Characteristics of .au Websites: An Analysis of Large-Scale Web Crawl Data from 2005

Robert Ackland
Centre for Social Research
Research School of Social Science
The Australian National University
Email: robert.ackland@anu.edu.au

Amanda Spink
Faculty of Information Technology
Queensland University of Technology
Email: ah.spink@qut.edu.au

Peter Bailey
Information Engineering laboratory, ICT Centre
Commonwealth Scientific and Industrial Research Organisation
Email: peter.bailey@csiro.au

Paper presented at the 13th Australian World Wide Web Conference,
30 June – 4 July 2007, Brisbane.

ABSTRACT

This paper presents a preliminary analysis of Websites contained in a large-scale crawl of the .au domain that was collected by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in late 2005. Our analysis is based on a 10 percent random sample of the CSIRO crawl dataset, which contains around 10 million Web pages from approximately 200,000 Websites. This paper represents the first step in a larger planned project, which we title the ".au Census", the aim of which is to use large-scale Web crawls to conduct research into commercial and social aspects of the Australian public Web. The primary aim of the present paper is to analyse the data's properties. We find that the dataset predominantly consists of commercial Websites (78 percent of sites are ".com.au") while .edu.au, .org.au and .net.au sites each account for around 5-6 percent of all sites (and .gov.au sites are approximately 2 percent).

INTRODUCTION

The Web is now an important element of life for many Australians. The Web is a major source of consumer information, with consumer- or business-related searches continuing to dominate commercial Web search engine queries (Spink and Jansen, 2004, forthcoming). Most Australian businesses maintain at least a minimal Web presence (for e.g. contact details), while some base their entire business model around the Web. The Web is also increasing important to Australian politics, with many Australians obtaining political information via party Websites or politically-oriented Weblogs. One can also expect that there will be social impacts of the Web, but compared with the impacts of the Web on Australian commerce and politics, any effects on Australian society may be less direct and take longer to eventuate.

Recent sociological research has found that while the "traditional" conceptualisations of Australian national identity that emerged in the 20th Century (e.g., the ANZACs, sporting success and the casual lifestyle) are still strong in the popular imagination (Phillips and Smith 2000; Smith and Phillips 2001), they are being challenged by social movements (e.g. environmentalists, feminists

and advocates of Aboriginal and multicultural constituencies) aiming to redefine the national identity in more inclusive or progressive ways (Wickes, Smith and Phillips 2006). One social impact of the Web may therefore be in its use as a tool for promoting alternative conceptualisations of Australian national identity; the use of the Web in this way would parallel its extensive use in activist networks such as the global justice movement and the environmental movement who endeavour to bypass the mass media, which tend to highlight mainstream views (Castells 1997, Garrett 2006).

Organisations such as the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and National Library of Australia (NLA) are collecting and archiving large-scale crawls of the Australian public Web. Internationally, the Australian public Web is also being archived by the Internet Archive project[1], which has been conducting large-scale crawls of the Web since 1997. The primary motivation behind the large-scale crawl activities of CSIRO is information retrieval research, while for the NLA and the Internet Archive, the major goal is the preservation of digital heritage. It is notable that the National Science Foundation-funded Cornell Cybertools project[2] is devising methods and tools for conducting empirical social science research using data from the Internet Archive; social science-oriented research of large-scale crawls is therefore gaining attention internationally.

To our knowledge, there has been no attempt to conduct empirical research into Australian commerce and society using large-scale crawls of the Australian public Web. The majority of Australian social-science oriented Web studies have been qualitative analyses emerging from cultural and media studies. Long and Allen (2001) use a case study of a denial of service attack on an Australian internet relay chat network to study libertarianism on the Web. Allen (2002) reviews Australian policy relating to Web censorship. Holloway (2002) presents a case study on the digital divide in Western Sydney. There are chapters focusing on Australian internet history, policy and culture in Goggin (2004). The limited empirical research that does exist generally relates to Web use – analysis of survey data to show how Australians are using the Internet. The Internet Activity Survey (ABS 2006) provides details on aspects of Web access services provided by Internet Service Providers in Australia. Gibson (2003) uses the 2001 Australian Census (which included questions on computer and internet use at home and work) to study social and spatial inequalities in personal usage of ICTs. Madden and Savage (2000) and Madden and Coble-Neal (2003) estimate an econometric model of individual internet use and show how the price of internet services, sociodemographic factors and connection capacity affect internet use.

The present paper is the first stage of a planned larger project (the ".au Census"), the aim of which is to conduct research into Australian commerce and society using large-scale Web crawl data. While we will use automatic text analysis methods in this research, we envisage that such methods will not tell us enough about a given Website, and consequently, the Web pages will also need to be viewed and classified by human analysts. If we wanted to characterise the Web presence of companies in footwear retail, for example, it would not be enough to know that a site has "footwear" as a meta keyword – it could be a Website for a footwear manufacturer, rather than retailer. To make the distinction it would be necessary to look at the other meta keywords (and probably the entire Web page).

As human classification of 10 million Web pages is clearly not feasible, we need a sampling approach. Existing techniques for sampling from the Web are well studied (e.g. Henzinger *et al* (2000), Broder *et al* (2006)) and aim to obtain close approximations to a true random sample. Various characteristics of the Web make this difficult however. A key innovation of the .au Census project will be our investigation of adaptive sampling (Thompson and Seber 1996) for constructing samples of Web pages that are subsequently used to infer characteristics for the underlying population (the Australian public Web). While adaptive sampling is commonly used to study hard-

1 http://www.archive.org

2 http://www.news.cornell.edu/stories/Sept05/NSFcybertools.dea.html

to-reach populations where there is clustering of observations, for example because of the existence of social networks, the only example of its use for analysis of Web data is Ackland (2005). The reason why a "traditional" sampling approach is not appropriate is the rarity of the underlying populations: only a small proportion of the pages in the CSIRO dataset will be directly relevant to footwear retailing, for example, and a "standard" 1% sample of Web pages would not include enough relevant pages.

In the .au Census project, we will be taking the CSIRO crawl data as our underlying "population" of Web pages. Given this, a true random sample can be taken from the "population", and "hard-to-reach" sub-populations can be much more readily identified.

However, the "population" itself clearly is not random as crawlers are not a random sampling technology. For example, it is not feasible for CSIRO to crawl the entire .au domain for at least four reasons. First, there is the question of resources – even efficiently-designed Web crawlers such as that developed by CSIRO take time to crawl the Web and there is a necessary trade-off between extensiveness or coverage of an individual crawl and the frequency of new crawls (which are carried out to obtain up-to-date data snapshots). Researchers who are constructing large-scale crawls such as the CSIRO crawl are generally going to aim for a "sweet spot" whereby the crawl is limited only to relevant or interesting parts of the Web, which are then regularly crawled. However, achieving this balance is the subject of current research in information retrieval, and hence we cannot take it as given that the 2005 CSIRO crawl is representative of the parts of the Australian public Web that are of interest for commercially- and socially-oriented research. The second reason why the CSIRO crawl will not be a complete snapshot of the .au domain relates to the existence of Web pages which are not connected to other parts of the Web (this is sometimes called the "dark Internet" http://en.wikipedia.org/wiki/Dark_internet) – if a Web page is not linked to by any other Web page, and in particular not linked to by other pages in the .au domain (or linked to by only a few pages) then it will be very difficult for a crawler to find the page and hence it is unlikely to be included in the crawl. The third reason is due to the policies imposed by the crawler. For example, to ensure reasonable coverage across many servers, individual large servers are not necessarily crawled in entirety; the limit is set by crawler policy configuration. Similarly, the crawler adheres to robots.txt conventions which allow Web server administrators to set restrictions on what parts of their site and/or what crawlers are allowed to access. The settings on individual servers could preclude large amounts of their data from being crawled. The fourth reason is that large quantities of spam (junk) Web pages can be created by individuals with intent to put their information in front of search engine users for the purposes of generating revenue through advertisements or for other more illicit purposes. These crawler "traps" are constantly being changed and updated and any particular crawl may contain an unknown number of "spam" pages.

DATA COLLECTION

We use a large-scale crawl of the Australian public Web (defined here as publicly-available Web pages with root URLs containing the .au country-code Top-Level Domain) that was collected by CSIRO in 2005.[3] The crawl dataset contains around 10 million Web pages from approximately 200,000 Websites (a Website is defined here as Web pages with same root URL). Only the text information from Web pages is stored in the CSIRO crawl dataset (images are not collected, although the links to the images are preserved). If a Web page contains a link to a PDF or an Office application file (such as Word or Excel files), the CSIRO crawler stored a text representation of the application file (we do not make use of this data here).

The analysis in the present paper is based on a 10 percent random sample of the Websites in the CSIRO crawl dataset – 19,492 Websites. To clarify, the random sample is based on selecting from

3 The root URL of an HTTP web page is the portion of the URL between the "http://" and the next "/". For example, the root URL of the web page http://www.example.com/mydir/mypage.html is "www.example.com".

the total set of Websites (not from their Web pages). Nor was any bucketing by the size or other properties of the Websites carried out prior to random samples being taken. The consequence of this (due to the typical power law distribution of server sizes on the Web) is that many more small Websites will be selected than medium or large Websites.

DATA ANALYSIS

Number of Australian Web domains

There are 10 types of Australian Website domains (.AUDA second level domains) in the 10% sample:

- .com.au – For Australian commercial entities, such as companies (with ACN as registered with ASIC), and businesses (registered with state governments). Note that the CSIRO crawl dataset does not include .com Websites originating from Australia that does not include ".au".
- .edu.au – For Australian educational institutions registered at the federal or state level.
- .org.au – For Australian charities and non-profit organisations.
- .net.au – For Australian commercial entities, such as companies (with ACN as registered through ASIC), and businesses (registered with state governments)
- .gov.au – For Australian federal, state and local government bodies.
- .asn.au – For Australian incorporated associations, political parties, trade unions, sporting and special interest clubs..
- .csiro.au – For the sole use of the Commonwealth Scientific and Industrial Research Organisation (CSIRO).
- .oz.au - Australian Websites, not limited to any particular type.
- .info.au - Australian Websites providing information in a particular subject area.
- .id.au - For individuals who are Australian citizens or residents.

The majority (78 percent) of .au Websites are .com Websites (Table 1). Educational, organisational and network Websites each account for about 5-6 percent of the sample, 2 percent of the sample are government Websites and 1 percent are association Websites. The remaining 4 domains (.csiro.au, .oz.au, .info.au, .id.au) form less than 0.5 percent of all .au Websites.[4] The largely commercial nature of the Australian Web has evolved over the last ten years as business content increased and also began to dominate Web searching (Spink and Jansen, 2004). The social use of the Web may not predominate in an analysis of Websites, but can be seen more in networking, blogs, email, etc. Note that the number of Websites in a category does not correlate with the number of Webpages in that category (see below).

[Place Table 1 Here]

Number of Web pages

On average, around 50 Web pages were crawled in each Website. There is marked variation in the size of Websites: government Websites are the largest with approximately 140 pages per site, while Websites in the remaining 5 domains (that we consider here) had an average size of at most around half of the government sites. Websites in the two domains most used for commercial purposes (.com.au and .net.au) were the smallest, with an average size of 40-50 pages.

4 In the analysis that follows, we do not discuss these four domains because the number of websites in each domain is too small for statistically reliable conclusions to be made.

Number of external hyperlinks

We define an external hyperlink as a hyperlink to a Web page with a different root URL to the page containing the hyperlink. Overall, the mean number of (unique) external links per Website for any domain was 42 and as we found for Web pages, .com.au sites make a lower-than-average number of external links (35).[5] This compares with .org.au sites (56 links), .edu.au sites (78 links) and .gov.au sites (96 links). The above analysis for the number of Web pages and external hyperlinks indicates that commercial sites are, on average, smaller and contain fewer external hyperlinks. However, as commercial Websites are by far the largest Web domain, 70.4 percent of all Web pages and 64.9 percent of all external Web links are from commercial Websites. In other words, most Websites, Web pages and Web links on the Australian public Web are commercial in nature.

Meta keywords

Meta keywords are terms that Website authors use for a variety of reasons. Sometimes (in particular with .gov.au sites due to Australian Government best practice guidelines) there is an honest attempt to describe their site across various standard metadata fields to the rest of the world and as such, they provide useful information for identifying the nature of sites. In other circumstances, particularly commercial organisations, there can be an incentive to exaggerate or purposefully to attempt to mislead. Meta keywords are in general not viewed by individuals as they appear in header information which is not rendered by Web browsers. Due to the inability of search engines to trust the self-representation of commercial organisations, meta keywords are uniformly ignored from an indexing and ranking perspective. Some commercial organisations use meta keywords to accurately identify information; this is primarily from an internal business process perspective as it enables them to create their own enterprise search facilities that accurately find their products when users carry out searches with the organisation's search engine. However, this practice cannot be relied on.

Overall, the average number of (unique) meta keywords per site was 27.[6] In contrast to what we found for external hyperlinks, the average number of meta keywords on commercial sites is about the same as the average over all domains, while .org.au and .edu.au sites on average use less keywords. Government sites on average contain the most meta keywords – 56 keywords per site.

In Table 2, listings of the top-40 meta keywords (in terms of frequency of use) are presented for the 6 major domains.[7]

[Place Table 2 about here]

The following are some initial conclusions:

- To the extent that the appearance of a city or state name as a meta keyword on a commercial site indicates that the Website is related to a commercial activity in that location, the distribution of city/state name meta keywords accords with expectation: Sydney/NSW and Melbourne/Victoria are highly-ranked, as is Queensland (the destination of much travel in Australia). Adelaide is ranked much further down the list, and Hobart doesn't even make it into the top-40. Commercial sites on the Australian public Web are primarily focused on holidays/tourism, restaurants, real estate, entertainment and sport.

5 If a hyperlink to an external web page was contained on more than one page within a given website, it was only counted once.

6 If a particular meta keyword was contained on more than one page within a given website, it was only counted once.

7 Variations in spelling and capitalisation were taken account of. For example, the frequency count of 1187 for "accommodation" includes the following variations: Accomodation (17), accomodation (367), Accommodation (96), accommodations (18), accomodations (10), accommodation (676), ACCOMMODATION (3).

- The fields/disciplines of study that are most prominent (in terms of keyword frequency) on the Australian public Web are: science, technology, environment, business, engineering and management.
- The striking feature regarding the .org.au sites is the predominance of keywords relating to religion: christian (40), church (39), god (23), jesus (20), religion (20). Otherwise, .org sites appear to be most focused on: community, education, health, children, environment, sport, art.
- As expected, the keywords for .net.au sites are focused on technology and the internet.
- The areas of government that are represented in the keywords on .gov.au sites are: employment, environment, community, service, law, education, health, law, tourism and training.

DISCUSSION AND CONCLUSION

In this paper, we have presented a preliminary analysis of a 10 percent sample of data from CSIRO's 2005 large-scale crawl of the Australian public Web. Our main aim was to analyse properties of the CSIRO crawl data. Another aim was to establish a benchmark for comparative studies over time (CSIRO are presently conducting further crawls of the .au domain).

Does the quantity and nature of Web material in the dataset match our expectations regarding how Australians are using the Web (either as producers or consumers of Web material)? One of our key findings is the predominance of commercial- and business-oriented material on the Australian public Web. Based on our analysis, the non-commercial Web represents at most 22 percent of Websites and 30 percent of Web pages (these figures include .net.au sites that may in fact be commercial). On this basis, we would have to conclude that the Australian Web is in fact mostly a business Web.

The government, CSIRO and educational Websites represent a particular group in that they share some common characteristics. They are of course non-commercial, and are much deeper and denser than commercial Websites. This may reflect that nature of their genres as more stable in nature than commercial businesses, and the policy requirements to communicate effectively at a distance.

While our meta keyword analysis provided some initial insights into the nature of Australian Websites, including the surprising finding that many .org.au sites appear to be of a religious nature, this analysis also revealed the limitation of basic frequency counts of meta keywords. In future analysis, we plan to use an adaptive sampling approach that will enable us to carefully study a small sample of Websites more carefully (thus allowing us to move beyond basic frequency counts of meta keywords).

REFERENCES

Ackland, R. (2005). "Estimating the size of political Web graphs," mimeo, The Australian National University. http://acsr.anu.edu.au/staff/ackland/papers/political_web_graphs.pdf

Allen, M. (2001). "Reviewing Australia's attempts at Web censorship," *Association of Web Researchers v2 Conference 2001, Minneapolis, October*

Australian Bureau of Statistics. (June 2006). *Web Activity – Australia*. [http://www.abs.gov.au/ausstats/abs@.nsf/e8ae5488b598839cca25682000131612/6445f12663006b83ca256a150079564d!OpenDocument]

Broder, A., Fontura, M., Josifovski, V., Kumar, R., Motwani, R., Nabar, S., Panigrahy, R., Tomkins, A. and Xu, Y. (2006). "Estimating Corpus Size via Queries" In *CIKM '06: Proceedings of the 15th ACM international conference on Information and Knowledge Management* (Arlington, Virginia, USA). ACM Press. New York, NY, USA, 594-603.

Castells, M. (1997). *The Power of Identity. The Information Age: Economy, Society and Culture Vol. II*. Oxford: Blackwell.

Garrett, R. Kelly. 2006. "Protest in an information society. A review of literature on social movements and new ICTs." *Information, Communication & Society 9*(2): 202-224.
Gibson, C. (2003). "Digital divides in New South Wales: A research note on socio-spatial inequality using 2001 census data on computer and Web technology," *Australian Geographer, 34*(2), 239-257.
Goggin, G. (2004). *Virtual Nation: The Web in Australia*. Sydney: UNSW Press.
Henzinger, M., Heydon, A., Mitzenmacher, M. and Najork, M. (2000) "On Near-Uniform URL Sampling," In *9th International World Wide Web Conference (WWW9)*, Amsterdam. The Netherlands, 295-308.
Holloway, D. (2002). "Disparities in Web access: A case study of the digital divide in Western Sydney," *Australian Journal of Social Issues, 37*(1), 51-69.
Long, J. and M. Allen (2001). "Hacking the undernet: Libertarian limits; commercial containment," *Australian Journal of Communication, 28*(3), 37-54.
Madden, G. and G. Coble-Neal (2003). "Web use in rural and remote Western Australia," *Telecommunications Policy, 27*(3-4), 253-266.
Madden, G. and S. J. Savage (2000). "Some economic and social aspects of residential Web use in Australia," *Journal of Media Economics, 13*(3), 171-185.
Phillips, T and P. Smith (2000). "What is 'Australian'? Knowledge and attitudes among a gallery of contemporary Australians," *Australian Journal of Political Science*, *35*(2), 203-224.
Smith, P. and T. Phillips (2001). "What is UnAustralian?" *Journal of Sociology*, *37*(4), 323-339.
Spink, A. and B. J. Jansen (forthcoming). "Trends in searching for business and e-commerce information on Web search engines," *Journal of Electronic Commerce Research*
Spink, A. and B. J. Jansen (2004). *Web Search: Public Searching of the Web*. Berlin: Springer.
Thompson, S. K and G. A. F. Seber. (1996). *Adaptive Sampling.* New York: J. Wiley & Sons.
Wickes, R., P. Smith and T. Phillips (2006). "Gender and national identity: Lessons from the Australian case," *The Australian Journal of Political Science*, *41*.

Table 1: Analysis of 10% sample of 2005 CSIRO crawl data

	com.au	edu.au	org.au	net.au	gov.au	asn.au	id.au	csiro.au	oz.au	info.au	All
Number of Websites	15245	1324	1228	1032	399	209	38	10	5	2	19492
(% of Websites in sample)	78.2	6.8	6.3	5.3	2.0	1.1	0.2	0.1	0.0	0.0	100.0
Number of Web pages	727705	93330	93106	44481	56482	14516	2385	1321	621	163	1034110
(% of Web pages in sample)	70.4	9.0	9.0	4.3	5.5	1.4	0.2	0.1	0.1	0.0	100.0
Mean number of Web pages per Website	47.7	70.4	75.8	43.1	141.5	69.4	62.7	132.1	124.2	81.5	53.05
Number of external hyperlinks	536772	102617	69201	56928	38242	10270	4145	1872	6386	994	827427
(% of external links in sample)	64.9	12.4	8.4	6.9	4.6	1.2	0.5	0.2	0.8	0.1	100.0
Mean number of external hyperlinks per Website	35.2	77.5	56.3	55.1	95.8	49.1	109.1	187.2	1277.2	497.0	42.4
Number of meta keywords (unique within a Website)	428377	24283	19115	18229	22368	3430	639	614	57	51	517163
(% of meta keywords in sample)	82.8	4.7	3.7	3.5	4.3	0.7	0.1	0.1	0.0	0.0	100.0
Mean number of unique keywords per Website	28.1	18.3	15.5	17.6	56.1	16.4	16.8	61.4	11.4	25.5	26.5

Table 2: Meta keywords

.com.au	frequency	.edu.au	frequency	.org.au	frequency
australia	3069	education	168	australia	179
accommodation	1187	australia	139	victoria	55
sydney	832	school	108	community	50
queensland	823	research	102	new south wales	49
new south wales	805	learning	75	education	46
melbourne	744	student	65	health	42
holiday	702	university	60	sydney	42
restaurant	676	The University of Western Australia	50	christian	40
service	655	course	52	church	39
victoria	641	science	52	history	31
western australia	627	children	48	volunteer	31
business	598	staff	47	melbourne	30
hotel	547	website	45	news	30
south australia	524	teaching	44	board	28
real estate	509	technology	41	club	28
property	503	public schools	40	queensland	27
brisbane	473	new south wales	38	south australia	27
perth	459	contact	37	information	26
travel guide	427	sydney	35	support	25
entertainment	422	information	34	children	24
shop	419	postgraduate	34	environment	24
motel	417	training	32	forum	24
tourism	408	college	31	service	24
design	398	primary	31	training	24
photo	394	centre	29	download	23
event	388	study	29	free	23
vacation	386	job	28	god	23
sport	385	resource	28	sport	23
travel	385	undergraduate	28	database	22
shopping	378	academic	26	resource	22
information	367	conference	26	event	21
home	363	environment	25	comment	20
wedding	360	business	24	game	20
professional	359	career	24	hacker	20
club	354	engineering	24	jesus	20
news	351	management	24	link	20
community	339	employment	23	religion	20
adelaide	334	health	23	school	20
link	326	library	23	western australia	20
history	324	preschool	23	art	19

Table 2: Meta keywords (cont.)

.net.au	frequency	.gov.au	frequency	.asn.au	frequency
australia	146	australia	61	australia	35
new south wales	46	government	39	western australia	12
sydney	45	employment	35	association	10
western australia	42	environment	32	school	10
queensland	35	information	31	south australia	10
accommodation	31	job	31	club	9
internet	29	community	30	training	9
brisbane	28	service	30	perth	8
community	28	victoria	30	support	8
melbourne	28	law	28	education	7
south australia	28	map	27	event	7
victoria	28	queensland	27	newsletter	7
business	24	history	25	championship	6
perth	24	research	25	disability	6
design	23	council	24	resource	6
free	22	education	24	sport	6
music	20	health	24	victoria	6
service	20	legislation	23	volunteer	6
software	20	new south wales	22	adelaide	5
training	20	publication	22	committee	5
website	20	development	21	conference	5
holiday	19	business	20	contact	5
hosting	19	funding	20	council	5
technology	17	local government	20	disabled	5
tourism	17	privacy	20	horse	5
link	16	south australia	20	queensland	5
network	16	tourism	20	report	5
news	16	training	20	walk	5
rural	15	career	19	water	5
adelaide	14	management	19	carer	4
email	14	news	19	children	4
management	14	western australia	19	competition	4
web design	14	act	13	dance	4
book	13	children	18	endurance	4
computer	13	library	18	endurance riding	4
download	13	program	18	history	4
event	13	report	18	information	4
health	13	school	18	leadership	4
property	13	tender	18	map	4
rental	13	website	18	national	4