This article is downloaded from

http://researchoutput.csu.edu.au

It is the paper published as:

Author: B. Pymm and J. Wallis
Title: Archiving the web: does whole-of-domain archiving = information overload?
Conference Name: ALIA 14th Exhibition and Conference
Conference Location: Sydney, NSW Australia
Publisher: Australian Library and Information Assoc
Year: 2009
Pages: 1-11
Date: 20-22 January 2009

Abstract: This paper aims to generate discussion on the question of very large data stores and their usefulness as research corpora as exemplified by the whole of domain web archiving undertaken by the National Library of Australia (NLA). Is the effort of creating such huge datasets, maintaining them over the long term and building appropriate access pathways providing a valuable resource for current and future researchers? Or is the highly selective approach, exemplified by the NLAs PANDORA archive, potentially more useful. Two basic issues were identified as of national concern and searches were undertaken on these terms across the 2007 whole of domain archive and in PANDORA. The relevance of the results obtained were then compared with reference to various indexing approaches, searching behaviour and desired outcomes. The need for further research into this area is highlighted by the conclusions of this study.

Author Address: r pymm@csu.edu.au
jwallis@csu.edu.au

http://researchoutput.csu.edu.au/R/-?func=dbin-jump-full&amp;object_id=8078&amp;local_base=GEN01-CSU01

CRO identification number: 8078
Archiving the Web: does whole-of-domain archiving = information overload?
Dr Bob Pymm, Charles Sturt University (rpymm@csu.edu.au)
Jake Wallis, Charles Sturt University (jwallis@csu.edu.au)

Abstract

This paper aims to generate discussion on the question of very large data stores and their usefulness as research corpora as exemplified by the whole of domain web archiving undertaken by the National Library of Australia (NLA). Is the effort of creating such huge datasets, maintaining them over the long term and building appropriate access pathways providing a valuable resource for current and future researchers? Or is the highly selective approach, exemplified by the NLA’s PANDORA archive, potentially more useful. Two basic issues were identified as of national concern and searches were undertaken on these terms across the 2007 whole of domain archive and in PANDORA. The relevance of the results obtained were then compared with reference to various indexing approaches, searching behaviour and desired outcomes. The need for further research into this area is highlighted by the conclusions of this study.

Introduction

In 1996, the Internet Archive (IA) in the United States pioneered an automated approach to web harvesting and archiving by conducting regular crawls of the entire Web, or at least its surface, using automated crawler software following pre-established rules. Taking this approach provides a relatively straightforward method for acquiring very large amounts of data – millions of files – with modest effort. This has resulted in the building of an archive currently estimated at two petabytes (PB) of data, and growing at a rate of 20 terabytes (TB) a month (Internet Archive, 2008). An automated approach such as this, while having the advantage of being simple and straightforward, does have the major disadvantage of being open to the possibility of legal challenge. No permissions have been sought from the site owners to acquire and archive their sites and in many countries, including Australia at the federal level, the provisions of legal deposit legislation do not cover web sites. Thus the Internet Archive, in an effort to minimize its legal exposure, provides an extensive Terms of Agreement clause to which those accessing the site must agree and provides mechanisms for rights owners to seek removal of their sites from the Archive. Charlesworth (2003, p. 42) however, notes that the Internet Archive is in a weak position in relying on such clauses and is essentially dependant upon the goodwill of rights holders. Yet over the 12 year life of the Internet Archive legal disputes have been limited and generally resolved without recourse to the courts (Simes and Pymm, 2009). Thus it would seem that this approach to web archiving pioneered by the IA, while it can present some potential problems regarding the material captured and provision of access to it, does provide a practical and manageable option for institutions interested in large scale web archiving whether or not enabling legislation exists.
While the Internet Archive was the first, and is still the most comprehensive, attempt at archiving the Web, others have adopted their approach of automatically gathering very large numbers of web sites within given parameters. Commonly, the main parameters refer to a national domain. Thus for example, the National Library of Norway has been attempting to archive “the entire Norwegian web space” since 2002 (Rustad, 2006) and the National Library of Australia (NLA) has been undertaking a similar approach with respect to the Australian domain since 2005. This is in addition to its established web archive, PANDORA which has been in existence since 1996. The PANDORA archive has been built up over this period by careful manual selection, gaining permission of owners and harvesting and checking results; a time consuming and relatively resource intensive process shared amongst the major state libraries and related institutions. PANDORA now comprises over 19,800 titles comprising 2.34TB of data (National Library of Australia, 2008).

The nature of the highly selective PANDORA archive means that it will only ever hold a tiny fraction of Australian web sites. However, the sites it does hold, having been carefully identified and selected for acquisition, suggest that it is a rich resource of quality information, highly relevant to current and future researchers. Also, given that permission for harvesting has been acquired, there are no legal issues regarding access and use of this resource.

However, the concern that PANDORA includes such a small proportion of the Australian domain led to the NLA investigating the possibility of undertaking a crawl to capture the entire “.au” domain. Thus the entire “.au” environment was trawled in 2005, 2006, 2007 and 2008 by the National Library of Australia (NLA), capturing over 500 million files in 2007 comprising 19TB of data (Koerbin, 2008). These crawls were undertaken by the Internet Archive following parameters provided by the NLA. They ran for around six weeks before being halted. This time frame was set based on the financial resources available for the crawl and the experience learned previously that by this stage, few new relevant sites were being discovered. The contents of the resultant archive were predominantly text/html objects, but also included large numbers of the other MIME (Multipurpose Internet Mail Extensions) media types such as audio, video etc. (Fellows et al., 2008). In addition to the raw file, accessible only by URL, a keyword index was created in order to facilitate access to this dataset.

For the NLA or any other body undertaking this exercise, considerable resources are required to undertake the initial crawl. These will also need to be supplemented on an ongoing basis in order to store and make the data accessible over the long term. Backups need to be created, software platforms acquired or maintained and at some stage a strategy for migrating the data to new media or to take advantage of updated software will have to be developed and implemented. In addition, staff resources have to be committed to monitor access, respond to queries and generally manage this significant ‘collection’. Thus whole of domain harvesting, while it may be technically quite feasible, does pose ongoing as well as once-off costs for the organization. In a publicly available paper on the National Library’s 2005 crawl which was outsourced to the Internet Archive in the US, cost categories were split into the initial crawl, indexing the contents, delivery and installation and National Library administrative costs. The actual cost of these tasks was not provided but this breakdown does give some idea of the one-off cost centres involved in the initial archive creation (Koerbin, 2005).
Little appears to have been written about the economics of web archiving. Generally it has been seen as a necessary part of preserving a society’s culture with a feeling that the Web may actually be changing traditional images of what, for instance, it means to be Australian (Ackland, Spink, Murphy, 2007). While this may be speculative and difficult to prove, the fact that there is a general feeling prevalent that the Web is impacting directly on individual and societal behaviours and attitudes (see for instance discussion of social networking sites such as Facebook (Kaufman, 2007)) means that the desire to preserve some or all of it would generally be seen as an appropriate response on the part of memory institutions. Interestingly, the great medium of influence in the latter half of the twentieth century, television, while it generated significant discussion as to its impact and influence on behaviour and society more generally, received far less attention with respect to preservation and long term accessibility of programming. Preservation of the Web and its contents seems to have become a much more visible issue.

Yet is this whole of domain approach worth the cost, involving as it does reliance upon a pre-programmed crawler working through this unwieldy mass of data? Is the resultant huge ‘haystack’ of data useful in any meaningful way? Does it enable researchers to discover ‘needles’ of information or identify broader trends within this undifferentiated stack?

The ubiquity of Web 2.0 means that, more than ever before, there is potential for everyone to become a publisher of digital content. The growth of online data is exponential, hence the efforts of the NLA and other national libraries to explore comprehensive preservation strategies through web harvesting. In addition to cost issues, there are other significant hindrances to the effectiveness of this approach, including:

- the implications of copyright restrictions on public access to material collected through automated harvesting as noted earlier;
- difficulties in capturing dynamically generated, non-HTML content (eg. images and multimedia objects served by JavaScript) and
- the complexities arising from the dispersed nature of the web
  - Australian content that resides on servers physically located outside of Australia and hence not within the .au domain (for example many blogs will be located on servers in the United States)
  - The difficulty of identifying Australian digital content using automated web crawling applications with limited semantic intelligence
  - The size of the resulting dataset and issues surrounding indexing for access.

The complexity of the issues arising from the size of harvested datasets is reflected in the difficulties around the facilitation of access. A shift from selective practices to forms of automated harvesting and indexing is mirrored in alternative approaches to data access. Subject terms and headings are replaced by keyword searching across dataset indexes. This is the approach that is strongly favoured in accessing the Web, the largest of all datasets. Search engines such as Google use web crawlers to map and index the web. It is these indexes, rather than the ‘live’ web that is searched when typing keywords into Google with the result that if a site, for whatever reason, is not yet indexed by Google, it is effectively invisible to most users.
Similarly, the content of the ‘deep’ web is beyond the coverage of most web crawlers and requires mediated and dynamic access points.

The importance of effective indexing to successful retrieval of information from such a large dataset is fundamental to the usefulness of this mass of data. One of the reasons for Google’s success is its methodology for returning ranked listings of results. It is the efficiency and effectiveness of its ranking algorithm that has ensured its popularity as the primary access point to online content. However, this comes at a substantial cost to the company with Google investing significant resources into the ongoing development of the software and its related patents (Google, 2007).

An alternative to Google’s ranked listing approach, is to present results arising from searching such huge quantities of data in a graphical manner. This approach is used by search engines such as Kartoo (www.kartoo.com) to offer a visual representation of web search results. The combination of web crawling and graphical representation is used for online social research by an international community of researchers using the Issue Crawler tool based at the University of Amsterdam (at www.issuecrawler.net). Issue Crawler is a ‘virtual collaboratory’ combining web crawling and visual tools for the representation of resulting datasets. It provides an effective means by which social researchers can undertake analysis of online networks, given the size and complexity of the web as a dataset.

The figure below demonstrates the graphical output of web crawling using Issue Crawler. This type of cartographic representation can be helpful in analyzing datasets resulting from web crawling in that the layout identifies relationships and clustering within online social networks around specific issues (in this instance indigenous health).
These or other methods for indexing, searching and presenting results from large datasets require sophisticated software development and expertise. They are of paramount importance to successful navigation and data discovery (finding the right ‘needles’ in the haystack) and are therefore the key component in providing access to non-programming users.

Given the importance of these indexing tools to accessing the ‘live’ web, it would appear that logically, accessing an archived version of the web, even if it comprises only a subset (such as the “.au” domain), will require equally sophisticated searching and interrogation tools if it is to be queried on anything other than domain name. Thus, as part of an ongoing research project underway at Charles Sturt University (CSU), focusing on the NLA web harvests and their contents, it was decided to pursue a simple approach to better understand the accessibility of the 2007 whole of domain crawl to researchers in comparison to the mediated and well indexed PANDORA archive. This approach involves focusing on two high-profile research topics and evaluating results from comparative searches.

This study will help evaluate the potential of automated harvesting and indexing in generating a publicly accessible collection of Australian digital heritage in order that future researchers may feel confident that they can discover relevant and appropriate material from this corpus. In addition, it may help indicate whether there is a balance between the current and ongoing costs of creating such a corpus and the potential benefits of future accessibility.
Methodology

The researchers selected two issues of national prominence, as represented by public concern and the government policy agenda, using the final report of the Australia 2020 Summit (http://www.australia2020.gov.au/) as a guide to significant and topical areas. One social and one environmental issue were selected and the search terms indigenous health and landcare program were identified. Using PANDORA and the 2007 whole-of-domain web harvest, the researchers investigated these two topics. These particular topics were chosen due to their ongoing social relevance, as emphasized by the importance placed upon them in the recent Federal government 2020 Summit Report, where both topics were identified as significant priority policy areas for the Commonwealth (Australia 2020, 2008).

Both subjects provide broad yet socially relevant areas of potential research interest that have high visibility and are thus likely to have been selected for archiving in some way within the National Library’s PANDORA archive. As such, they offer the opportunity for a comparative study of the value, efficiency and relevance of results produced from interrogation of PANDORA and the Crawl dataset.

Searches were conducted using the simplest approach – keyword searching. PANDORA is publicly available and allows for keyword and more sophisticated searching, as well as facilitating browsing behaviour through the use of subject headings. Its searching algorithms also involve the ranking of results based upon the terms that appear in the ‘curated title’ of the captured site.

The Web Crawl is not publicly accessible but was made available to the researchers with a simple keyword and URL index. It was therefore decided to use the straight keyword approach as applicable to both datasets.

Thus the terms indigenous health and landcare program were used as simple search terms to identify holdings within the PANDORA archive.

The first five records returned by PANDORA in response to these searches were considered to be important and authoritative resources on the subjects. These web sites had been carefully selected, harvested, catalogued and indexed with a view to their long term preservation. From this, it was felt a reasonable assumption could be made as to their quality and relevance for any researcher investigating this particular topic. The sites were also checked to ensure they had been acquired by PANDORA prior to August 2007, the date of the Web crawl.

The same terms were then entered into the 2007 Web Crawl. The resultant hit list was then searched to ascertain if the sites provided by PANDORA were returned in the first three pages of the results list. The limiting of the search to the first three pages is supported by research that suggests few users go beyond the third page of a hit list if they have not found anything relevant to their search. A recent study reported that 80% of searchers stop after the third page and review their search strategy rather than continue on through the dataset returned (Mastodonte,
Thus the researchers felt that three pages deep within the retrieved dataset represented a valid replication of user information seeking behaviour.

Finally, the Web Crawl was searched using the URLs of the top five PANDORA sites returned for each search term in order to ascertain the existence (or otherwise) of the sites in the Crawl.

**Results**


For the term *landcare program* Pandora returned a results set of 29 records. Again, the first record returned referred to a Federal government department site (the Department of Agriculture, Fisheries and Forestry), [www.affa.gov.au/content/publicationsde9d.html](http://www.affa.gov.au/content/publicationsde9d.html), focusing on the community landcare program. The site had been archived in 2004. The next three sites were linked to the then federal Department of Primary Industry and Energy and were archived in 2004. The fifth site had been archived in 2008 and was thus excluded from this research.

A search of the 2007 Web Crawl was then undertaken for the same terms. *Indigenous health* returned a dataset of 768,402 records. The first three pages of this dataset (30 records) were searched for occurrences of the sites identified as the top five in the PANDORA dataset. None of the PANDORA sites featured in these pages. *Landcare program* returned a dataset of 83,843 records. Again, the first three pages of this results list were searched for occurrences of the sites identified as the top four in the PANDORA dataset. One match was found for the first PANDORA site, a later version of the Department of Agriculture, Fisheries and Forestry’s site.

Intensive searching of the Crawl dataset was necessary in order to confirm the presence of the URLs identified in PANDORA. Finding and confirming the existence of these URLs proved to be a challenging task not least given the apparent truncation of URL strings provided to the Crawl search interface. However, all of the PANDORA sites noted above were eventually located in the Crawl dataset via their URL.

**Discussion**

The digital information environment is expanding at an unprecedented rate. Estimates vary – inevitably – but EMC Consulting (2008), suggest that in 2007, the digital universe comprised 281 exabytes and that this growth will continue exponentially. Making sense of such vast quantities of data becomes increasingly complex and costly as it grows in volume. Collecting and preserving even a fraction of such output requires new models and practices. Emerging paradigms for looking at data may help in gaining better understandings of what needs to be kept. One such paradigm has been identified by Chris Anderson (2008) as the Long Tail which
complements long held views exemplified by the Pareto Principle (see for example Chiorazzi, 2005), where a small proportion of available information meets the vast majority of needs. This can be shown graphically with Anderson’s chart clearly indicating the principle that a small number of instances meet a large number of requests. The tail stretches to infinity, meeting an ever decreasing level of need.

While Anderson uses the Long Tail view as a way of understanding consumer behavior and online marketing, it also provides a useful schematic with which to view web searching. Most users find what they are looking for (or give up) within the first few pages of results – no one is going to scroll through the millions of hits virtually any but the most explicit search of the web returns. However, searching a truncated version of the web, such as provided by PANDORA, delivers small numbers of pages to users which, should, by the selection process involved in their acquisition, and the weighting applied through the sophisticated indexing tools to assist searching, make them highly relevant. It can therefore be hypothesized that to some degree, PANDORA represents the head of the tail – high relevance to most people – while the entire web or whole of domain crawl represents the long tail – of interest to a small and diminishing number of users.

It therefore becomes a question of the cost/benefits involved in creating large datasets, such as domain harvests, in order to meet what inevitably will be a very small need. The finding that searching the Web Crawl did not readily return the key sites identified by PANDORA suggests simple keyword indexing is limited and that there is a need for more sophisticated – and thus expensive – indexing and ranking software. Alternatively, individual researchers can programmatically access the file (with appropriate permissions) but this would seem to heavily proscribe access to those with the necessary level of technical skills. Thus in addition to the establishment and ongoing support costs identified in the earlier discussion, in order to properly exploit this huge dataset, a significant investment would need to be made in index creation and sophisticated searching tools, incorporating an interface which meets the usability requirements of researchers whose information needs may be ill-defined and a results display in an appropriate
format to aid understanding. Whether this is appropriate given the likelihood of limited use of
the dataset has to be weighed against other organizational priorities.

The issue of copyright and possible privacy infringement is significant in this context.
Permission is not pre-sought in the automated harvesting process and the NLA has to weigh up
the risk inherent in providing public access to content for which no permissions have been
obtained. The institution may be willing to carry such liability, given the limited recourse to
legal action in negotiating the removal of content from the Internet Archive (as noted earlier).

If whole of domain harvesting is not undertaken, or is done only at reduced intervals (with or
without the development of sophisticated indexing software), the there is a high level of reliance
placed upon the selective archive, in Australia’s case PANDORA, to hold a representative
sample of sites highly relevant to Australia’s digital culture. It therefore increases the need for
serious levels of quality control in both the selection and capture of sites to ensure their
completeness and relevance. In particular, it gives prominence to the issue of selection
guidelines for PANDORA which traditionally, have focused on more ‘authoritative’ sites such as
those produced by government departments. Inevitably, the PANDORA approach is not able to
cover every single issue, blog or digital object type, given the limitation of resources available
for this work-intensive selective archiving approach and this needs to be acknowledged and open
for discussion if it becomes the primary web archiving approach. These are all areas for further
research, vital to better inform the debate over the way in which digital culture is preserved.

Conclusion

We have noted the exponential growth of digital information. This growth reflects the shift
towards the online environment as a vibrant cultural space, a public and media sphere of
increasing social significance. The growth of the Web as a social space relates directly to its role
in providing the cultural milieu for Generation Y; where news and entertainment are consumed
online, social relationships and networks are established and maintained, and learning and work
are often online activities. This online environment also offers an effective tool for groups
marginalized from the mainstream political process. The dispersed networked environment, by
its very nature, provides an ideal vehicle for collective mobilization and shared identity around
single issue politics. This capacity has been capitalized by organizations such as GetUp
(www.getup.org.au) which draws together a range of issues under one umbrella. GetUp uses the
affordances of the networked environment (e-mail petitions and campaigns, the dissemination of
information on specific causes via the organisation’s web site, the use of digital audio and video
to promote causes) to build participation and online community around political issues.

Hara (2008) also sees the Internet as extending the potential for collective action and broadening
the traditional media confines of the public sphere by creating a more fluid link between our
private and public selves. Online activities such as blogging, social networking and the
formation of online communities certainly support Hara’s analysis. The Australian Centre for
Public Communication’s (2008) study of the use of new media in the recent Australian Federal
Election campaign for example notes the transitional nature of changing practices around both
political activity and engagement as the Australian population gradually gains maturity in the use of the online environment as a cultural and potentially democratic medium.

Aware of the growing importance of the online sphere, the NLA has grappled with this emergent digital culture through adjustments to its collecting practices. The institution is employing Web 2.0 philosophy, technologies and applications to broaden its inclusiveness. One might go so far as to describe the NLA’s attempts to broaden participation in its collection practices as inherently democratic. Partnership with the online photo hosting platform Flickr (www.flickr.com) allows anyone to upload digital photographs to a national collection (Picture Australia – www.pictureaustralia.org).

Given this increasing attention to the role of the Web as a key component in informing and leading in the social and political spheres, and the sometimes complicated and opaque nature of online publication, the web harvesting approach is an attempt to capture the transitory virtual content that represents Australian digital culture. The size of the dataset creates issues around which there needs to be careful thinking if the investment in data capture is to prove worthwhile. Searching PANDORA is easy and straightforward in terms of the effectiveness of its interface and simple keyword search, delivering a results list that, because it has been evaluated as to its integrity and quality, provides immediate access to high level, high relevance data. The process of mediated selection, incorporating a level of semantic understanding which is currently impossible to replicate through artificial intelligence, ensures that the resultant sites selected for permanent retention are significant in their representation of Australia’s cultural heritage. Whether the selective approach or the whole of domain crawl approach is best suited to meet the needs of future researchers is difficult to convincingly demonstrate. Either has its possibilities, limitations, legal and ethical challenges and it is hoped that this research encourages further enquiry and debate into this complex acquisition, preservation and access dilemma.
References


