Title: Separating the wheat from the chaff: Identifying key elements in the NLA .au domain harvest
Author: G. Fellows, R. Harvey, A. Lloyd, R. Pymm and J. Wallis
Author Address: gfellows@csu.edu.au; rossharvey@csu.edu.au; anlloyd@csu.edu.au
Conference Title: Fourth Research Applications in Library and Information Studies Seminar
Conference Name: Research Applications in Library and Information Studies Seminar
Year of Conference: Conference Location: State Library of Victoria, Melbourne, Australia
Editor: URL:
http://www.rmit.edu.au/browse;ID=9n0rxvwez2l1;STATUS=A;PAGE_AUTHOR=Lauren%20Furlong;SECTION=4;
Abstract: Separating the wheat from the chaff: identifying key elements in the NLA .au domain harvest
The National Library of Australia (NLA) carried out two whole domain web harvests in 2005 and 2006 which complement the selective web archiving approach taken by PANDORA. These whole domain harvests use crawler software to gather all web sites within a specified domain: in this case .au sites, or sites containing significant Australian content. Web harvests of this size (many terabytes) pose significant challenges, such as stakeholder perceptions and the provision of legal constraints to access. Despite these challenges, such harvests present fascinating research opportunities, allowing for example the possibility of investigating the extent to which Australian society has begun to engage with the online environment, or exploring the online presence and representation of social issues on the Australian public agenda (such as indigenous issues, extremist politics, religious fundamentalism, terrorism, pornography). The NLA has provided Charles Sturt University’s POA (Preservation for Ongoing Accessibility) research group with access to these web harvests and associated keyword indexes. This group is investigating three broad questions in relation to the NLA’s web harvest: 1) What are the contents of the harvest? 2) How can access be provided to this content? 3) What is the value of the domain harvests in relation to the NLA’s overall web preservation interests? This paper will describe the datasets and the harvesting process, briefly compare the 2005 and 2006 harvests, and describe some methodologies the POA research group is considering. It will use the example of blogs to investigate the extent to which Australian society has embraced the online environment as public sphere. Specifically, it will address how to identify blogs within the harvests, how to estimate the success of the harvest in gathering a relevant cross-section of Australian blogs, and the extent to which those captured represent a broadening of voices within the Australian mediascape.
CSU ID: CSU285401
Separating the wheat from the chaff: 
Identifying key elements in the NLA .au domain harvest

Geoff Fellows, Ross Harvey, Annemaree Lloyd, Bob Pymm, Jake Wallis 
(School of Information Studies, Charles Sturt University)

[Corresponding author: Bob Pymm, School of Information Studies, Locked Bag 675, Wagga Wagga NSW 2650. Telephone: 02 6933 2467 Email: rpymm@csu.edu.au]

Abstract

The National Library of Australia (NLA) carried out two whole domain web harvests in 2005 and 2006 which complement the selective web archiving approach taken by PANDORA. These whole domain harvests use crawler software to gather all web sites within a specified domain: in this case .au sites, or sites containing significant Australian content. Web harvests of this size (many terabytes) pose significant challenges to their use, such as the provision of, and legal constraints to, access. Despite these challenges, such harvests present fascinating research opportunities, allowing for example the possibility of investigating the extent to which Australian society has begun to engage with the online environment, or exploring the online presence and representation of social issues on the Australian public agenda (such as indigenous issues, extremist politics, religious fundamentalism, terrorism, pornography).

The NLA has provided Charles Sturt University’s POA (Preservation for Ongoing Accessibility) research group with access to these web harvests and associated keyword indexes. This group is investigating three broad questions in relation to the NLA’s web harvest:

1) What are the contents of the harvests?
2) How can access be provided to this content?
3) What is the value of the domain harvests in relation to the NLA’s overall web preservation interests?

This paper describes the 2006 harvest and uses the example of blogs to address how to identify material within the harvest and determine issues that need further investigation.

PANDORA

Australia established one of the world’s first web archiving programs when it introduced PANDORA in 1996. PANDORA has been very successful in its aim of involving a range of institutions around the country in identifying specific Australian web sites for retention and capture. The PANDORA archiving procedure entails going through the relatively complex business of obtaining permissions to capture the site, followed by the actual archiving process itself using the PANDAS software. This has resulted in a well-constructed archive, publicly available, of over 1.7 TB of data.
comprising over 32,000 archived sites as at September 2007 (National Library of Australia 2007). However, as was acknowledged by many of the participants in the PANDORA project there are concerns regarding the small numbers of sites actually being acquired through this process (Pymm and Lloyd 2007, p.175).

Selecting web sites for archiving

The PANDORA model of careful selection and archiving is one approach to preserving web sites. Other approaches include topic or subject archiving, for example, the Digital Archive of Chinese Studies, or domain-centred archiving, based upon content location (Masanes 2005, p.75). This approach, which is being used in a number of countries around the world, focuses on acquiring the contents of an entire domain (in Australia, the ‘.au’ domain) in order to build the archive using a completely automated approach. This raises issues of the quality of the data captured, its coverage, and importantly, the absence of permission to archive the resultant files. However, it does provide a relatively straightforward method of acquiring very large amounts of data – millions of files – with modest effort, which can provide a comprehensive snapshot of a country’s web presence at a particular point in time.

The Internet Archive (IA) has been taking this approach at an even broader level, looking at the entire web, and has been endeavouring to capture ‘the web’ and archive its contents, again since 1996. Using their Heritrix software the IA has been taking a broad approach to capture, archiving any sites the crawler can access. This has resulted in an archive reported to hold 30 billion web pages in 2004 (Tyburski 2004), growing at the rate of around 20 TB/month according to Wikipedia. The IA makes it clear that while it captures and stores every site it can, and makes it publicly available, it respects the rights of copyright holders to ask for the removal or blocking of their site from the Archive. Links to an exclusion policy can be found on their site and this explains how rights holders can contact the IA if they wish to seek removal or blocking of their site (Internet Archive 2007).

With the success of the Internet Archive, large scale harvesting has increasingly been seen as a viable option, despite the costs and concerns over the reliability of the crawl. Thus for the first whole of domain crawl undertaken in 2005 by the National Library of Australia’s (NLA), Phillips (2005) noted that:

40 percent of harvested titles could be incomplete or defective in some way. Nationally significant material is likely to be missing, and the archive administration will not be aware of it. (p.62)

Despite this, a number of country-based domain harvests had been undertaken, with European countries including Sweden, Norway, and Austria leading the way (Phillips 2005, p.59) and the NLA conducting a further whole-of-domain harvest in 2006, with another scheduled for 2007.
The .au harvests

The first whole of the ‘.au’ domain harvest was conducted by the National Library of Australia in June-July 2005. The Internet Archive were contracted to undertake the crawl over a four week period. They were provided with a set of specifications outlining the breadth and depth of the crawl, including rules to control harvesting and a seed list upon which the crawl was based. It ran for four weeks before being halted, resulting in 6.69 TB of data comprising over 185 million unique files being harvested. These results were then analysed at the NLA, although they cautioned in their report against the difficulty of drawing strong conclusions about the depth of the crawl due to a range of reasons, including an uncertain knowledge of what actually does comprise the totality of the ‘.au’ domain (Koerbin 2005, p.14). Yet the number of registered ‘.au’ domain names was significantly less than the number of domains actually crawled (585,000 compared to 811,000 crawled) indicating that the crawl included a large number of non-Australian-registered sites (Koerbin 2005, p15). This is an area requiring further analysis, and points to one of the big problems associated with examining such information – the sheer volume of data presented, requiring sophisticated programming and considerable computing power in order to obtain a better grasp of the nature and scope of the harvest’s contents.

The extent of the harvest was affected by a range of issues, including:

- the rules delivered to the Internet Archive: these included recognition of the robots.txt file, which seeks to limit web crawlers from accessing particular files or directories on a website;
- certain file types which proved difficult for the search engine to properly handle;
- password-protected sites;
- difficulty in accessing the ‘deep’ web, for example, dynamically-generated pages resulting from an enquiry; and
- the use of CAPTCHA systems on web sites designed to restrict access to humans and reject automated crawler software.

This first Australian harvest was followed by a similar one undertaken a year later during August-September 2006. Building on the experience of the first harvest, the specifications and seed lists supplied to the Internet Archive were more sophisticated. This resulted in a crawl lasting five weeks capturing 596 million files, over 19 TB of data – a massive increase over the 2005 harvest. Analysis of this crawl is currently underway and this paper forms part of that analysis process.

Web harvests of this size pose significant challenges to their use, such as the provision of, and legal constraints to, access. Despite these challenges, such harvests present fascinating research opportunities, allowing for example the possibility of investigating the extent to which Australian society has begun to engage with the online environment, or exploring the online presence and representation of social issues on the Australian public agenda (such as indigenous issues, extremist politics, religious fundamentalism, terrorism, or pornography).

The NLA has provided Charles Sturt University’s POA (Preservation for Ongoing Accessibility) research group with access to these web harvests and associated...
keyword indexes. This group is investigating three broad questions in relation to the NLA’s web harvest:

1) What are the contents of the harvests?
2) How can access be provided to this content?
3) What is the value of the domain harvests in relation to the NLA’s overall web preservation interests?

The rest of the paper focuses on blogs to illustrate some of the issues that the POA research group are encountering. It describes the 2006 harvest and uses the example of blogs to address how to identify types of material within the harvest and determine issues that need further investigation.

The .au harvest: technical issues

The .au harvests pose technical challenges, some of which are exemplified in a description of the process of attempting to identify blogs in them.

The Internet Archive supplied the 2006 web site harvest archive data on twenty ordinary PCs housed in red boxes, each about half the height of a desktop PC, in a tower rack. The data were collected in the US, transferred to the computers in the rack, and the rack with computers installed was delivered to the NLA in Canberra. All computers run the GNU/Linux operating system. One of the computers in the rack (the home server) was connected to the NLA network and as well connected to all the other servers. This home server contains the metadata, including indexes to the archive data housed on the other servers.

Initial analysis of the 2006 archive was done using the 146GB metadata file stored on the home server. Each line of this plain-text file has metadata for one web object in the archive. The metadata fields of interest for this paper are:

- the Uniform Resource Locator (URL);
- date and time of web object capture (ISO 8601 format);
- IP address of the web server host; and
- MIME media type of the web object.

The URL is the unique address commonly used to access web documents. For example, typing the URL ‘http://pandora.nla.gov.au’ will access the main page of the NLA’s PANDORA project.

Originally email was designed for simple text messages. When people decided they wanted to send and receive richer content, such as images and sounds, a method of translating the data of these media files into text had to be added to the basic mail system. The standard for this translation is called the Multipurpose Internet Mail Extensions (MIME). A MIME media type consists of nine content types: application, audio, example, image, message, model, multipart, text and video. Each content type has many subtypes, such as html or jpeg. Normal HTML documents should have the MIME media type ‘text/html’.
Using Linux text processing tools (described at TLDP n.d.) on the home server, many approximate statistics for the 2006 harvest were determined. The metadata file consists of about 620 million lines. There are around one thousand different MIME media types found in the metadata file, with most of these not found in the MIME media type registry (IANA 2007). Of objects with image MIME media types there were 81 million JPEG, 33 million GIF, two million PNG files. Of the non-standard images 480 thousand Progressive JPEG, two hundred thousand JPG and one hundred thousand BMP image files were found.

There were also seven million Portable Document Format (PDF) files generally meant for printing on paper and about three million simple plain-text files. See Table 1 for the exact figures.

Table 1: MIME media types represented in the 2006 .au harvest

<table>
<thead>
<tr>
<th>MIME media type</th>
<th>Number of items</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text/html</td>
<td>483,522,308</td>
<td>77.8</td>
</tr>
<tr>
<td>Joint Photographic Expert Group (JPEG)</td>
<td>80,886,507</td>
<td>13.0</td>
</tr>
<tr>
<td>Graphic Image Format (GIF)</td>
<td>33,016,258</td>
<td>5.3</td>
</tr>
<tr>
<td>Portable Network Graphic (PNG)</td>
<td>2,765,353</td>
<td>0.4</td>
</tr>
<tr>
<td>Progressive JPEG</td>
<td>489,344</td>
<td>0.1</td>
</tr>
<tr>
<td>JPG</td>
<td>190,389</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>BMP</td>
<td>99,447</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Portable Document Format (PDF)</td>
<td>7,356,818</td>
<td>1.2</td>
</tr>
<tr>
<td>Plain-text</td>
<td>2,960,303</td>
<td>0.5</td>
</tr>
</tbody>
</table>

By selecting the MIME media type field from each line in the metadata file, sorting the result, reducing repeated lines to unique lines and counts of repeats, then sorting by the count (largest first), the following were determined:

- The most common web objects (about 480 million) were those with MIME media type text/html. Typically a web document will contain HTML, which results in a web browser displaying text and graphics and links to other web documents. The web server sends each of these items, along with their MIME media type. While items such as images or movies end up being pictures on a screen and sound through loudspeakers, the data describing the graphics is encoded using several different methods. Because storing and transferring data costs money, many methods for encoding graphics have been invented. On the World Wide Web one of the most common forms of encoding for images is the Joint Photographic Expert Group (JPEG) standard. This standard aims to significantly reduce storage size without comprising human perception of the image. Other formats include Graphic Image Format (GIF), Portable Network Graphic (PNG). BitMaP (BMP) is the image format common in Microsoft Windows and supported in the web browser product Internet Explorer provided in Windows.

- Dynamic web pages are produced mostly using JavaScript (ECMA 1999). About 1.2 million objects with MIME media types related to JavaScript were
found, though often JavaScript is embedded in HTML documents. Very few web objects related to Java were found. Crockford (2001) provides an excellent discussion of the differences between JavaScript and Java; in essence, most of the new dynamic features of web pages often called Web 2.0 are done using JavaScript).

- Using the URL field from the metadata file, a search for the string ‘blog’ was made. The URL field was selected from each line and the web server name extracted. The result was sorted and reduced to unique lines with counts of copies. These were sorted by the counts, with largest first. About 1200 separate blog sites were found, although on examination the URLs came from many other countries as well as .au sites. In Australia many internet names, but not all, end in .au. This means it is difficult to tell if a server is definitely sited in Australia, even if it has Australian content. Geographic DNS names end in a country code (such as .au for Australia) with US organisation names omitting the country code, with only the type code at the end, eg. .com, .mil, .org, .edu, .net and so on. For the 1200 blog site URLs identified in the 2006 NLA web harvest, many came from non .au sites, confusing the issue. During the gathering process, The Internet Archive used the server’s IP address and geo-location lookup (Koerbin 2005 p5) to determine if a site was based in Australia so it is not clear as to the reason why a number of non-Australian sites were acquired – again, further investigation is required.

- The counts of web objects for each internet name containing the word ‘blog’ yielded some of significant size, eg. The Visual Lab of the University of Sydney (blogs.vislab.usyd.edu.au) which had the most objects, around 55,000. Confusingly, another large site, with about 27,000 objects, had a Tuvalu domain name. This was followed by bloghost.pixelsquare.com.au with about 25,000.

Each of the text processing jobs above took about ninety minutes to run as batch jobs. These were done using the harvest metadata file. Another possible avenue is to include the actual web data as part of the processing for more detailed analysis. Taking this approach however, would increase the processing time significantly. For example, if a process took one second per web page the total elapsed time would be 420 million seconds to go through the entire harvest, about 13 years. So either more efficient algorithms need to be developed or a statistical sampling approach taken.

**Blogs and their value**

A blog (short for weblog) is an online publishing tool which is low-demand in terms of the technological capability required to set it up. The blog provides a self-contained publishing environment in which individuals (or groups) can publish content (predominantly text-based, but gradually converging with multimedia formats) on a chronological basis. This content generally resides online and access can be public or controlled.
Farkas (2007, pp.12-14) argues that we can define blogs more through their structural components than their content. These components are:

- Archives – listings of past blog entries are usually organized by date, although some software also lets readers browse by category;
- Dated entries – all entries are dated and usually also have a time stamp;
- Permalinks – links that let others link directly to each post;
- Reverse-chronological postings – posting are arranged in reverse chronological order; and
- Two- or three-column format – most blogs have a large column for blog posts and one or two columns on the sides of the screen for additional information about the blog or its author.

And in most but not all blogs:

- Categories – archiving blog posts under certain categories makes it easier to retrieve them at a later date;
- Comments – users can post comments to specific blog posts;
- Search functionality – many bloggers today let readers search their archives for specific terms; this functionality is built into most blog software; and
- TrackBacks – these are messages sent from one server to another, letting a blogger know when someone else has referenced one of their blog posts.

Blogging has become a popular online activity. Uses of the blogging medium vary across a wide spectrum, from delivering personal journals to corporate news. Australia is no stranger to this trend. With a wide diaspora overseas and relatively widespread and affordable Internet connectivity, blogging has become an effective mechanism for maintaining networked relationships across time and over distance. The NLA harvest endeavours to capture a representative sample of Australia’s blogging output over the two periods of the harvest, but there are many issues associated with identifying whether blogs are Australian. There are few clues that the harvest can pick up on in terms of identifying an Australian blog – often these will reside on distributed servers, housed for example in the US (that is, not within the .au domain) so capturing this content is not necessarily a natural consequence of the harvest.

Blogs create a social space in an environment of temporal and spatial ambiguity; creating communities without propinquity across a distributed global culture. The network facilitates relationships which:

- Are issue-focused – for example, Greenfoot (http://www.greenfoot.com.au/) on environmentally sustainable lifestyles;
- Relate to communicative action around both consent and resistance, and inclusion and marginalization, which may be created at the moment of need – for example, GetUp (http://www.getup.org.au/blogs/) on grass-roots community advocacy and action around issues such as indigenous health;
- Facilitates social networking around communities with specific needs and interest – such as Ask Sam (http://blogs.smh.com.au/lifestyle/asksam/) on dating etiquette; or
• Are expressions of community values or points for agreement or disagreement, and are therefore inherently democratic – for example, Australia’s Migrant Heritage (http://blogs.smh.com.au/australiasmigrantheritage/) on Australia’s migrant history and multicultural future.

These examples highlight the potential of the blogging to provide multiple voices and multiple points of view in an arena which is open to scrutiny. The ease of publishing allows for the expression of a wide range of perspectives and opinions. In this way blogging provides a holistic representation of the diversity of Australian attitudes and culture. Ultimately, blogging represents a snapshot into the public sphere of democratic thought at a particular point in time.

With the benefit of two datasets to work from, there is potential to address the evolution of Australian digital culture as reflected through a sample of blogs gathered from the NLA harvest. This leads us to consider both

• the potential of web harvests to capture community as it forms and interacts, for example, through engaged commentary; and
• this approach as foundational in conceptualising a framework for digital anthropology.

Blogs represent a new and important collecting point for digital cultural heritage, in that they provide a snapshot of society at a particular point in time. The democratic social commentary that blogging represents creates a social dynamic not easily captured in current preservation selection strategies (Philips 2005, Pymm and Lloyd 2007). This offers new challenges and issues which will have implications for the digital preservation community to consider.

Preserving blogs: selection and other issues

A web harvest such as the .au domain harvest is, of course, part of a preservation plan for web sites. Preservation of blogs, whether part of a domain harvest or not, presents specific problems. In 2004 Entlich posed the question ‘Are any special efforts being made to preserve their contents?’ (Entlich 2004). Blog content, he contended, was lost as a consequence of many factors: sudden decisions by authors to stop blogging, or hosting services ceasing activities or terminating their agreement with blog authors. He suggests that as blogs become part of mainstream publications (examples used are the New York Times and the Columbia Journalism Review) they will be selected for preservation because of that association, but this still leaves the majority of blogs to be considered for preservation. Identifying blogs by automated means poses challenges, as has been indicated earlier in this paper. Entlich describes the status of Web archiving activities in 2004 as ‘still in their infancy’ and suggested that few librarians and archivists had identified blogs as meriting preservation. He concludes: ‘It's pretty easy to make a case for selective archiving of blogs. They represent a recognizably distinct form of communication that is having a measurable impact on human affairs in the early 21st century’. He also notes that this is ‘not yet getting much attention from librarians and archivists’. (The NLA is, perhaps, an exception: although its selective approach to web collection for PANDORA excludes blogs, an
exception is made for ‘those that support tertiary institutions publications category’ (Phillips 2005, p.67).

Perhaps the most important consideration when archiving blogs is the need to also keep reader commentary and linked material. Commentators, suggests Entlich, in effect ‘acquire a kind of ‘contributor’ status on the blogs they frequent’, so ‘full and complete capture of a blog’ is the aim. Additionally, blogs often contain links to other material and are an integral part of the blog. The problems of maintaining links and of deciding the extent to which linked material is also captured is well understood (Brown 2006 gives examples).

Recognition of the need to archive and preserve blogs is not just limited to the traditional custodians of heritage materials. The e-science/e-research community is also actively discussing the issues as traditional scholarly publishing activities change. For example, Murray-Rust’s blog entry ‘Blogs as scholarly record? Should we reposit them?’ (Murray-Rust 2006) suggests that some blogs may need preservation because they are ‘becoming the grey literature of our time’. One response notes the issues of using existing systems which are developed for the preservation of static documents for dynamic material, of which blogs are an example. Another raises the question of which blogs should be archived and notes the importance of ‘veracity of metadata’ for those which are deposited. Another e-science view is in Miller’s posting ‘What do you think about blog preservation?’ on the panlibus blog (Miller 2007). He notes the changing way in which scholarly publishing is changing, from the traditional scholarly paper being the goal of the work to a ‘far more fluid and collaborative’ research process where peer review was engaged at various stages of the research, rather than only at the end of the process with the published paper. In this new world, blogs become one of a range of mechanisms that assume importance.

Researchers from the University of North Carolina at Chapel Hill's School of Information and Library Science have recently concluded a survey of bloggers about their attitudes towards digital preservation of blogs (Hank 2007). The aim is to inform development of recommendations for impacting stewardship of weblogs at the level of creation, and the development of strategies for capturing the content of blogs for perpetuity’. Preliminary results (from 215 responses) are starting to appear on the blog established for this research project (Blogger Perceptions 2007). Some of the preliminary conclusions that are of relevance to this discussion are:

1. **Blogs are highly linked**: 89% of bloggers use links to other blogs
2. **Blogs are very dynamic**: ‘Nearly 99% of bloggers report editing blog posts after initial posting (often, sometimes, or rarely), and about 2 out of 5 (41.3%) report deleting blog posts after publication’; ‘80% of respondents have changed the look of their blogs on purpose, with 20% doing so 2 to 3 times per year and 11% changing appearance every few months’
3. **Blogs can be deliberately deleted**: ‘about 23% have deleted an entire blog, 87% of these reporting intentional deletion’
4. **Links to blogs are important to preserve**: elements identified as import to capture included linked content (13.8%), imported images – e.g., flickr (10.5%), imported tags – e.g., del.icio.us (10.5%), RSS feeds (10.5%).

Taken together, these preliminary results paint a picture of a highly fluid genre that
poses serious challenges to conventional web archiving practice.

Providing access to blogs in the .au harvests

Another issue to note when dealing with archived web sites, particularly in the large numbers captured in a domain harvest such as the .au harvest, is the difficulty of providing access to them. There are currently a range of approaches to this, and also a range of reasons why access is restricted. At one end of the spectrum is the way in which PANDORA provides access to its selected web sites. At the other is the restrictive approach taken by the Danish Net Archive (netarchive.dk) which allows access to its domain harvest only to researchers who hold a masters degree. This is for data protection and privacy reasons (Jacobsen 2007).

Noted above for the .au harvest have been some of the technical issues associated with identifying blogs, a first step towards providing access to them. The POA research group intends to investigate this further.

Summary: four issues

There are at least four major issues for blog archiving and preservation. The first is how to identify Australian-specific material, as indicated in this paper. The second issue is: what to capture? There are two aspects to this question. The first is that of the selection criteria to be applied if a selective approach is used – but this is not the main concern of this paper. The second is that we while need to capture the blog postings and comments, how important is it also to capture all of the material linked from the blog? As noted above, blogs are highly linked and bloggers themselves consider that linked content is important to capture.

The third major issue is how to ensure blogs are captured at sufficiently short intervals to provide an accurate representation of the blog. It can be argued that this requires constant harvesting, that is, every time a change is made to a blog it is harvested. Domain harvests do not by definition have the aim of doing this, but are they, therefore, inadequate?

These three issues are not limited only to blogs; they also apply to other kinds of web resources. Techniques for capturing and processing static web documents (those which do not change much or at all – an example is a PDF document) are well developed, but this is not the case for dynamic documents (those which change frequently, such as many blogs). Snapshots of the web, such as the .au harvests, are a positive interim step towards preserving the web, but they are only an interim step.

The fourth major issue is providing access to the harvested blogs. This was briefly noted above, and will be the subject of further investigation by the POA research group.

There can be no doubt that the .au harvests represent a major step forward in attempts spearheaded by the NLA to preserve the national memory. However, as this paper has demonstrated, the harvesting process itself raises many technical issues. The example
of blogs in the .au harvest illustrates some of the other issues. The POA research group intends to continue its investigation into blogs. For example, we will compare their presence in the 2005, 2006 and 2007 .au domain harvests and develop methods to identify them and provide better access to them in these harvests. We intend to use the example of blogs to investigate the extent to which Australian society has embraced the online environment as public sphere. Specifically, we hope to estimate the success of the harvests in gathering a relevant cross-section of Australian blogs, and the extent to which those captured represent a broadening of voices within the Australian mediascape.

References


