The Australian web domain harvests: a preliminary quantitative analysis of the archive data

Paul Koerbin
Manager Web Archiving
National Library of Australia

15 April 2008 (revised)

Acknowledgements

Most of the data used in this report was derived from the harvest logs by Alex Osborne and Tamara Ross.

Published by
National Library of Australia
Parkes Place, Canberra ACT 2600 Australia

Contents

Boxes, figures and tables

1 Introduction

The aim of this report is to present some analysed quantitative data about the content of three Australian web domain harvests conducted between 2005 and 2007. It provides a substantial update to the data about the Australian web domain and domain harvests that has been publicly available in the *Report on the crawl and harvest of the whole Australian web domain undertaken during June and July 2005*[1].

This report does not extend to an evaluative analysis of the data and any conclusions suggested in this report are only preliminary. The more modest ambition of this report is to provide information that may be of interest to others for the purpose of comparison with other large scale web archiving activities.

2 Size of the Australian web domain harvests

Three large scale harvests of the Australian web domain have been undertaken by the National Library of Australia in collaboration with the Internet Archive. The first harvest was conducted between 13 June and 18 July 2005, the second between 18 August and 22 September 2006 and the third between 28 August and 24 September 2007.

The harvest in 2005 was scoped as a time limited crawl, set to harvest broadly and deeply within the .au top level domain for a crawl time of four weeks. Subsequent harvests in 2006 and 2007 were scoped as size limited crawls. Both crawls had a target of a minimum of 500 million unique URLs.

In addition to harvesting content from the .au top level domain, some content identified during the crawl process as being on non .au domains but hosted on Australian located servers was also included.

Table 1 – Size of the Australian domain harvest data sets

Domain Harvest Date	2005	2006	2007
Unique documents (files) crawled	185,549,662	596,238,990	516,064,820
Total documents (files) crawled	189,824,119	621,664,876	523,510,945
Hosts	811,523	1,260,553	1,247,614
Raw data size	6.69 TB	19.04 TB	18.47 TB
Compressed ARC file size	4.52 TB	10.48 TB	10.18 TB

The National Library has conducted selective web archiving for the PANDORA Archive since late 1996. The chart below provides a comparison of the PANDORA Archive content with the domain harvests. The comparison is based on unique hosts. Figures for the number of unique files are not available for the PANDORA Archive as much of the content is re-gathered as the result of scheduled harvesting with the result that there is considerable duplication of content

[1] Available at http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf

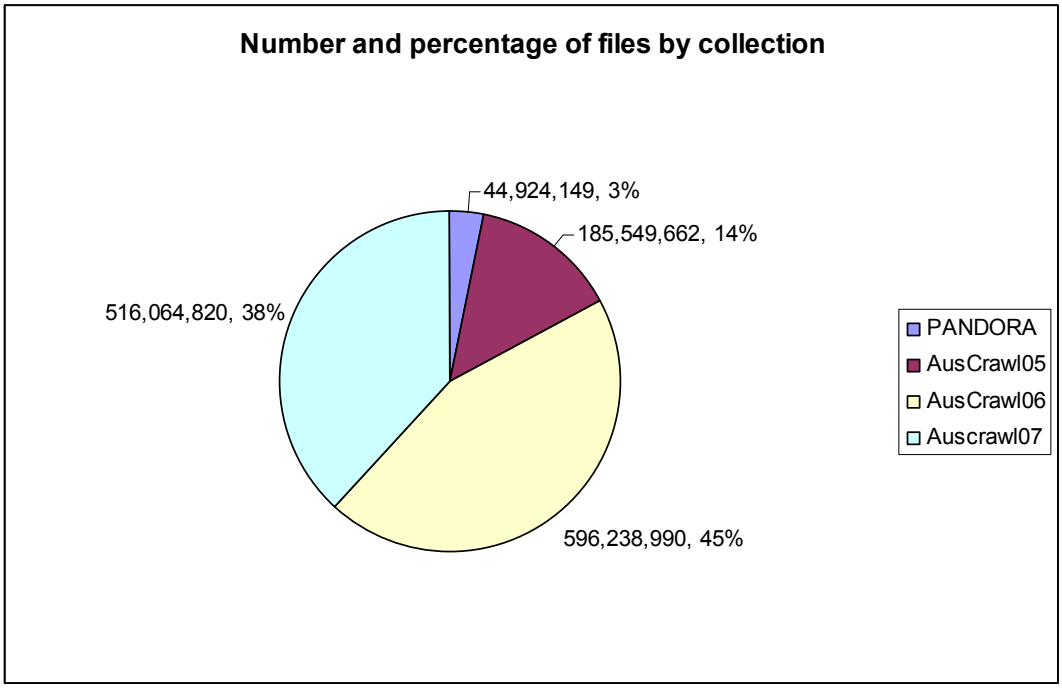
within the PANDORA Archive. While the figures used for comparison with the domain harvest are from October 2007, the PANDORA Archive continues to grow at a current average monthly rate[2] of 55 gigabytes or 1.25 million files[3].

Table 2 – Size of the PANDORA Archive as at October 2007

PANDORA Archive		Percentage of all archive data
Files	44,924,149	3.35%
Hosts (all)	42,936	n/a
Hosts (au)	10,037	n/a
Size	1.79 TB	3.89%

The following two graphs show the relative sizes of the National Library's web archive collections: the PANDORA Archive and the three domain harvest collections.

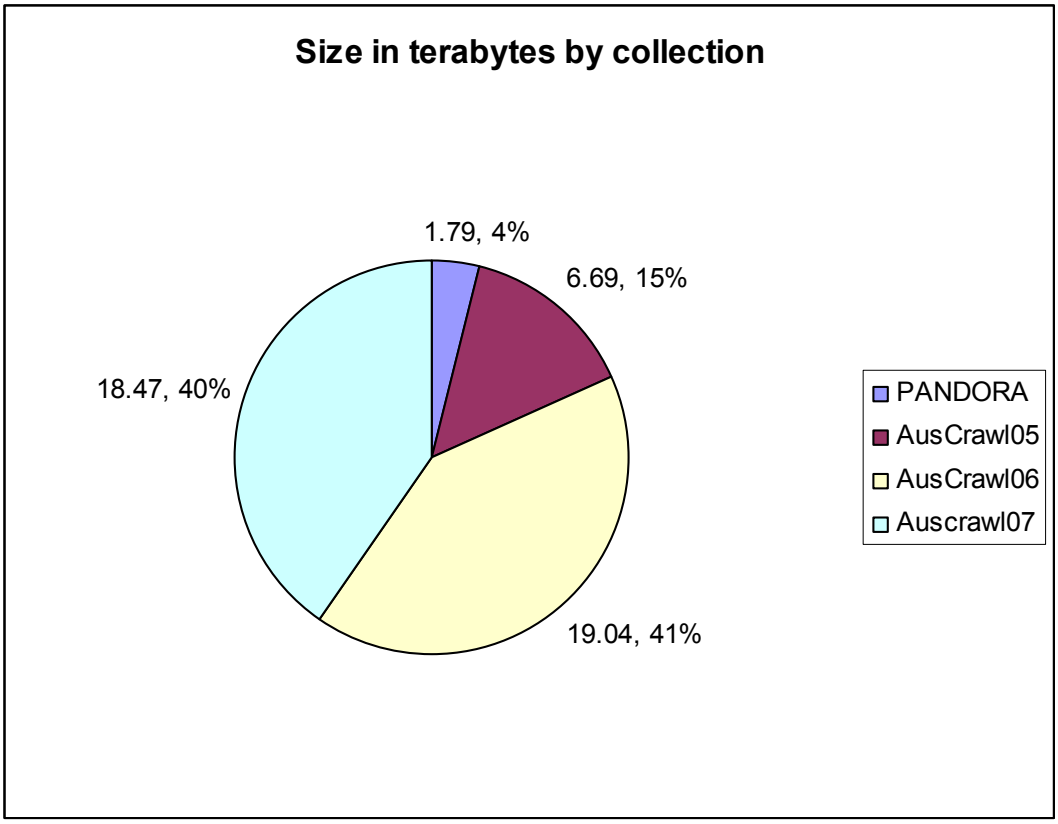
Figure 1 – Number and percentage of files by web archive collection



[2] Based on growth for the six months to February 2008.

[3] The current size and growth rate of the PANDORA Archive is posted at http://pandora.nla.gov.au/statistics.html

Figure 2 – Size in terabytes by web archive collection

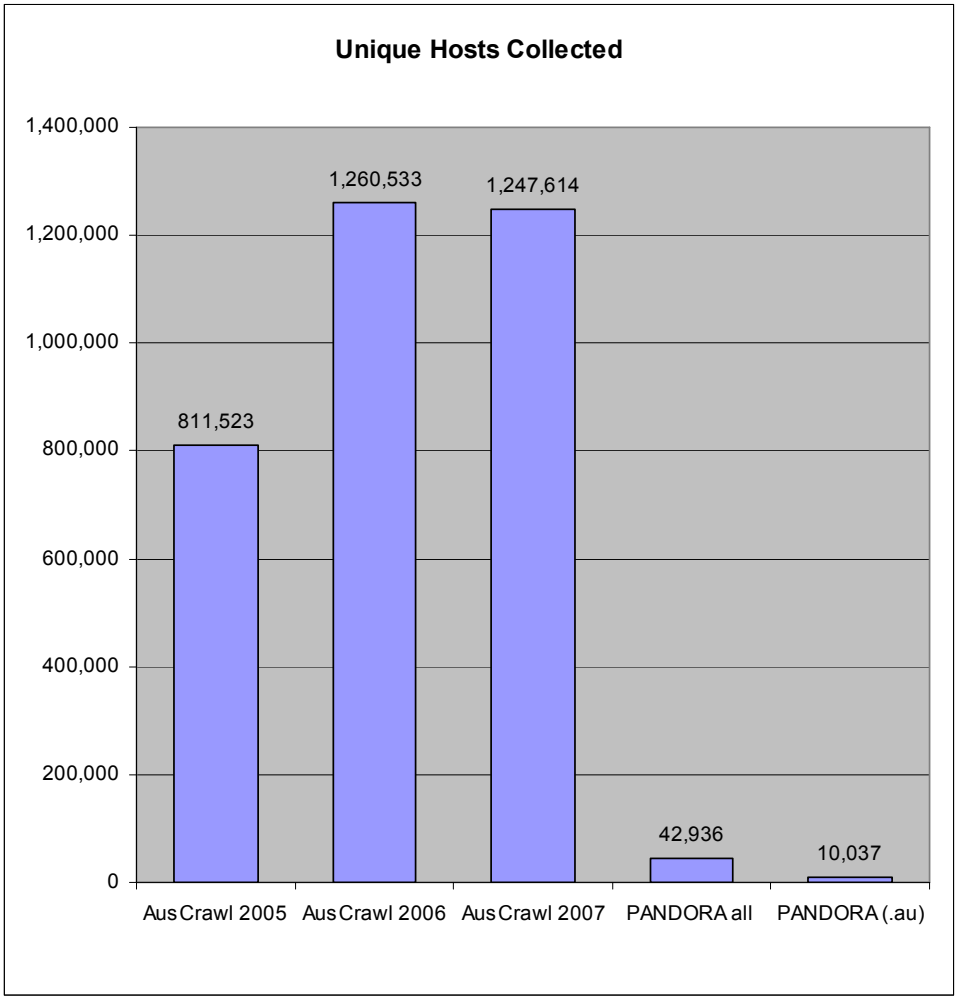


The following graph shows the relative sizes of the three domain harvest collections and the PANDORA Archive based on the number of unique hosts present in the collections.

Because there is much duplication in the content of the PANDORA Archive at the file level due to the regime of scheduling the re-archiving of selected content within the one collection without de-duplication mechanisms, a file level comparison is perhaps less useful than a host level comparison. An interesting observation is that there is in the order of three times the amount of content in the PANDORA Archive that is derived from non .au host domains as is derived from .au host domains. However much of this content is likely to have a very small ratio of files to hosts since it will represent images and other non text content picked up from embedded page links in the harvesting process.

The number of unique .au hosts found in the PANDORA Archive represents 0.8% of those found in the domain harvest; and even when comparing all hosts in PANDORA, this represents only around 3.5% of those found in the domain harvest. Taking into consideration that somewhere in the order of 20% of the domains in the domain harvests are non .au domains the percentages represented by PANDORA host domains in respect to those in the domain harvest are still around 1% (.au host domains) and 4.3% (all host domains).

Figure 3 - comparison of hosts in Australian domain harvests and the PANDORA Archive



3 Number and percentage of hosts by 2nd level domain

The following tables and graph show the relative representation of hosts in the domain harvest collections analysed by the top seven second level domain. In all three harvests more than 80% of hosts belong to the .com.au second level domain.

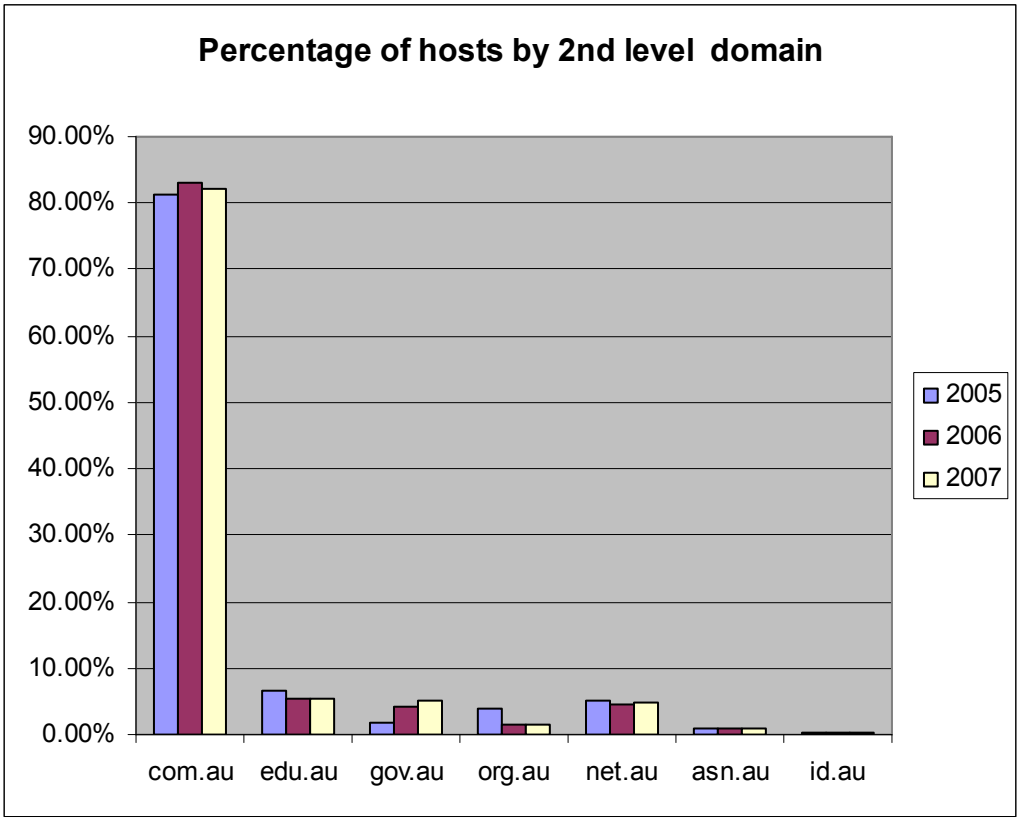
Table 3 – Number of hosts by 2nd level domain

Domain	Hosts		
	2005	2006	2007
com.au	273,066	480,529	480,469
edu.au	21,827	30,888	31,398
gov.au	5,924	25,203	29,314
org.au	13,504	8,574	9,095
net.au	16,979	26,005	27,705
asn.au	2,554	4,474	4,662
id.au	1,322	1,962	2,047

Table 4 – Percentage of hosts by 2nd level domain

Domain	Hosts		
	2005	2006	2007
com.au	81.40%	83.14%	82.13%
edu.au	6.51%	5.34%	5.37%
gov.au	1.77%	4.36%	5.01%
org.au	4.03%	1.48%	1.55%
net.au	5.06%	4.50%	4.74%
asn.au	0.76%	0.77%	0.80%
id.au	0.39%	0.34%	0.35%

Figure 4 - Percentage of hosts by 2nd level domain



4 Number and percentage of URLs by 2nd level domain and response code status

The URL counts used in this analysis are only those for which a 200 status response code was returned (i.e. a successful GET request was fulfilled). The percentages represented by the top reported response codes from the three domain harvests are given in the following table.

Table 5 – Percentage of top HTTP response status codes returned in all harvests

Response code	2005	2006	2007
200 (OK: found and returned)	88.57%	87.42%	86.27%
302 (temporary redirect)	6.26%	7.69%	7.75%
404 (not found)	5.78%	7.08%	5.58%
400 (bad request)	0.31%	0.80%	0.62%
301 (moved permanently)	0.53%	0.40%	0.53%
500 (internal server error)	0.37%	0.41%	0.35%
403 (forbidden)	0.15%	0.26%	0.14%
401 (authorisation required)	0.24%	0.14%	0.12%
303 (response under other URI)	0.00%	0.01%	0.03%
503 (service unavailable)	0.03%	0.04%	0.02%
414 (request URI too long)	0.02%	0.01%	0.00%
204 (no content)	0.01%	0.01%	0.01%

The com.au second level domain constitutes the major portion of the domain harvest URLs by a very large percentage. The 2005 harvest had some scope weighting towards the gov.au and the edu.au domains and in this harvest the com.au domain registered just below 70%. In the 2006 and 2007 harvests the percentage of com.au URLs in the harvest is just over 70%. The edu.au domain is the second most prominent with just under 10% in the 2006 and 2007 harvests. The gov.au domain shows some evidence of a small growth in the percentage (1.3%) of the harvest it constitutes.

Table 6 – Number of URLs by 2nd level domain

Domain	URLs		
	2005	2006	2007
com.au	92,116,410	261,208,879	311,519,044
edu.au	15,849,405	33,369,094	39,609,285
gov.au	12,894,393	22,493,055	32,858,864
org.au	5,582,075	18,521,229	24,395,200
net.au	4,801,907	15,380,412	17,780,849
asn.au	1,554,214	3,250,447	3,474,140
id.au	552,190	2,271,808	2,222,004

Table 7 – Percentage of URLs by 2nd level domain

Domain	URLs		
	2005	2006	2007
com.au	68.83%	73.11%	72.01%
edu.au	11.84%	9.34%	9.16%
gov.au	9.63%	6.30%	7.60%
org.au	4.17%	5.18%	5.64%
net.au	3.59%	4.30%	4.11%
asn.au	1.16%	0.91%	0.80%
id.au	0.41%	0.64%	0.51%

Figure 5 – Number or URLs by 2nd level domain

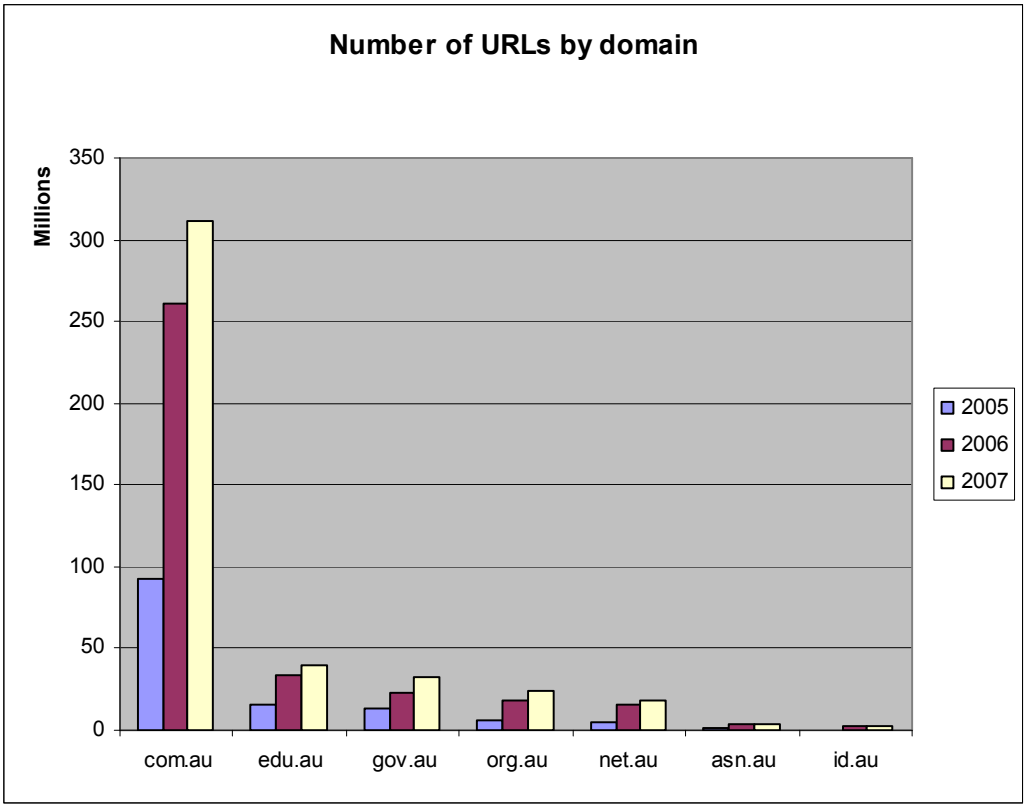
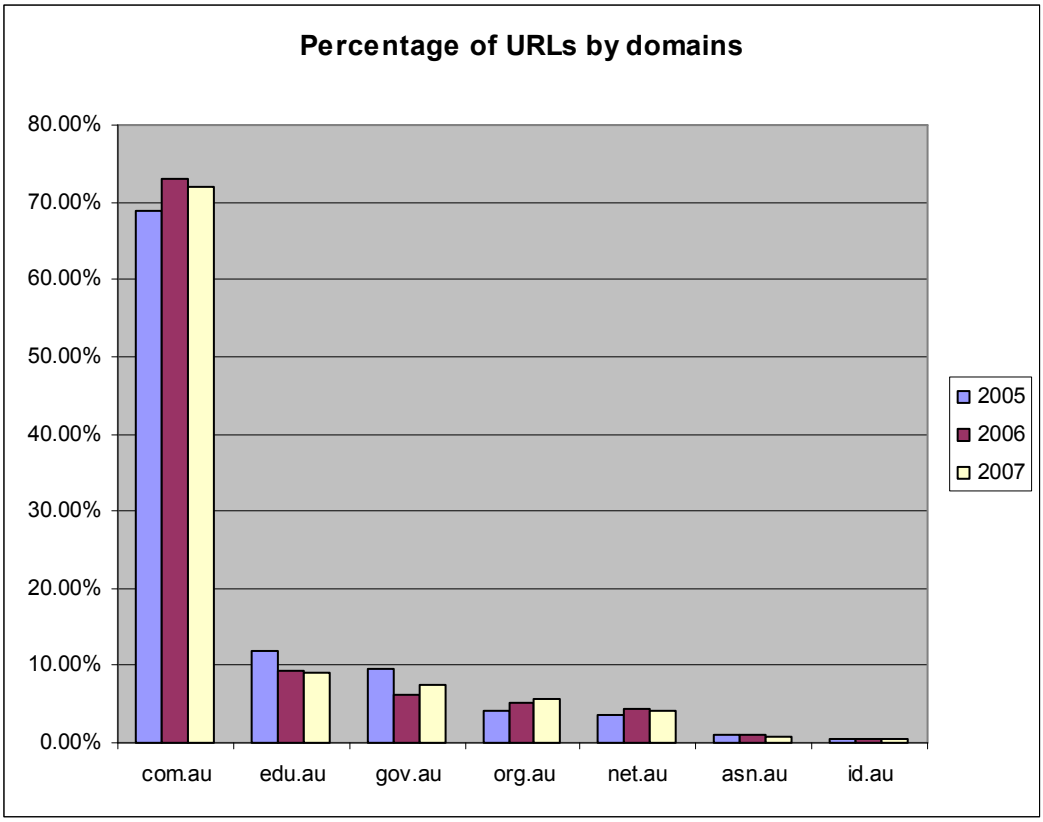


Figure 6 – Percentage of URLs by 2nd level domain



5 MIME types

Not surprisingly the most prominent MIME (Multipurpose Internet Mail Extension) type found in the harvest collections is the text/html MIME type representing as this does the textual pages of the Web, whether static HTML pages or dynamic generated content such as asp, php et al. Around three quarters of the files in the harvest are text/html, while image MIME types account for around 22% of the files and PDF for just 1.7% of the files.

The following tables do not include audio, video and Shockwave (Flash) MIME types. These media types represent a relatively small part of the harvested content. The MIME types associated with media files are less reliable and consistent than other formats such as text, images and PDFs and as a result it is difficult to be completely certain in the counting of these file types. Nevertheless in the combined total of files for all collections (2005, 2006 and 2007) audio MIME types represent around 0.0005% of files, video MIME types 0.0003% and the MIME type application/x-shockwave-flash constitutes 0.001% of files.

Table 10 and Figure 8 give a breakdown of text, image and PDF MIME types by second level domain. The info.au domain shows a noticeably higher percentage of PDF and image files than other domains while the id.au and com.au domains have the smallest percentage of PDF files. The gov.au domain has a fairly substantial percentage of PDF files (8.49%) but the edu.au

domain a little surprisingly only reveals 2.93% of files as PDF. This may perhaps be accounted for by the existence of institutional repositories in which much of the PDF output of the higher education sector will be located. The content of these repositories are for the most part not crawled by the harvester robot.

Table 8 – Number of files by main MIME types

	app'n/pdf	image/gif	image/jpeg	image/png	text/css	text/html	text/plain
2005	2,688,619	12,834,418	22,642,434	547,749	583,553	82,024,727	1,361,584
2006	6,628,094	24,468,976	60,358,321	1,768,620	1,118,150	307,828,555	1,616,618
2007	6,049,067	19,809,643	54,375,599	2,107,822	1,544,500	291,398,577	1,531,772
All years	15,365,780	57,113,037	137,376,354	4,424,191	3,246,203	681,251,859	4,509,974

Table 9 Percentage of files by main MIME types

	app'n/pdf	image/gif	image/jpeg	image/png	text/css	text/html	text/plain
2005	2.19%	10.46%	18.46%	0.45%	0.48%	66.86%	1.11%
2006	1.64%	6.06%	14.95%	0.44%	0.28%	76.24%	0.40%
2007	1.61%	5.26%	14.43%	0.56%	0.41%	77.33%	0.41%
All years	1.70%	6.32%	15.21%	0.49%	0.36%	75.42%	0.50%

Figure 7 – Percentage of files by main MIME types (2005-2007 data)

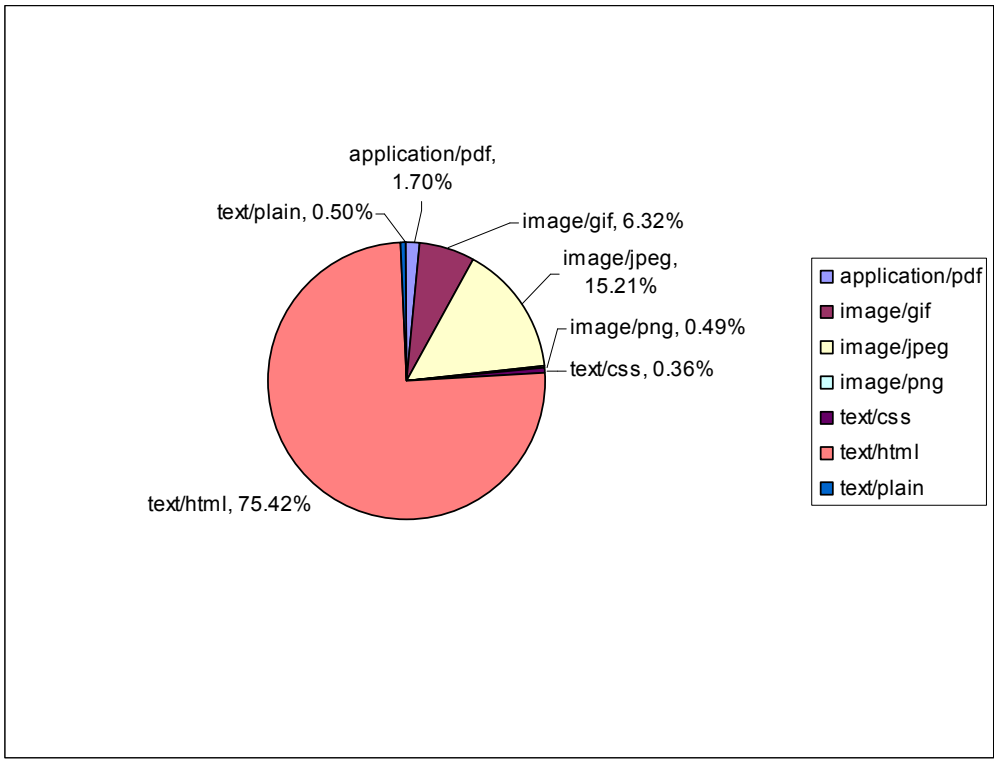
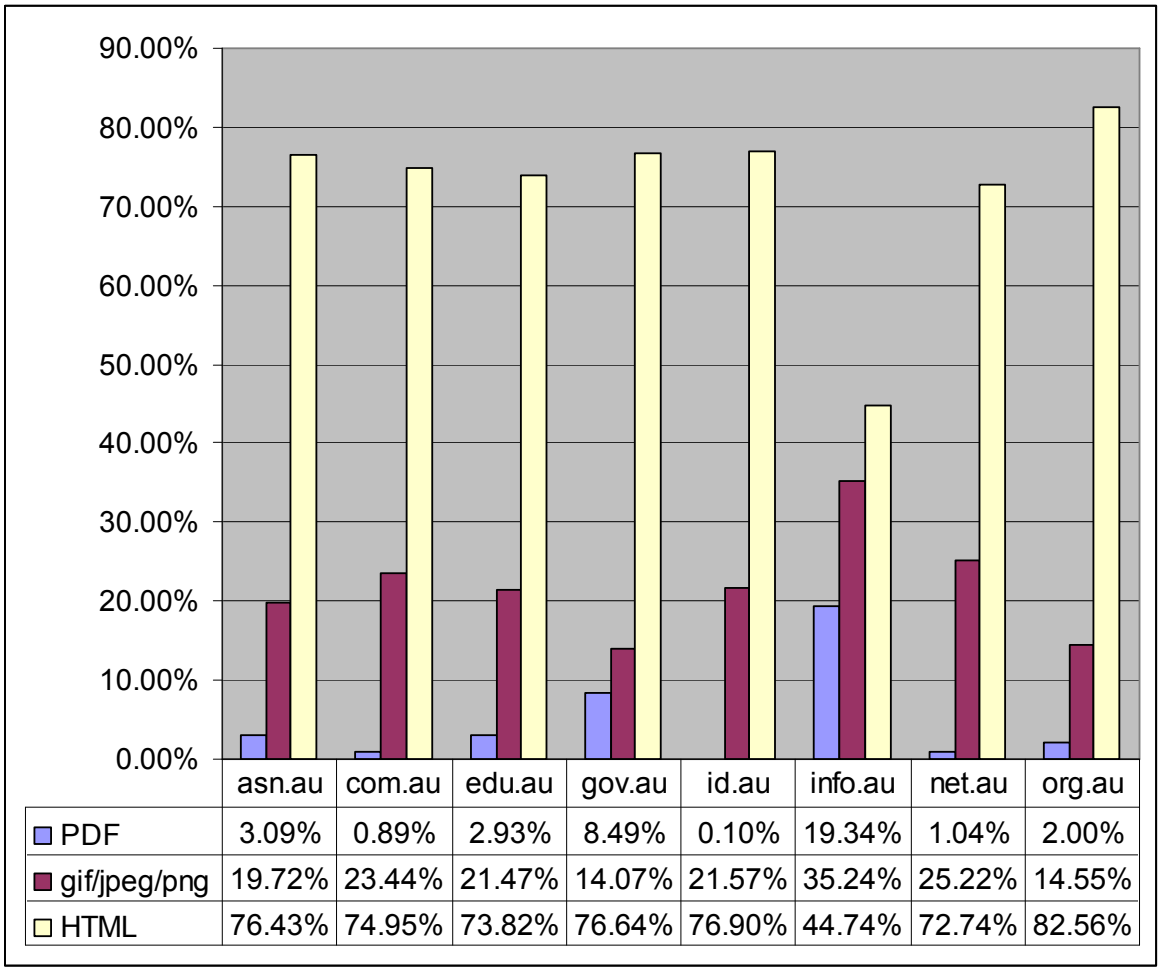


Table 10 – Main MIME types by 2nd level domain (2005-2007)

	app'n/pdf	image/gif	image/jpeg	image/png	text/css	text/html	text/plain
asn.au	3.09%	5.16%	14.16%	0.40%	0.35%	76.43%	0.41%
com.au	0.89%	6.31%	16.78%	0.35%	0.36%	74.95%	0.35%
edu.au	2.93%	7.77%	12.20%	1.49%	0.38%	73.82%	1.40%
gov.au	8.49%	6.80%	6.92%	0.35%	0.42%	76.64%	0.37%
id.au	0.10%	2.31%	18.56%	0.70%	0.14%	76.90%	1.28%
info.au	19.34%	8.41%	25.63%	1.20%	0.55%	44.74%	0.12%
net.au	1.04%	7.07%	17.52%	0.63%	0.38%	72.74%	0.63%
org.au	2.00%	3.96%	10.11%	0.48%	0.27%	82.56%	0.61%
asn.au	3.09%	5.16%	14.16%	0.40%	0.35%	76.43%	0.41%

Figure 8 – PDF, image and HTML MIME types by 2nd level domain (2005-2007 data)



6 Changes between the 2006 and 2007 harvests

As the National Library has now acquired three large scale domain harvest of the Australian web domain, some comparison of the data sets is possible. The figures presented here focus mainly on a comparison of data from the 2006 and 2007 harvests as the scope of those two harvests was the same (based on a target of 500 million unique URLs). The 2005 harvest was a time-based crawl with a smaller data set so a comparison based on numbers is less valid; however the 2005 harvest has been included in the percentage comparison table and graphs below.

The comparative figures suggest a high rate of change at the URL level (more than 50%) with slightly more content added than removed between the 2006 and 2007 harvests. The rate of change at the host level is considerably less at around 5-6%. When analysed at the domain level there is evidence that more content is being added than removed at double the rate in the edu.au and gov.au domains and considerably more in the net.au and org.au domains. The org.au domain shows nearly 10 times the rate of added content over removed content. The com.au domain shows a more even ratio of added and removed content suggesting perhaps that content is replaced or changed rather more than in other domains where the figures suggest more content being added and less removed.

Table 11 URLs and hosts added/removed/unchanged between 2006 and 2007 harvests

	.au Hosts	URLs
Added	83,282	334,665,044
Removed	65,763	313,430,438
Unchanged	317,092	112,728,584

Table 12 Percentage of URLs and hosts added/removed/unchanged between all harvests

	.au Hosts		URLs	
	2005 to 2006	2006 to 2007	2005 to 2006	2006 to 2007
Added	8.08%	6.68%	62.91%	64.85%
Removed	5.14%	5.22%	63.62%	52.57%
Unchanged	35.26%	25.14%	27.51%	18.91%

Figure 9 Percentage of .au hosts added/removed/unchanged

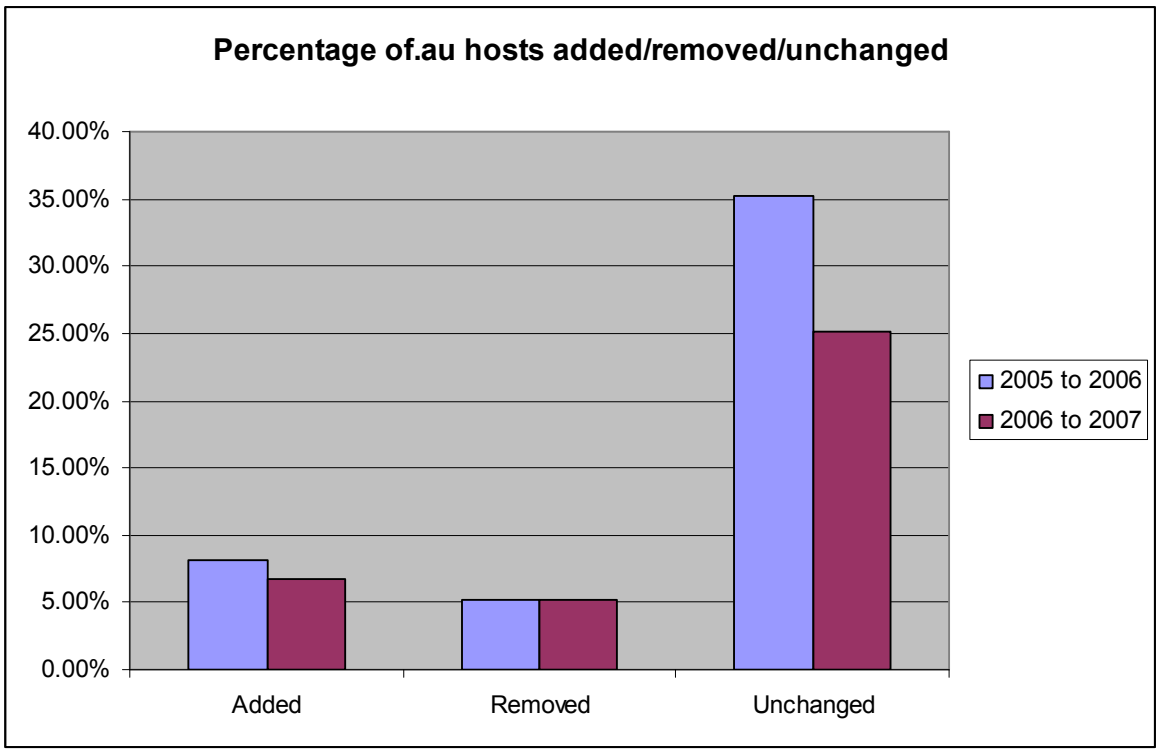


Figure 10 - Percentage of URLs added/removed/unchanged

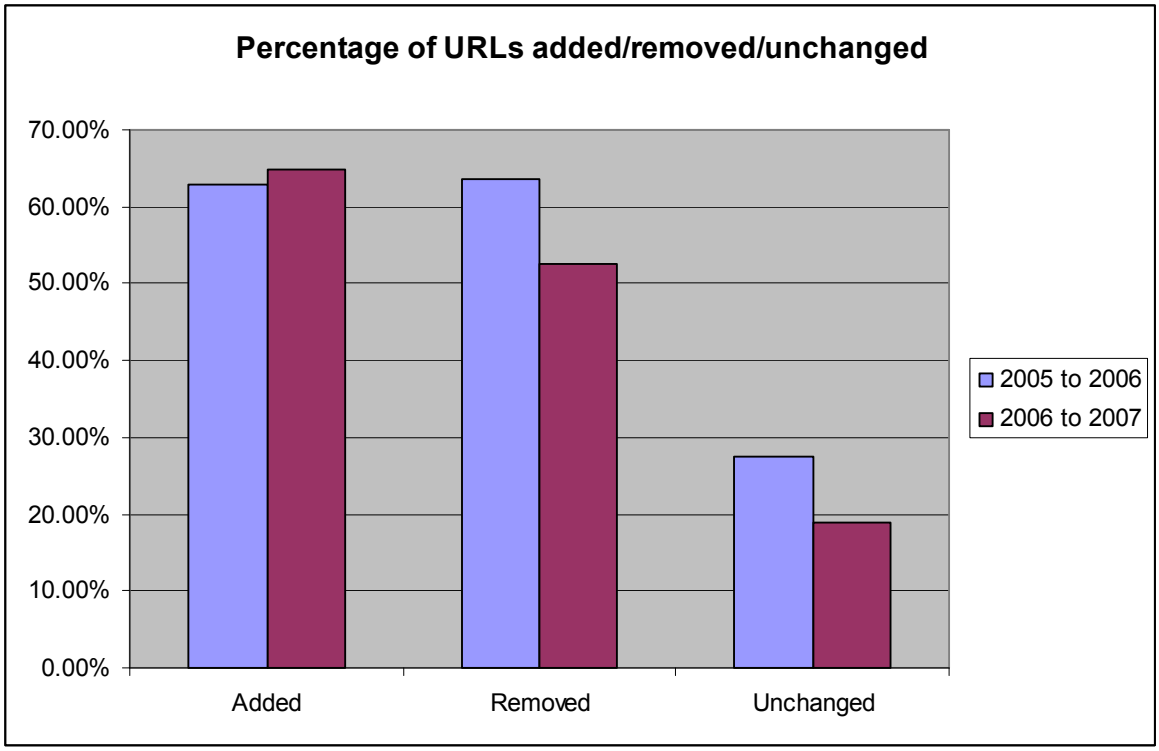


Table 13 URLs added/removed between 2006 and 2007 harvests by 2nd level domain

	Added		Removed	
	Number of URLs	% of total domain	Number of URLs	% of total domain
com.au	40,753,051	13.08%	33,497,107	12.82%
edu.au	3,812,010	9.62%	1,579,410	4.73%
gov.au	2,207,624	6.72%	773,410	3.44%
org.au	9,483,210	38.87%	750,184	4.05%
net.au	2,588,785	14.56%	901,793	5.86%

Figure 11 - Percentage of URLs added/removed by 2nd level domain between 2006 and 2007