Report on the Crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005

Paul Koerbin

10 October 2005

1. Executive Summary

In June and July 2005 the National Library of Australia conducted a project to undertake the first large scale harvest of the Australian web domain. The objective was to collect, for preservation purposes, as much, in terms of breadth and depth, of the *.au* domain as possible within a defined and limited crawl duration and to gain more understanding of the issues associated with this approach to digital archiving. The Internet Archive, being the only agency experienced is such large scale web harvesting for preservation purposes, was contracted to undertake the crawl on behalf of the Library. The harvest crawl was run continuously for six weeks and captured more than 185 million documents, or files, the equivalent of 6.69 terabytes of data. This report is based on an initial analysis of the content by using the Internet Archive's WayBack Machine interface, pending receipt of the indexed content from the Internet Archive which will allow more detailed analysis. Limitations of the WayBack Machine interface means that this first stage analysis was somewhat constrained. However, it does suggest that the harvest was very successful particularly in regards to the breadth of the crawl. Particular issues identified and lessons learned are outlined in parts 9 and 10 of the report. The domain crawl provides a valuable benchmark and foundation for the National Library to develop future web archiving strategies and also provides the basis for the development of a significant collection of Australia's documentary heritage for future generations.

2. Background

In 1996 the National Library of Australia established the PANDORA Archive to selectively archive Australian web resources. Web resources (web sites, web publications) are included in the PANDORA Archive on a highly selective basis following the selection guidelines for online resources (available at: http://pandora.nla.gov.au/guidelines.html). As a result of this activity, as at August 2005, 9,315 individual titles had been archived. As some titles are scheduled to be archived periodically this equates to 18,928 archived instances. In terms of size (excluding multiple preservation copies) this amounts to more than 26 million files or 925 GB of content data acquired over a period of nine years.

Resources archived in PANDORA are quality checked for completeness and functionality and consequently represent the highest possible archiving standards given the technical limitations of the harvesting robots. Permission to archive a publication is sought from the publisher prior to harvesting and the archived resources are subsequently made accessible from the PANDORA Archive portal (although access restrictions may apply in some cases). The

PANDORA Archive therefore is a high quality and accessible archive of selected Australian web resources.

The National Library considers this approach to have been the most suitable approach to web archiving to achieve the best outcomes with the resources available. However, the Library has always recognised the limitations of the selective approach to web archiving. Selective archiving is labour intensive and the number of resources able to be archived represents only a very small proportion of Australian web publishing. The Library has therefore considered it important to attempt a large scale harvest of Australian web resources.

The opportunity to undertake and achieve the first large scale harvest of the Australian web domain arises from the available expertise of the Internet Archive (http://www.archive.org/), a not-for-profit organisation that has unique and extensive experience crawling and harvesting content from the entire World Wide Web since 1996. The Internet Archive (IA) has developed a crawl robot called Heritrix, which is specifically designed to undertake large scale crawling and harvesting for archival and preservation purposes. The National Library had a pre-existing collaborative association with the IA through common membership of the International Internet Preservation Consortium (IIPC) (see http://www.netpreserve.org).

3. Objectives

As National Library staff have limited practical understanding of large scale domain harvesting, the principal objectives of this project, therefore, were to:

- Test the ability to automatically crawl a large web domain based on non-exclusive seed lists and specified parameters (primarily the *.au* domain), including:
    - Analysing the success of the breadth of the crawl.
    - Analysing the success of depth of the crawl.
- Observe and analyse the performance of the Heritrix crawl robot.
- Analyse the extent to which restrictions on access to content limit automated harvesting.
- Obtain a clearer understanding of the composition of the Australian web domain.
- Obtain real content which could also be used as a test bed to determine approaches to delivering access to identifiable and specific content from a large scale harvest.
- Obtain a better understanding of the processes and costs involved in large scale domain harvesting.

4. Other Domain Crawl Projects

Large scale, automated domain crawls have been undertaken by a number of organisations for some years, notably in Sweden and Norway. The Royal Library of Sweden began crawling the Swedish domain in 1997 and completed their 12th crawl in January 2005. More recently a number of other institutions in countries such as Slovenia, Greece and Portugal have undertaken country domain crawls.

From the statistics available in regard to the domain crawling activities in other countries, the Australian domain crawl appears relatively large. For example the latest Swedish crawl amounted to around 46 million files (URLs) or 1.6 TB of data (about one quarter the size of the Australian domain crawl: see section 7 below). The entire bulk collection of the Swedish domain crawls from 1997 to 2005 is 306 million URLs or 10 TB; that is, less than twice the size of the Australian domain crawl. The Portuguese crawl in 2003 resulted in around 3.8 million files (URLs) or 78 GB. The second harvesting round for the National Library of Norway's Paradigma Project in August 2003 resulted in 4.1 million URLs.

Common issues that arise from recent publications on these activities include:

- The problem of characterising the web domain and translating this to configure automated harvesting.
- The automatic analysis and extraction of metadata from the archival content to support access.
- Rendering the archived content.

Close contact will be maintained with other institutions carrying out whole domain harvests to keep up with developments in addressing these issues and to exchange experience and learning. Ongoing involvement in the International Internet Preservation Consortium (IIPC) is one useful way of achieving this.

Some recent papers and documents of interest on these issues include:

- Drugeon, Thomas. "A technical approach for the French Web Legal Deposit". A paper presented at the 5th International Web Archiving Workshop, 2005. http://www.iwaw.net/05/drugeon.pdf
- Gomes, Daniel and Silva, Mario J.. "Characterizing a National Community Web". This refers the Portuguese web and is forthcoming in the *ACM Transactions on Internet Technology*.
- The Umich.edu Archives Project, http://si.umich.edu/mirror/. Papers and technical documents for this project include some interesting work on processing ARC files for access to Heritrix crawled content.

5. Agreement with the Internet Archive and Cost of the Project

5.1 Letter of Intent and Quotation of Costs

The agreement with the Internet Archive to undertake the crawl on behalf of the National Library of Australia was in the form of a Letter or Intent signed by the Director-General dated 12 May 2005.

The Letter of Intent includes general statements as to what the National Library sought the Internet Archive to provide including:

- The undertaking of a four week crawl of the whole Australian domain pursuant to technical, time and resource limitations;
- The provision of reports;
- The provision of a Wayback Machine style interface to the archived content;

- The building of a search index; and,
- Initial hosting of the collection content.

The Letter of Intent stated the expected costs based on the quotation provided by Michele Kimpton, Director Web Archive, Internet Archive. These quoted costs were:

- Set-up, crawl and indexing: XXX
- Hosting fee, per month: XXX

At the time that the Letter of Intent was sent, no decision had been made as to, if, how and when the archive content would be delivered to the National Library. However, the estimated costs for options for delivery of the content were stated as follows:

- Content on plug and play set of Petaboxes ('redboxes'): XXX to XXX; or
- Content on set of SATA or IDE drives: XXX to XXX.

5.2 Final Quotation from the Internet Archive for Cost of the Crawl and Delivery and Installation of the Harvested Content

Following the completion of the crawl the decision was made not to proceed with the option for the Internet Archive to host the content but rather that the content would be delivered to the National Library on Petaboxes as soon as possible. Prior to shipping the content, the IA would complete the full-text indexing of the content. These changes were agreeable to the IA and a revised quotation of the cost was provided by them (see the table below). The IA continued to allow access to the content via the Wayback Machine interface, at not additional cost, while the content was being indexed and prepared for shipping. The indexed content was delivered to the National Library in late November 2005.

	Estimate in USD provided by the Internet Archive	*Estimated AUD equivalent as at 13 Sept. 05*
Crawl (cost based on a four week continuous crawl)	XXX	XXX
Indexing	XXX	XXX
Delivery and Installation (including requisite hardware, middleware, software, and engineer to install)	XXX	XXX
Total	XXX	XXX

5.3 NLA Administration Cost

NLA Project Officer at EL1.3 level for 92 days	XXX

6. Technical Description of the Crawl

6.1 Crawl Duration

The crawl was planned to run for four (4) consecutive weeks beginning on 13 June. However, the Internet Archived (IA), at no additional cost, allowed the crawl to continue for two extra weeks. The completed crawl duration, therefore, was six (6) weeks from 13 June 2005 until 25 July 2005. Prior to the crawl proper starting, a test crawl was undertaken by the IA during the week of 6 June to 10 June. The test crawl period did not involve harvesting any content.

6.2 Crawl Specifications

A specification document was written to describe the scope and requirements of the crawl (Appendix 1). In summary, the specifications included the following:

- Breadth of crawl: the entire *.au* domain (or as much as could be crawled within the agreed crawl period).
- The use of an experimental and not fully implemented mechanism developed by the IA to undertake automatic DNS (Geo-location) lookup of non *.au* pages linked from the crawled pages to identify non *.au* domains on Australian IP addresses. Where these could be so identified, they would be included in the Australian domain harvest.
- Depth of crawl: all content of the identified websites that could be crawled was to be harvested.
- A limit of 100 MB per file was set. Single files greater in size than 100 MB were excluded from the harvest and a report on these files was provided.
- Files taking longer than 20 minutes to download were excluded.
- Robots.txt rules (exclusions) were obeyed.

6.3 Seed Lists

The main seed list used was that provided by the IA based on previous Alexa crawls of the Australian domain done as part of the broad Internet archiving work of Alexa and the IA. The last broad crawl done by Alexa (and available to the IA) was in November 2004 which yielded 342,296 *.au* hosts. These formed the bulk of the seed list and provided for the broad nature of the crawl.

In addition to the Alexa/IA derived seeds, a seed list of around 530 URIs was provided by the National Library to form the basis of the crawl bias. These URIs were predominantly government (Commonwealth and state) and higher education sector URIs. These seed URIs were given priority then filled out by the broader Alexa/IA list.

The IA identified from previous Alexa crawl logs a list of 230 sites with more than 50,000 pages (the sites listed actually ranged from 49,920 pages to 844,688 for the *www.ezydvd.com.au* site). This list was reviewed and around 40 were excluded from the crawl. The purpose of this was so as not to tie up crawler time with harvesting large commercial sites and so as not to compromise the bias given to the crawl towards government and higher education sites. A number of other sites from this list were identified and marked as low priority seeds.

6.4 Notifications and Target Server Issues

A Notice to Webmasters page explaining the crawl project and providing contact details was created and hosted on the PANDORA web site at: http://pandora.nla.gov.au/crawl.html

This crawl notification page could be identified by webmasters from their server logs so that if they identified crawl robot activity on their site that was causing problems for their server, or if they were otherwise curious, they could contact the NLA project officer and the IA crawl engineer immediately.

It should be noted that the Heritrix crawl robot was configured in such a way as to harvest content in a measured and adaptive pace so as to minimise any prospect of disrupting or otherwise compromising the normal activity of the crawled servers. As stated above, robots.txt exclusions were obeyed in all cases.

7. Summary Reports and Statistics of the Crawl Outcomes

7.1 Statistics for Size and Extent of the Crawl

Total number of hosts crawled:	811,523
Total documents (files) crawled:	189,824,119
Total unique documents (files) crawled:	185,549,662
Total raw data size in bytes:	7,360,187,145,622 (6.69 TB)
Total compressed ARC* file size:	4,964,818,275,410 (4.52 TB)
Total compressed DAT** file size:	90,888,523,022 (84.65 GB)
Total compressed data size:	5,055,715,039,125 (4.60 TB)

* ARC files are an archival file format used by Heritrix to store the archived content.
** DAT files are metadata files describing the content of the ARC files.

7.2 Statistics of Number of MIME Types

The domain crawl identified 976 MIME (Multipurpose Internet Mail Extension) types. Typically, as has been found in the experience of PANDORA Archive, a large number of these represent badly or incorrectly formed MIME type identification. Of the 976 MIME types reported, 836 are associated with less than 1000 files, 686 of those with less than 100 files and 476 with 10 or less. The top MIME types as reported are listed in the table below. More complex analysis would be required to obtain exact figures for the types of files (e.g. all image files or all audio files) since all 976 reported MIME types would have to be analysed and sorted.

MIME Type	No. of files	% of total
text/html	126,587,753	67%
image/jpeg	32,414,376	17%
image/gif	20,716,296	11%
application/pdf	3,071,252	1.6%
text/plain	1,521,619	0.8%
image/png	913,104	0.48%
text/css	808,571	0.42%
application/x-javascript	429,700	0.22%
application/msword	392,140	0.21%
application/x-shockwave	355,840	0.18%

7.3 Full List of Reports Provided by the Internet Archive:

A range or reports were provided by the IA at the completion of the crawl.
File exclusion reports were also provided after two weeks of the crawl to allow for files excluded on the basis of size and download time to be reviewed and included in the crawl if desired. At this stage some of the larger sites that had originally been excluded from the crawl – so as not to take up resources and allow the crawl to go as broadly as possible – were included.

- Crawl Report: size, number of files etc. (4 KB)
- MIME Type report (88 KB)
- Response Code Report (4 KB)
- Per Host Summary Report (59 MB)
- File Exclusion Report 1: files taking longer than 20 minutes to download (5.3 MB)
- File Exclusion Report 2: files discovered but not captured as they are larger than 100MB (799 KB)
- File Exclusion Report 3: files excluded due to robots.txt rules (3.4 GB)
- Report on hosts with outstanding URIs queued when crawl was stopped (960 KB)

8. Management of the Process Prior To and During the Crawl Period

8.1 Contacts Between the National Library and the Internet Archive

The Project Officer for the crawl period was Paul Koerbin, Digital Archiving Branch. The two principal contacts at the IA were Michele Kimpton, Director Web Archive, and Igor Ranitovic who was the crawl engineer for the project. Contact was mostly by email although four phone conferences were also held with Ms Kimpton. Some of these phone conferences also involved, variously, Monica Berko, Director Web Applications, Margaret Phillips, Director Digital Archiving Branch and Keith Jeffers, Director Business Systems Support. Initial discussions concerning the crawl were conducted on behalf of the NLA by Monica Berko with Ms Kimpton in person at meetings of the IIPC in 2004 and 2005. Pre-crawl email contact with Ms Kimpton was also undertaken by Margaret Phillips.

During the crawl period contact was generally with Mr Ranitovic who provided exceptional attention and immediate responses to questions and requests. Mr Ranitovic adjusted his own work hours to better align his working hours in San Francisco with business hours in Canberra. He also provided very helpful clarifications and explanations of technical aspects of the crawl mechanisms.

8.2 Contacts from the Public

During the crawl and test crawl period contacts from 11 webmasters were received. Of these contacts:

- Two could be characterised as complaints of an aggressive nature. One was merely abusive; and the other, which involved some ongoing correspondence, was from a commercial site that was entirely antipathetic to the objectives of the crawl.
- Two more contacts were straightforward requests to stop crawling;
- One webmaster initially requested a stop to crawling, but after the project was explained allowed the crawl to continue; three
- Three webmasters reported errors in their server logs and requested that the crawl agent stop;
- Three contacts could be characterised as curious webmasters noting activity on their logs and interested to know more. These did not result in the need for any action.

All complaints and requests to stop crawling were acted on in a very timely manner and all contacts were responded to and satisfactorily resolved. Beside the two aggressive complaints, the contacts were generally polite and even positive. Two comments from webmasters were: “your attention to this has been outstanding” and “thank you very much for your prompt response”. In two or three cases the webmasters were very helpful in identifying problems and providing information and detail about the activity of the crawl robot on their server. In at least once case this lead to Mr Ranitovic, the IA crawl engineer, identifying a minor technical bug in the Heritrix crawler.

8.3 Documentation

Correspondence with the Internet Archive is filed on TRIM file: NLA05/391
Correspondence relating to contact from the public is filed on TRIM file: NLA05/853

9. Analysis

9.1 Methodology

The analysis of the domain harvest content undertaken so far, involving a visual sampling of the content, and described in this report, should only be seen as a first phase analysis. It is envisaged that a second phase of analysis, which should involve IT staff, would adopt a more technical analysis of the content with the objective of informing the development of access options and methods. This second phase of analysis would need to be conducted once the content and indexes have been delivered to the Library, expected to be in late November 2005.

9.1.1 General Issues and Limitations

The principal method used for analysis was to access and examine the harvested content. While the Per Host Summary Report provides statistical figures on the number of files harvested per host, exactly what has been harvested and its functionality could only be determined from viewing the content.

This approach does however have inherent limitations as the only access currently available to the content is through the Wayback Machine interface provided by the Internet Archive. This interface only allows access by means of a specific URL. It also employs a method of rendering pages to the browser that uses JavaScript to append a *sWayBackCGI* string to re-write original links in the page to deliver archive content and adds a BASE tag to resolve relative links to deliver objects from the archive. In this way, links are not permanently rewritten in the source code as is the case in the PANDORA Archive display copies. However, as JavaScript code and binary resources are not rewritten, content rendered in the browser that may appear to be delivered from the archive may in fact be delivered from the live site.

Browsing is possible to some degree once an initial archived page is accessed, but with evident limitations. For example, content may have been captured for sites that have dynamically generated content, but the links may not function in the harvested version. In general, the Wayback Machine interface functions best when exact known URLs are sought. For these reasons it is not a straightforward task to make definitive conclusions in as to whether content has been harvested or not.

9.1.2 Methodology for Analysis of Breadth of Coverage

In analysing the content for the breadth of coverage, sites (or specific pages within sites) were identified in two ways:

1. Through known lists (portals) of web sites such as lists of government and educational sites. The portals themselves were accessed in the harvested content as a starting point for browsing to specific sites. In addition, specific sites identified from the live version

of the portal were identified and directly searched through the Wayback Interface. In cases where browsing from the harvested version of the portal did not work, pages on the live site were identified and searched for in the archived content. In most cases a site that appeared not to have been harvested when attempting to browse for it, could be located (or a least a sample page) when searching for specific pages.

2. By doing Google searches to locate random sites (or specific pages). This was done by limiting the Google search to Australian content using the *google.com.au* site, and then using random search words (names of places, people, topics) and also applying additional limiting conditions such as *-site:.au* (to limit to content not on a *.au* domain) and *site:.asn.au* (to limit to just sites on the *asn.au* domain).

It was also necessary in many cases, particularly in respect of checking the coverage of non *.au* sites and weblogs to check whether or not the site's IP address was located in Australia (using the tools on the *www.dnsstuff.com* site). This was so as to determine whether or not a site could be expected to have been harvested or not. Robots.txt files were also checked on a number of occasions to determine if robots exclusion rules were the reason for missing content.

9.1.3 Methodology for Analysis of Depth of Coverage

The methodology in regard to an analysis of the depth of the crawl is somewhat more problematic. By depth we mean the extent of the content archived, so that we can say the crawl has harvested all public content files and files that support the visual representation and functionality of the site such as stylesheets and JavaScript (JS) files.

There are major limitations in respect to drawing strong conclusions about the depth of the crawl. These include:

- Lack of statistics to compare the number of files for a live and archived site.
- Reliance on visual checking and the impracticality of analysing a large enough sample through this method.
- Reliance on visual checking when internal absolute links in archived pages have not been re-written, giving the impression of successful harvesting (which may or may not be the case and can only be confirmed by identifying individual files embedded in pages, stylesheets and JS files and searching for them through the Wayback Machine). The complex system the Wayback Machine employs to deliver content to the browser means that it is not possible to make definitive assessments as to whether the content has been archived from the visual checking alone.
- Inefficiency of, and problems with, browsing (that is, following links within) the archived content using the Wayback Machine interface.

The last point requires some further explanation. While it is possible to access a page via the Wayback Machine interface and begin browsing (or navigating) through a site (or indeed to other sites) from that starting point, this ability to browse regularly breaks down. However, in a number of cases a page that is not found from this browsing may be found by directly accessing the file.

These limitations work against obtaining quantitative evidence as to the performance of the crawl in terms of depth. However, the other aspect of content crawling performance analysed involved a qualitative approach to the depth and success of the crawl's performance in respect of known issues and problems and doing spot checking of pages within large sites. This involved checking some sites known to cause problems because of the existence of certain file formats or delivery mechanism, such as sites using Flash technology. This procedure also proved problematic in regard to being able to draw definite conclusions on performance, both for reasons already mentioned and because it was not possible to determine how content was located by the crawler (this last point is explained in more detail under the section 9.2.5 Problematic File Formats). This aspect of the analysis in particular needs more work in order to draw useful conclusions.

9.2 Comments on the Analysis in Relation to Specific Domains and Formats

9.2.1 Government and Education Domains Sites

Government and higher education sector websites were strongly represented in the priority seed lists supplied to the IA and so the coverage of all levels of government sites appears to be extensive. Many government and higher education sites are, however, large, consist of dynamically delivered content and utilise complex process to deliver style and functionality, consequently the depth and completeness of the harvesting cannot be assured. Browsing such sites in the archive is not always possible to deeper levels. This may represent a break in functionality in the version delivered from the archive or it may indicate content has not been harvested – both scenarios were identified in the analysis.

9.2.2 Associations Domain (asn.au) Site

A random sample of *asn.au* sites were checked. The purpose of this was to look at a specific (and identifiable) sub-domain of the *.au* domain that had not been specifically seeded by the NLA to assess the broadness of the coverage. A small sample of 40 sites was randomly selected using a Google search using the Google limit parameter *site:.asn.au* and the search term "Australia". Of the 40 randomly selected sites (or pages) from the Google result list all but 2 could be located in the domain harvest, giving a 95% success rate on this sample.

9.2.3 Non .au (i.e. .com) Sites

A similar approach to sampling non *.au* sites was taken. The purpose of sampling the coverage of non *.au* sites was to assess the success of the automatic DNS looking functionality trialled by the IA during the domain crawl. During the crawl links identified in crawled pages that were not on the *.au* domain were logged and passed through a GeoIP database to determine if the site was located in Australia. The IA stressed that this functionality was only at a trial stage and the automated mechanism only partially implemented in Heritrix and as such they could not guarantee the success of this mechanism. Again a small sample of 40 sites was selected using a Google Australia search limited to "pages from Australia" plus the condition *–site:.au* (to filter out *.au* domain sites) and the search term "web" to obtain a search result. As with the sample of *asn.au* sites only 2 of the 40 randomly selected sites from the Google results list could not be found in the domain harvest, giving a 95% success rate on this sample.

9.2.4 Weblogs (Blogs)

Weblogs (blogs) are a form of web publishing that is currently of great interest, however they do present a problem in respect of domain harvesting. For this reason some analysis was done of the coverage of blogs. In many respects blogs are like any other web site and do not generally provide major problems technically for archiving in the PANDORA Archive. The main difficulty with blogs is identifying them as Australian since, in many cases, this depends not on the geo-location of the server but on the content or identity of the author.

Some blogs of course do reside on *.au* domains and therefore could be expected to be harvested and some examples found indicate this is the case. Other blogs may be on non *.au* sites but be located on an Australian IP address and therefore have a chance of being picked up by means of the automated DNS look up functionality. Again, examples found in the domain harvest demonstrate this to be the case. However, many blogs by Australians or of Australian interest will be located on sites designed specifically to host blogs such as Blogger (http://www.blogspot.com/). Major popular blog hosting sites such as this Google owned site are located outside Australia and therefore "Australian" blogs hosted on such sites have not been harvested. Oddly, however, using a directory of blogs – Blogwise – to identify a sample of Australian blogs, a number of these blogs using non-Australian blog hosting sites were located in the domain harvest but with only the first page harvested.

9.2.5 Problematic File Formats

Some analysis was done of how well the harvest robot performed on file formats known to be problematic. The results of the analysis so far are inconclusive and this is an aspect of the analysis that requires further work. However, from the limited analysis so far possible indications are that Heritrix performed at least as well in regard to problematic file formats as HTTrack (the robot utilised by PANDAS for PANDORA archiving) and probably better.

Some examples known to be problematic when archiving for PANDORA were found to be similarly incomplete in the domain crawl. For example, like HTTrack, it did not successfully harvest a site with an *.ehtml* extension. Also, some sites with menu options governed by JavaScript files were no more successful in the domain crawl than when attempted with HTTrack.

There is some evidence that Heritrix may have performed better in some circumstances. By way of example, in relation to Flash sites, where Flash technology is used to deliver content and links to content, in some cases the whole domain harvest appears to perform much the same as HTTrack. That is, generally some content, but not necessarily all, can be harvested but cannot be successfully re-delivered from the archive (although for PANDORA Flash files may be edited with a HEX editor to achieve functionality). However, while analysis of such performance by Heritrix using only visual checking is problematic and inconclusive there is some evidence of a better performance by Heritrix than HTTrack. For example, in relation to the Bell Shakespeare Company site – a Flash enabled site – 465 files were harvested in the domain harvest, while a test harvest using HTTrack appeared to be able to harvest only around 6. This would seem to be an indication that Heritrix performed better. While it is not possible by means of visual checking to determine if this was due to the better ability of Heritrix to parse embedded links, the Internet Archive were able to confirm this to be the case in this example by reference to harvest logs using the Heritrix 'hoppath script'.

Nevertheless the scope of the harvesting done for the whole domain crawl and that done for PANDORA are not strictly comparable. Harvesting done for PANDORA using HTTrack is done on a site by site basis with specified parameters (gather filters) and generally using only a single seed URL, while that done for the whole domain harvest is not targeted or limited in the same way. Thus it is quite possible for a particular site in the whole domain harvest to have been identified by various referring URLs which could include referring URLs providing access to content that may not be able to be parsed in a crawl initiated from a single starting URL (as in the case of crawling done for PANDORA). This is an aspect of analysis that could be further explored once the content and the logs are available to the National Library.

9.2.6 Restricted Pages

From the HTTP response code report provided by the IA, 444,214 URLs reported a 401 response code. This is the 'unauthorized' response code which includes the server challenge response for user ID and password. This may only be an indicative figure but as such it represents around 0.234% of all URLs crawled.

10. Findings

10.1 Strengths

The whole domain harvest represents more than seven times the extent of the entire nine years of harvesting for the PANDORA Archive (as at July 2005 the number of files was around 26 million, compared with the whole domain harvest's 185 million unique files). The extent of the whole domain harvest also compares very strongly against the last Alexa (Internet Archive) crawl of the Australian domain done in November 2004, which constitutes only around 21 million files.

The crawl appears to have been very successful in the breadth of coverage of the *.au* domain. As many sites were accessed during the analysis period but not counted, an overall exact count cannot be given. However in respect to some of the specific types of sites looked at, including non *.au* sites and *asn.au* sites, a count was kept and in both cases 95% of the randomly selected sample URLs were located in the archive. While the samples in relation to the entire content were very small, it was random without any obvious bias. The sample is really too small and the visual analysis process too limited to draw any definite conclusions; however this figure may represent an optimistic estimate of overall coverage of the *.au* domain.

Another strength of the harvest is that a certain number of non *.au* hosts located on Australian IP addresses could be identified automatically and included in the harvest. While this does represent a good additional contribution to the harvested content, it must be assumed that the percentage of such host included in the harvest is relatively small.

Clearly, given the size of the harvested content considerable depth (as well as breadth) has been achieved. Some large sites give evidence of considerable depth. For example the ASIC site (a complex NSF site delivering content dynamically) with 11,857 files. However, some sites like the ABS site (also and NSF site) has only 5,641 files. The ABS site however has robots exclusion on /ausstats/ directories and this would seem to be at least one reason for the limited harvesting of content. Around 13,612 files from the *nla.gov.au* domain appear to have been archived.

It is worth noting that undertaking this harvest through the services of the IA and the technology they have developed, specifically the use of the Heritrix robot and the ARC file format means the content is in a form that is emerging as the standard preservation archival format and as such is compliant with IIPC objectives.

10.2 Weaknesses

One weakness of the harvest is a result of the inability to apply title specific gather filters. While the nature of this broad harvest does not require filters to limit crawling, filters are also used to facilitate harvesting to gather content, auxiliary files (such as style sheets and JavaScript files) and in some cases frameset files that the gather robot is unable to successfully identify and harvest. In other words, there is an over-riding dependence upon the crawl robot to be able to parse code and follow links.

The compliance with robots.txt exclusion rules also weakened the harvest outcomes. The robots.txt rules in many (probably most) cases assist the crawl robot, or at least do not compromise it, in identifying relevant content to harvest. However, robots.txt files are not specifically directed towards harvesting robots, but rather towards robots generally, including indexing robots. There is evidence from harvest that a number of webmasters exclude such things as image directories or even (as in the case of the Australian National University site) whole host sites that provide the stylesheets and style objects. Since the crawl robot cannot analyse the rules and determine which ones to obey and which ones not to obey, by complying with robots.txt exclusions content that would have been desirable has in some cases been excluded.

While the harvest was able to obtain a certain amount of content on non *.au* domain by doing ‘on-the-fly’ DNS lookup or parsed links, this identified putative Australian sites on the basis of IP location. The weakness of this harvest method is that sites with Australian content or authorship located outside Australia could not be automatically identified and included in the harvest.

10.3 Metrics Constraints

A constraint on the analysis of the domain harvest is that there is no existing useful benchmark or metric against which to assess it.. That is, there was no directly applicable metric to determine what percentage of the actual entire Australian domain the crawl represents. Nor are figures readily available to assess the completeness of the crawl in respect to individual sites. Consequently, it is not possible to make definite, statistics based conclusions as to the success or otherwise of the crawl.

The most readily available metrics for the Australian web domain are based either on domain name registrations or on IP allocation counting. For September 2005 *AusRegistry* reports 585,756 registered *.au* domains (see: http://www.ausregistry.com.au/pdf/PUBLIC-200509-print.pdf). This figure is smaller than the 811,523 domains crawled, suggesting a relatively large number of non *.au* domains were included in the crawl. In respect of IP metrics, the September 2005 *Whois Source* count of registered Australian IPs was 26,848,679 (see: http://www.whois.sc/internet-statistics/country-ip-counts.html). The *Internet Systems Consortium* also uses a count of IP addresses that have been assigned a name which, it should be noted, can include any number of addresses that do not actually exist. In order to get a truer figure, they ping a 1% sample of the total IP addresses to estimate the total ping-able hosts. The ISC figure for Australian hosts for July 2005 was estimated at 5,351,622 (see: http://www.isc.org/index.pl?/ops/ds/). This figure is considerably more than the 811,523 hosts crawled for the Australian domain crawl. However, the metrics are not directly compatible and only very cautious inferences should be drawn based on these simple figures. One is perhaps that there are many non *.au* Australian IP addresses that have not been identified and crawled.

10.4. Access Limitations

While it is not strictly a weakness of the crawl, the crawl method and re-delivery (display) mechanisms do provide limited access to the content. Access will be improved once the full-text indexing of the content is completed prior to delivery of the content to the National Library, since this will provide keyword access in addition to Wayback Machine exact URL access. Difficulties in relation to browsing the content will remain; however, this may be considered a limitation that may be able to be reduced by the subsequent development of access tools to the content.

11. Lessons Learned

While it is not possible to determine accurately how much of the Australian domain has been identified and harvested, it is seems reasonable to conclude that a very large percentage of it, particularly in the dimension of breadth, can be successfully crawled within six weeks when starting with a large seed list and utilising the Heritrix crawl robot.

Providing seed lists to bias the crawl supplemented by the large seed lists available from the IA, does seem to have been effective in directing the crawl to content we specifically wanted while also allowing for a broad crawl. To this end, more concentration on developing target seed lists could be an effective way to focus future harvests.

The coverage of this crawl is broad and was specified to focus on the identifiable *.au* domain and what other content that could automatically identified as being located on Australian web servers. In this initial project therefore it can be said that Australian domain is defined geographically. It is evident that this approach while broad and substantial and relevant in content, lack the dimension that takes into account the actual content of the sites, since much web content related to Australia or produced by Australians may be published on web servers outside Australia. This is perhaps most obvious in relation to weblogs, but is undoubtedly much broader especially, it may be assumed, on the internationally utilised *.com* and *.net* domains.

Obeying robots.txt exclusions caused unforeseen limitations in regard to harvesting publicly available content. The use by webmasters of robots.txt exclusions to direct indexing robots to content they want indexed while limiting unnecessary load on their servers can result in content such as images and other resources constituting the look and style of the web sites may not be harvested.

Reliance on a visual analysis of the content is more difficult and limited than may have been anticipated. This is mainly due to the limitations of access to the content through the Wayback Machine and the manner in which the content is rendered.

Engaging the Internet Archive to undertake the crawl and harvest was successful in regard to being efficient and providing an outcome in line with the National Library's specification. The IA were also demonstrated very good communication, support and advice. It should be noted that at the time of writing this report the full-text indexing has not been completed and content has yet to be delivered to the National Library.

12. Options for the Future

This domain crawl provides a benchmark and metrics for future large scale crawls and forms the establishment collection of an archival snapshot of the Australian web domain. This collection provides a good basis on which to build future snapshots. Such future snapshots could include:

- Periodic domain crawls of a similar specification which would provide a chronology of snapshots as well as providing metrics on the growth and composition the domain.
- Narrow domain crawls focusing on particular sub-domains (such as *.gov*, *.edu*, *.org*) which could be run on shorter duration or for a duration appropriate to achieve an exhaustive coverage.
- Broader domain crawls that recognise the Australian domain can be defined by dimensions other than the geographical location of the host server.

A collection of this substance provides a basis to advance the efficiency of web archiving in Australia. For example:

- The existence of this significant collection of Australians digital publishing heritage supports the need for amendments to the Copyright Act to allow public access to the content. In other words such a collection is no longer merely theoretical, but rather is actual and could usefully be made accessible to researchers and indeed the Australian public.
- It provides a real and substantial collection for the National Library to utilise to develop its future strategies for web archiving. For example, it provides a basis on which to consider whether reliance upon large scale, technology dependent harvesting can produce outcomes in accordance with the Library's strategic objectives; and where we

should direct our future efforts, for example in implementing necessary access tools to reveal and disseminate the web archive content.

To explore these options further and to define the Library's future strategies for web archiving, an ongoing commitment of resources from the Collection Management and Information Technology Divisions will be required. At the broad level, the following actions are required:

- Reinvigorate action to extend Legal Deposit to electronic publications.
- Conduct more detailed data analysis once the harvested content has been indexed with the view to gaining a better understanding of the issues associated with collecting and providing access
- Implement tools to provide public access to the harvested content once this is permitted
- Consider policy and process implications of domain harvesting for future development of the archive of Australian online resources.

13. Recommendations

It is recommended that a Working Group with representatives from Divisions 1 and 4 be established to develop a Work Plan that includes the following actions:

- Further and more thorough analysis of the content of the harvest when full-text indexing is complete and content has been delivered to the National Library.

- Investigation and, where appropriate, implementation of tools that will enhance access to the ARC file content.

- Using the existence of the domain harvest archive in the promotion of changes to the Legal Deposit provisions of the Copyright Act.

- Consideration of the legal, technical and organisational issues and costs of providing public access to the domain harvest content. This should include:

    - Investigating with PANDORA partners whether there are components of the content that we could legally provide access to now (e.g. Northern Territory publications and New South Wales government publications).
    - Investigating whether the National Library can provide limited onsite access from the Library's reading rooms.

- Defining metrics based on this domain harvest so it can be used as a benchmark for the refining, redefining and scheduling of further large scale domain harvesting.

- Investigating efficient and practical methods of identifying and defining the Australian web domain based on content and other non-geographical dimensions.

- Scheduling a second domain harvest for the second half of 2006.

Appendix A

NATIONAL LIBRARY OF AUSTRALIA
SPECIFICATION
for a
WHOLE AUSTRALIAN DOMAIN CRAWL
by
THE INTERNET ARCHIVE (CRAWLING SERVICES CONTRACTOR)

1. **SCOPE.** An extensive crawl of the publicly accessible Australian web domain. The Australian web domain is broadly defined as those resources belonging to the top level *.au* domain and other resources identified as being hosted within Australia.

2. **TASK DESCRIPTION**

   2.1. **Size of Collection:** It is not possible to give a completely accurate estimate of the size of collection. However, as an indicative figure, the latest (November 2004) Alexa crawl of the *.au* domain identified 342,296 hosts and 21 million URIs. The Crawling Services Contractor has estimated that four (4) weeks crawling would net in the order of 100 million documents at a size of up to three (3) to four (4) terabytes.

   2.2. **Collection Period:** Collection will be made for a period of four (4) weeks commencing on 12 June 2005.

   2.3. **Acquisition parameters:**

   2.3.1. **Breadth:** The entire *.au* domain or as much of it as can be crawled during the collection period. In addition to the *.au* domain the following to be crawled: re-directs from seeds to URIs outside the *.au* domain; and embedded resources, even if they are outside the *.au* domain. (The Crawling Services Contractor will also implement an experimental crawl process that may be able to identify linked resource located on an Australia IP addresses that do not have an *.au* domain.)

   2.3.2. **Depth:** Capture the complete site within existing technical constraints and with regard to other stated exclusions.

   2.3.3. **Filters:** Filters may be supplied to exclude certain URIs or otherwise refine the crawl.

   2.3.4. **File size limits:** A size limit of 100 Mb for an individual file should be set.

   2.4. **Website URLs:** A seed list of URLs will be provided separately. Additional seed URLs may be supplied during the crawl period up to one week prior to the crawl period completing.

   2.5. **Robots.txt:** The robots.txt and Robots META tag exclusions on crawled sites will be obeyed.

3. **NOTICE.** No notification services will be required from the Crawling Services Contractor. The National Library of Australia will conduct all notifications and permissions. Any inquiries from crawled sites directed to the Crawling Services Contractor should be redirected to the project officer at the National Library at <pkoerbin@nla.gov.au>. The Crawl Notification (Notice to Webmasters) page is located at on the PANDORA Archive web site at: http://pandora.nla.gov.au/crawl.html

4. **ACQUISITION PROCESS.**

   4.1. **Crawl Time:** The Crawling Services Contractor will notify the National Library project officer when the crawl has commenced.

   4.2. **Seed File Exchange:** The National Library will provide seed lists to the Crawling Services Contractor which will be used to prioritise (add bias to) the crawl. These seed lists will be supplemented by the list derived from previous Alexa broad crawls of the Australian domain that are in the possession of the Crawling Services Contractor.

   4.3. **Spot Checking:** The crawl engineer will perform basic spot checking of the crawl and take prompt action to resolve any large scale issues. The crawl engineer may contact the National Library project officer if a decision is required in relation to the crawling of specific URIs.

   4.4. **Corrective Actions:** The Crawling Services Contractor shall take prompt action to investigate any problem identified by the National Library; and, within technical constraints, shall rectify the problem and, whatever the resolution, notify the Library of status of the problem as soon as it is known.

5. **METADATA.** For digital objects retrieved from each URL during the crawl, the Crawling services contractor shall generate metadata. Metadata shall include though not be restricted to: URI retrieved; time and date of capture; HTTP status code; size of object in bytes; referring URI (except for seed files); MIME type; and HTTP response headers.

6. **ACCESS AND REPORTS FOR PROCESSING QUALITY ASSURANCE**

   6.1. **Crawl Report:** A report including the number of seeds crawled, total unique documents captured and total URLs captured.

   6.2. **Host Report:** A report listing by host the total number of URLs and total size of data captured.

   6.3. **MIME type report:** A report listing all the MIME types harvested and the number of objects for each type.

6.4. **Response Code Report:** A report listing the number of URLs returned for each response code.

6.5. **Robots Report:** A report listing all URLs excluded due to the robots.txt exclusion policy.

6.6. **File Exception Report:** A report listing all files discovered but not captured, that are over 100mb.

6.7. **File Exclusion Report:** A report listing files taking longer than 20 minutes to download.

6.8. **Quality Review Access:** The Crawling Services Contractor will provide a Wayback Machine style access commencing after the second week of crawling to the Library for quality review processing of the completed archived web sites. Standard reports will be produced at this time for the completed sites.

6.9. **Complete Technical Parameters:** the Crawling Services Contractor will provide the National Library with the (Heritrix) specifications used for the crawl.

7. **INDEX.** The Crawling Services Contractor will compile the first Wayback Machine index of the collection (that is, those sites that have been completed) after the second week of crawling and a full index after the completion of the entire crawl. (This is not a search index; see section 8.)

8. **SEARCH AND RETRIEVAL INTERFACE.** The Crawling Services Contractor will build a search engine for the completed collection. This index to be built for a one-off set-up plus hardware cost as stated in the Letter of Intent. This index will include at least html and PDF documents (and Word documents if possible).

9. **STORAGE/HOSTING/DELIVERY OPTIONS.**

9.1. **Preferred initial option and cost:** The IA will store the archive collection and build a URI request interface. This interface will be restricted to access by those authorised by the National Library of Australia. This agreement to be on a month-to-month basis with the option to take up another delivery option.

9.2. **Other delivery options:**

9.2.1. The archive contents to be shipped on a set of SATA or IDE drives to the National Library.

9.2.2. The archive contents to be delivered with a plug-and-play rack set of Capricorn Technologies Petabox “red boxes”

9.2.3. The IA to host the archive, including setting up a Wayback Machine style access for the archive. This option would require a minimum 12 month agreement.

10. **PRICING.** The pricing for the requirements detailed in this specification as stated in the National Library of Australia's Letter of Intent to the Internet Archived dated 12 May 2005.

11. **CONTACT:** The National Library of Australia contact for this project is:

    Paul Koerbin
    Digital Archiving Branch
    National Library of Australia
    Canberra ACT 2600
    61-2-62621411
    pkoerbin@nla.gov.au