webb *et al.* 2013)

Web collections, such as PANDORA, present special challenges compared to other digital collections primarily due to the volume, diversity and complexity of the content over which the preserving institution has absolutely no control. At the same time, preservation planning for web collections are not as well addressed globally compared to other digital collections. For the NLA, its web collections are considered a high priority, and are also the most challenging. Its web collections fall into three categories: (i) PANDORA's selective web archive; (ii) 'whole domain harvests' capturing a broad periodic snapshot of the Australian web domain; and, (iii) coherent bulk collections, such as the Australian government web domain (.gov.au). Due to legal reasons, the PANDORA content is the only one currently available for public access. For PANDORA, the intent of web archiving is as follows:

1. To retain in perpetuity all PANDORA digital preservation masters, including all associated metadata. Moreover, all technical properties are maintained to the full extent possible.
2. Provide primary importance to content, connections and context. How it is ultimately presented to the user is given secondary importance.
3. Give primary importance to the 'display master', i.e is the copy that includes the results of quality assurance and curatorial work. The original harvested master, that is the 'preservation master', is retained at least at the bit level.
4. Maintain only for as long as useful, the derivative copy derived from the display master and created for display and access. A new derivative version can be generated according to future access requirements.

Due to the challenging nature of web objects, the NLA accepts that what it is preserving is NOT a mirror representation of the content in the web, or even the website, but a snapshot of content that was once arranged and published as a website, with only limited functionality of the original. A key focus is to limit the loss of web content. As such, the manner in which the content is collected and displayed may mean that there is significant limitation on the



presentation of the archived object as an authentic record of the publisher's original data or even the version of that data as originally published on the web. (NLA 2013d; See Appendix 1 for the Statement of Preservation Intent for Selective Web Harvesting)

Currently, there are a number of file types in the PANDORA web archive that present access problems, given the way that the content is collected and managed in the Archive. The intent statements will help the NLA to decide whether it wants to keep the content accessible over an extended period of time; whether it wants to solely view the content; or, whether it wants to manipulate the content. Practically speaking, the intent classifications would lead the NLA to consider the following action options:

1. Take no action; or
2. Replace the failed access software; or
3. Migrate the content to another format that does preserve the presentation of the properties which have been deemed as significant;
or
4. Emulate the presentation of those significant properties of the content;
or,
5. Maintain the bits and documentation that will support action in the future when more effective option actions may become available.
(Webb *et al.* 2013: 14).

The NLA is currently redeveloping its digital infrastructure through its four-year Digital Library Infrastructure Replacement (DLIR) Project, which is aimed at better enabling the NLA 'to collect, generate, manage, and preserve and provide ongoing access to the rapidly escalating volume of digital information resources that fall within its collecting mandate'. (<http://www.nla.gov.au/about-us/dlir>) The *Preservation Intent Statements* will be key in setting in place the new preservation management systems (and likely a replacement to PANDORA) that will arise from the exercise.

A discussion of digital preservation cannot be done without consideration of its economic implications. One could indeed argue that digital preservation is



as much an economic issue as it is a technical/technological one. In their work, Lavoie (2008) and Bradley (2007) argue that the biggest single risk to sustained access to digital information is economic. Lavoie contends that the 'lack of economically sustainable models for digital preservation activities represents just as real threat to the long-term persistence of digital material as the more traditional scourges of media decay and technological obsolescence'. (Lavoie 2008: 2). In effect, without adequate sustained funding, institutions are left unable to pay for the 'digital mortgage' (i.e., ongoing costs) that results from their selection decisions. (Harvey 2012)

The PANDORA web archive finds itself in the enviable position where there is an institutional commitment by the NLA to allocate resources to support the long-term management of its web collections, and for digital preservation in general. That said, the NLA has no additional funding to undertake this activity, and funds the activity from the ongoing annual budget allocation that it receives from the Australian government. Even in its initial start as a pilot project in 1995, PANDORA did not receive any special short-term funding, which in hindsight proved advantageous as it forced the activity to be funded from existing resources and ensured its continuity when the pilot came to an end. The titles archived have always been freely accessible, and there have been no revenue-generating activities to date, nor is there any indication that this could be expected in the future.

Given the above, PANDORA has always faced significant resource pressures. From the start, the NLA's approach to PANDORA was 'to do what [it] could with the limited resources available', which meant 'proceeding and working at a scale that permitted outcomes'. (Koerbin 2012: 4) As a result, a highly selective approach to PANDORA was chosen which was in-keeping to what the NLA could collect. While the approach has a high unit cost (estimated in 2005 as A\$179 or US\$169 at current exchange rates; taken from Cathro 2009: 24), the NLA has maintained that the benefits in quality control and accessibility outweigh the advantages of other web archiving approaches. Perhaps equally important, it has permitted the NLA to work around the lack of legal deposit laws for digital materials. At the same time, workflows and the infrastructure were developed in-house as there were no 'off-the-shelf' or open-source systems available. The NLA also employed 'bright' university students to undertake the technical development, which was the labour-cost



that the NLA could afford. The students were also more creative and enthusiastic, with low turn-over. Finally, the NLA adopted a collaborative approach early on, though this is limited to curatorial collaboration. (Koerbin 2012) Today, in the face of resource pressures, the NLA allocates priorities for action, and decisions for preservation selection are made based on the relative significance of the digital material and on the technical complexity of preserving access to them.

Another important economic consideration is the value of digital preservation, in particular, the value of the digital content that is being preserved, and in turn, the risk or cost associated with losing the data. In this fashion, digital preservation can be seen as an 'economic good' where an economic benefit to society can be gained with the future re-use of the information. (Blue Ribbon Task Force 2010, Harvey 2012, Bradley 2007). From the author's perspective, the value of the PANDORA web archive stems from fostering accountability and good governance of the government and other organizations, as well as serving as an important repository from a public policy perspective. As articulated by Cunningham and Phillips (2005), government agencies as well as educational and research institutions are producing large volumes of information in digital formats only, and if left uncollected are at great risk of being loss. Given PANDORA's focus to collect this type of material, it serves a very important role in safeguarding this information so that it may contribute to future public policy debate and development.

The NLA's current focus to redevelop its digital library infrastructure will likely mean that PANDORA will be replaced. The future of the NLA's web collections will likely include a greater emphasis on bulk thematic harvesting as well as adding value to the digital content collected. At present, the value added by PANDORA to the digital content is similar to that traditionally undertaken for non-digital material, namely, selection, cataloguing and quality control measures that facilitate their identification and retrieval. It also provides valuable services to publishers and others through the creation of persistent identifiers and its citation service. While this work is critical for ensuring access, one cannot help but feel that the potential of PANDORA is limited when it is solely functioning as a 'repository' or 'storage' for archived web material. Seen in this way, the author would support Margaret



Hedstrom's argument that digital preservation will add little value if it serves only as an alternative form of storage. Using the scholarly field as an example, she argues that users will demand digital materials that are 'easily retrievable, manipulated, transmittable and transportable' from the repository to other relevant sites for research and teaching. (Hedstrom 1997: 191) At present, PANDORA does not support any analytical functions of the digital data available in its repository. If PANDORA, and other NLA web collections, are to remain relevant they will need to support analytical processes and other forms of data manipulation. The costs for adopting such an approach should be more manageable as the NLA no longer finds itself as one of the few undertaking and studying the issue of digital preservation. There are a number of systems now in place and innovative approaches are being undertaken. The NLA will likely capitalize on these as it redefines its digital library infrastructure through its well-institutionalized approach to collaboration and information sharing with international, regional and local partners and stakeholders.

The analysis above highlights the important contribution that PANDORA has made to web archiving. Although, it may be in its last phase, it continues to serve as an important approach to web archiving with many lessons to be learned. It is an example of what can be achieved within limited budget parameters. At the same time, it illustrates the value of safeguarding digital materials for future generations; a conviction that the NLA was early to adopt and which hopefully continues to convince others. Its latest initiative to adopt *Preservation Intent Statements* serve as another important and pragmatic tool to help navigate the management of digital preservation, and will serve as an important beacon as the NLA adopts new operational systems for its digital preservation.

References

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010) *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. Retrieved 23, October 2013 from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.



Bradley, Kevin. (2007) Defining Digital Sustainability. *Library Trends*, 56(1) (pp. 148-163). Retrieved 23 October, 2013 from <https://www.ideals.illinois.edu/bitstream/handle/2142/3772/Bradley561.pdf>.

Cathro, Warwick. (2009) Digital library economics: international perspectives - I The Australian perspective. In Baker, D. and Evans W. *Digital library economics: an academic perspective* (pp. 119-130). Oxford: Chandos Publishing.

Cunningham, Adrian & Phillips, Margaret. (2005) Accountability and accessibility: ensuring the evidence of e-governance in Australia. *Aslib Proceedings: New Information Perspectives*, 57(4) (pp. 301-317). Retrieved 27, September 2013 from.

Greenstein, Daniel. (2000) Digital Libraries and Their Challenges. *Library Trends*, 49(2) (pp. 290-303). Retrieved 14, October 2013 from <http://search.proquest.com.lib.costello.pub.hb.se/docview/220452940/fulltextPDF?accountid=9670>.

Harvey, Ross. (2012) *Current Topics in Library and Information Practice: Preserving Digital Materials (2nd Auflage)* (Introduction, Chapters 2, 4, 10). Walter de Gruyter: Berlin/Boston. Retrieved 25 October, 2013 from <http://site.ebrary.com.lib.costello.pub.hb.se/lib/boras/docDetail.action?docID=10521727>.

Harvey, Ross. (2005) Case Study 2 – PANDORA (Preserving and Accessing Networked Documentary Resources of Australia). In Harvey, Ross. *Preserving Digital Materials* (pp. 203-208). K.G. Saur: Munich. Retrieved 23, October from <http://site.ebrary.com.lib.costello.pub.hb.se/lib/boras/docDetail.action?docID=10275865>.

Hedstrom, Margaret. (1997) Digital preservation: a time bomb for Digital Libraries. *Computers and the Humanities*, 31(3) (pp. 189-202). Retrieved 17 October 2013 <http://www.webcitation.org/5S7VwhqbB>.

Koerbin, Paul. (2012) *PANDORA past, present and future – national web archiving in Australia*. Retrieved 3 October, 2013 from



<http://www.nla.gov.au/content/pandora-past-present-and-future-national-web-archiving-in-australia>.

Lavoie, Brian F. (2008) The Fifth Blackbird: Some Thoughts on Economically Sustainable Digital Preservation. *D-Lib Magazine*, 14(3/4) (pp. 1-10). Retrieved 23, October 2013 from <http://www.dlib.org/dlib/march08/lavoie/03lavoie.html>.

Lavoie, Brian F., and Dempsey, Lorcan. (2004) Thirteen Ways of Looking at ... Digital Preservation. *D-Lib Magazine*, 10(7/8) (pp. 1-16). Retrieved 23, October 2013 from <http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>.

National Library of Australia. (2013a) *Digital Preservation Policy 4th Edition*. Retrieved 23 October, 2013 from <http://www.nla.gov.au/policy-and-planning/digital-preservation-policy>.

National Library of Australia. (2013b) *Broad directions for preserving the Library's digital collections*. Retrieved 23 October, 2013 from <http://www.nla.gov.au/content/broad-directions-for-preserving-the-library-s-digital-collections>.

National Library of Australia. (2013c) *Implementation Principles*. Retrieved 23 October, 2013 from <http://www.nla.gov.au/content/implimentation-principles>.

National Library of Australia. (2013d) *Preservation Intent – Selective Web Harvesting*. Retrieved 23 October, 2013 from <http://www.nla.gov.au/content/preservation-intent-selective-web-harvesting>.

PANDORA website: <http://pandora.nla.gov.au/>

Pitschmann, Louis A. (2001) *Building Sustainable Collections for Free Third-Party Web Resources*. Washington, D.C.: Digital Library Federation Council on Library and Information Resources. Retrieved 14, October 2013 from <http://www.webcitation.org/68uJ7PfyG>.

Webb, C., Pearson D., & Koerbin, P. (2013) *Oh, you wanted us to preserve that?!' Statements of preservation intent for the National Library of Australia's digital collection*. Retrieved 23 October, 2013 from <http://www.nla.gov.au/our->



publications/staff-papers/oh-you-wanted-us-to-preserve-that-statements-of
preservation-intent.

Witten I.H., Bainbridge D., Nichols D.M. (2010) *How to build a Digital Library*
(pp. 408-420). Burlington, MA: Morgan Kaufman.



APPENDIX 1

Preservation Intent - Selective Web Harvesting

NLA Digital collections: Statement of Preservation Intent

Collection Area: Selective Web Harvesting

The Library's selective web harvesting program, currently consists of the PANDORA Web Archive collection which contains a selective collection of web publications and websites relating to Australia and Australians. The PANDORA Web Archive was established by the National Library in 1996 and therefore contains historic online materials harvested from 1996 to the current period. Online materials (ranging from discrete publications to complete web sites) are selected for inclusion in the collection with the purpose of providing long-term and persistent access to them.

Web Archiving intends that:

- All PANDORA digital preservation masters including all associated metadata (currently known as the preservation master, the display master and the metadata master) should be retained in perpetuity. All technical properties should be maintained to the full extent possible.
- Content, connections and context are of primary importance. How it is ultimately presented to a user is a secondary consideration.
- The original harvested copy, that is the preservation master that represents the initial and untouched collection of files gathered by the harvest robot, is of less importance than the display master which is the copy that includes the results of quality assurance and curatorial work. However, as the implications of curatorial and QA work upon harvested instances may not be known, the preservation master, (although a lesser version in terms of completeness), should be retained at least at the bit level.
- The derivative copy derived from the display master and created for display and access should be maintained only for as long as useful; a new derivative version may be generated according to future access requirements.

The Library understands that web archives are problematic in that:

- The Library has no control over the creation of the original content and consequently its format, standards or quality and can only harvest what is delivered in a single published form through a browser/server request (e.g. the original data in the publishers databases are not collected);
- Current methods for collecting and rendering are also not ideal in ensuring the complete capture of all files or retaining full functionality in the version delivered from the archive;
- We are only taking artefactual snapshots of web content.



Therefore, the NLA accepts that what is to be preserved is not a mirror representation of the web, or even a website, but a snapshot of content that was once arranged and published as a website, with only limited functionality of the original. The archival artefact is formed out of the web collection process which is inevitably a lossy process in itself. Our concern is to define and control this loss. In addition, the way in which the content is collected and displayed may mean that there is significant limitation on the presentation of the archived artefact as an authentic record of the publisher's original data or even the version of that data as originally published on the web.

Other Preservation Issues:

- The intention is long term access for all users. However, over time access to certain content may only be available on site due to technical considerations.
- The harvested web content being complex objects are contained in either a compressed package (i.e. tarball) or a container file (i.e. WARC file). While the tarball retains the directory structure of the original harvested website, the WARC file may contain random collections of files plus metadata which are managed and located by indexes.
- The PANDORA collection, having begun in 1996 includes content collected through various methods and has acquired a legacy of inconsistency in respect to URIs, metadata and quality assurance interventions. Processes are underway to move the content to a consistent archival format (WARC) although the underlying legacy variations may not necessarily be removed in this process.
- The PANDORA collection can be broadly categorised as consisting of about 75% text, 20% images (JPEG, GIF, PNG) and 5% multimedia and style elements (Java script, CSS files, Flash etc.) including linkages. Because of the variable nature of the collected entity (the PANDORA title and its archival instances) ranging from simple documents to complex multi-file objects, there are some parts of the collection where style elements are more important, and some parts where this is less so. Style elements are problematic from the outset since they are sometimes difficult to harvest and remain often impossible to render (e.g. Heritrix wayback sometimes does not render some older content at all). Because content is harvested through a browser type request on a server in many cases only a subset of possible style element files are delivered (those required for the browser request). Moreover, harvesters are not able to thoroughly parse complex JavaScript which may also result in the collecting process not identifying and missing many style elements (JS, XML etc).
- Contemporary browsers are fairly tolerant for accessing both current and legacy web content. However, due to the variability of this content (collected from 1996 to the present day) and factors such as being poorly formed (no standards) can mean that viewing content as it was at the time of creation can be problematic.
- The status of the visual accuracy of the harvested copy of a site has not been kept in any systematic way (although implied in QA workflow). Thus, the look of the original may only be surmised from the content collected, the context of embedded links, tags and file types and the context of technologies known to exist at the time of harvesting.



- The Library's objective is not to misrepresent the material in any way that would compromise its legal warrant to collect, preserve and make accessible the archival content. Thus particular care in retaining the integrity of the intellectual content including embedded links and domain related image material is a priority.
- This collection is currently in a state of transition in how it is stored, described and understood via technical metadata.
- Below is an anecdotal list of file types that are known problems in the archive. Primarily the problems are first and foremost in how they content is managed in the archive rather than the file format itself:
 - RealMedia files – Files in this format are a main concern given that there would be quite a bit of this content in the archive, it took a lot of work to obtain and the files are generally significant parts of the sites. The archive includes a number of sites that include .rm and .ra files. Since these files were delivered in their live environment using a metafile (.ram) to point to the actual media file hosted on a RealMedia server, the NLA required the publishers to send us the .rm or .ra files which we then hosted on an RTSP enabled server. The continued support for this server at some point broke down – without notice to the business area – and consequently those archived sites with RealMedia content delivered in this way no longer function and it is uncertain if the files have been discarded.
 - VRML – This content required a plug-in (e.g LivePictureViewer) to function which does not seem to have been incorporated into browsers like other media players. This content will not function in the archive.
 - Shockwave – Older sites with Shockwave such as Director (.dcr) files do not work. Seems to be missing necessary plug-in components.
 - Quicktime VR (Virtual Reality) does not seem to work.

Version 1.0 1 March 2013

<http://www.nla.gov.au/content/preservation-intent-selective-web-harvesting>

