

# New Zealand Web Harvest 2010

## Options Paper

20 January 2010



## Table of Contents

1. Introduction .....	3
2. What we want to do and why .....	4
3. Reviewing the first New Zealand Web Harvest .....	4
4. Results of the New Zealand Web Harvest 2008.....	5
5. Harvesting issues and options .....	6
5.1. Notification.....	6
5.2. Robots Policy.....	7
5.3. Location of harvester .....	8
5.4. Other issues .....	9
6. Feedback.....	9
7. Appendix: Selected details of the New Zealand Web Harvest 2008 .....	10
7.1. Overview of harvest size.....	10
7.2. Size of harvest by top-level domain .....	10
7.3. Size of harvest by New Zealand second-level domain .....	11
7.4. Harvest statistics for New Zealand domains .....	12
7.5. Level of traffic per host.....	13
7.6. HTTP response codes .....	13
7.7. MIME media types .....	14
7.8. Robots exclusion .....	14

## 1. Introduction

The National Library of New Zealand is planning a New Zealand Web Harvest in April 2010.

This options paper is being circulated to consult stakeholder groups in the networking and technical communities on concerns raised during the first harvest in October 2008, specifically the notification period, the robots policy, and the location of the harvester. Feedback on this paper will be built into the planning of the harvest, its public notification, and its operation.

**Important note:** This options paper is **NOT** a general notification of the 2010 harvest. This document is written to distribute to a technical audience to gather feedback on the way in which the harvest will be implemented. The specific details and the date of the harvest are not yet known. We will notify the wider New Zealand internet community closer to the date of the harvest when more details are known.

If you have feedback on the options presented in the document, or any other advice to the library about the 2010 web harvest, please send it to **web-harvest-2010@natlib.govt.nz** by 9am Monday 8 February 2010.

You can give your feedback in person to Gordon Paynter at the New Zealand Network Operators Group Conference 2010 (<http://2010.nznog.org/>) in Hamilton on 28 and 29 January 2010.

## 2. What we want to do and why

The National Library of New Zealand has a social responsibility to preserve New Zealand's social and cultural history, be it in the form of books, newspapers and photographs, or of websites, blogs and videos. The planned New Zealand Web Harvest 2010 harvest recognises the importance of the internet in all areas of New Zealand society and culture by taking a 'snapshot' of the entire .NZ domain as it exists on the web in April 2010.

The National Library was given a mandate, and a responsibility, to undertake this work by the National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003 (<http://legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html>). To meet these obligations, the Library has an ongoing programme of selective web harvesting of high value websites, and conducted a first domain harvest in October 2008. The 2010 harvest will complement these harvesting activities, and we plan to continue both domain and selective harvests in future years.

Our proposed timeline for the 2010 web harvest is:

- January – Initial consultation with stakeholder groups.
- February – Technical planning.
- March – Communications and notifications about the upcoming harvest.
- April – The harvest.

## 3. Reviewing the first New Zealand Web Harvest

The Library undertook its first large-scale web harvest over a period of ten days in October 2008, eventually collecting 4 terabytes of data from over 100 million URLs. We are pleased with the result of the 2008 harvest, but acknowledge problems with its execution affected the wider New Zealand internet community, and apologise for any inconvenience it caused.

We are consulting with key technical and networking stakeholders on options to address the concerns raised during the 2008 harvest, specifically:

- Notification: The harvest was initiated without prior notification to affected parties.
- Robots policy: The harvester was configured to ignore the robots.txt convention unless the website owner contacted the Library to request that it be honoured.
- Location of the harvester: The harvest was operated by the Internet Archive from the United States, and some website owners are charged more for international traffic.

We are working hard to improve our communications, and hope to work with site owners, administrators and other stakeholders to lessen the impact of our harvesting activity this year. Feedback on this options paper will be built into the planning of the harvest, its public notification, and its operation.

## 4. Results of the New Zealand Web Harvest 2008

The first New Zealand web harvest ran for ten days in October 2008 (followed by a smaller “patch crawl” to fill gaps). 106 million URLs were requested, and 4.6 terabytes of data were downloaded. The data is securely stored at the Library, and we are working on providing appropriate access, and on related policy and legal issues.

The tables below provide a few key statistics from the 2008 harvest. See the Appendix for details.

Statistic	October 2008 harvest
URLs requested	106,184,620
Data downloaded (bytes)	4,566,562,591,555
Hosts discovered	397,101
Average requests per host	267
Average data downloaded per host (bytes)	11,499,751

Approximately 91% of the URLs requested, data downloaded, and hosts discovered were from the .NZ top-level domain (with 6% from .com and no other top-level domains reaching more than 1%).

The number of URLs requested and amount of data downloaded was usually quite small: for 81% of hosts less than a megabyte of data was downloaded. Only 22 hosts provided over 10 gigabytes.

Number of HTTP requests	Number of hosts	Percent of hosts	Data downloaded	Number of hosts	Percent of hosts
1-9	270,598	68%	<1,000,000 bytes	322,951	81.3%
10-99	69,161	17%	1 to 10 MB	43,226	10.9%
100-999	40,619	10%	10 to 100 MB	22,082	5.6%
1000-9999	16,369	4%	100 to 1000 MB	8,365	2.1%
10000-99999	354	0.1%	1 to 10 GB	455	0.1%
<b>Total</b>	<b>397,101</b>	<b>100%</b>	10 to 100 GB	22	0.006%
			<b>Total</b>	<b>397,101</b>	<b>100%</b>

A controversial feature of the harvest was the decision to ignore robots.txt files. This table quantifies the harvested material affected by robots.txt (no qualitative analysis has been attempted).

Theoretical robots.txt result	Number of URLs	Percent of URLs
Access allowed	95,404,999	89.6%
Access disallowed	11,041,756	10.4%
Not checked	724	0.0%

A robots.txt file was found on 115,374 hosts (29% of the total).

## 5. Harvesting issues and options

This section outlines a series of options that we are considering to address the issues that were raised in the first domain harvest. In some cases the options are complementary, in other cases they are mutually exclusive.

We are seeking feedback on all the options presented, and any other solutions to the problems described. While we have started exploring the costs and benefits of the options, we have not yet made decisions on key technical issues such as our robots policy and the location of the harvester.

### 5.1. Notification

The National Library did not give sufficient warning of the 2008 harvest to site owners, administrators and other affected parties. In 2010, the Library will be making a concerted effort to alert these stakeholders to the harvest, its potential impact, and how that impact can be addressed.

We intend to use the following communications channels:

- **Option 1.1: Website.** The authoritative source of project information and media releases will be the Library website (<http://www.natlib.govt.nz>).
- **Option 1.2: Blog.** We propose to provide notifications and updates during the harvest through the LibraryTechNZ blog (<http://librarytechnz.natlib.govt.nz/>).
- **Option 1.3: Mailing lists.** We propose to post notifications to the NZNOG and NZ-WEB-DEV mailing lists.
- **Option 1.4: Social media.** We propose to operate a dedicated account on Twitter during the notification and harvest periods.
- **Option 1.5: 2008 correspondents.** We propose to notify individuals who provided feedback on the 2008 harvest directly by email.
- **Option 1.6: Email address.** The library will provide an email address for all feedback during the notification and harvest periods.
- **Option 1.7: Period of notification.** The Library proposes to begin formal notification four to six weeks in advance of the start of the harvest.

We are seeking feedback on the whether these channels are necessary and sufficient, whether there are other ways we could publicise the harvest, and whether the period of notification is sufficient.

## 5.2. Robots Policy

The robots.txt convention uses a text file to tell web “robots” about which pages they may and may not download from a host. Robots.txt files are usually used to guide search engines toward important content on the site (e.g. HTML pages) and away from other types of content (e.g. images, stylesheets, some dynamic content). However, an archival harvester has a different goal from a search engine harvester: preserving an accurate and complete representation of a website, including appearance and in many cases the dynamic content. Almost all archival harvesters ignore the robots.txt convention to a greater or lesser extent.

In the October 2008 harvest the Library decided to ensure that the best possible record of websites was captured so the robots policy was to ignore robots.txt unless we were contacted by the website owner. This is a relatively strong policy position by international standards, but is supported (and arguably required) by the National Library of New Zealand Act (2003). Many webmasters, particularly those who were using robots.txt to prevent robots from downloading large files or from falling into crawler traps, complained that this created an unnecessary burden in terms of webserver load and bandwidth.

At the time of the harvest, it was not known how much material was disallowed by robots.txt files in New Zealand (or internationally). Subsequent analysis of the October 2008 harvest shows that if the Library had honoured the robots.txt convention on every website, then 10.3% of the requested URLs would have been unavailable (i.e. 11 million of 106 million URLs). This is a quantitative measure and we have not yet done a qualitative evaluation of the value of the material concerned.

The Library is considering what robots policy to use for the 2010 harvest. The options include:

- **Option 2.1: Honour robots.txt strictly.** This will result in a less complete and less accurate harvest of many sites, but the Library can still expect to capture 90% of the URLs, and can potentially balance the missing content with a faster, deeper harvest. This option gives the webmasters the ability to control the harvester with targeted or default rules.
- **Option 2.2: Honour robots.txt partially.** Honour robots.txt except when downloading images and other elements that are embedded in other web pages. For example, if the harvester is allowed to download an HTML page on a website, then it would download all the supporting images and stylesheets regardless of their robots.txt status. This will probably result in a less complete harvest overall, but pages that are downloaded will be accurate representations. This option gives webmasters some control of the harvester with targeted rules.
- **Option 2.3: Most generous interpretation of rules.** Treat each URL as being allowed for the Library’s harvester if it is allowed for any robot. In other words, if a URL is disallowed for some robots but allowed for one or more robots, then we treat it as allowed for our harvester. This option gives webmasters limited control of the harvester.
- **Option 2.4: Ignore robots.txt selectively.** If the robots.txt has rules specifically for our harvester then follow those rules; if it does not then ignore robots.txt altogether. This option gives webmasters a way to control the harvester if they create targeted rules, but will result in a complete and accurate harvest if they do nothing.
- **Option 2.5: Ignore robots.txt.** Continue the 2008 policy to ensure that the best possible record of websites is captured. This option gives webmasters no direct control of the harvester.

Note that the Library’s is constrained to using a robots policy that can be implemented by the Heritrix web harvester (version 1.14.2). All the options above can be implemented in Heritrix, and while we will consider other suggestions it may not be able to implement them.

We are seeking feedback on the relative desirability of the options above, and other potential options.

### 5.3. Location of harvester

The Library contracted the Internet Archive to perform the 2008 web harvest on its behalf, and intends to do the same for the 2010 harvest. The Internet Archive is an acknowledged leader in the field of web archiving, and in addition to running harvests for many other national libraries they maintain their own web archive, the largest in the world (<http://www.archive.org/web/web.php>).

The Internet Archive is based in California and performed the 2008 harvest from the United States. This had an unexpected side effect for websites hosted in New Zealand: website owners may be charged by volume of traffic, and may be charged more for international traffic than for national traffic. For a website owner in this situation, a web harvest from the United States will cost the website owner more than the same harvest in New Zealand.

This situation is apparently unique to the New Zealand internet, and has not been reported when the Internet Archive has performed domain harvests in other countries.

A second point to consider when harvesting from abroad is that some material on servers in New Zealand is only made available to clients that request it from within New Zealand. A harvest from the United States will not gather this material.

The Library is considering where to locate the 2010 harvest. The options include:

- **Option 3.1: Harvest from the United States.** Continue to use the Internet Archive servers in the United States. This is the most convenient option, but is likely to have the same cost implications for websites that are charged more for international traffic as the 2008 harvest.
- **Option 3.2: Harvest from New Zealand.** Purchase or hire hardware in New Zealand, and locate it at the Library or another New Zealand facility for the duration of the harvest. This option is more complex to organise and more expensive (for the Library), but it directly addresses the concerns of websites that are charged more for international traffic. An advantage is that material that is only available to New Zealand clients can be harvested.
- **Option 3.3: Route harvest from the United States through a New Zealand server.** While this is an elegant solution that appears to offer the best of both worlds, our advice from the Internet Archive engineers (and their peers in New Zealand) is that it is not technically feasible.

We are seeking feedback on the options above, and particularly on the implications of locating the harvester at different network locations within New Zealand.

## 5.4. Other issues

This section lists a few other changes we are considering for the 2010 harvest. We are seeking general feedback on these topics.

### Zone files

For the 2008 harvest, we did not seek a copy of the .nz zone file. Instead we built a seed list based on our previous harvesting experience (and that of the internet archive). It included all the .nz hosts we had previously encountered, plus a few hundred hosts from domains other than .nz selected on a relatively ad-hoc basis.

- Option 4.1: We **have applied** to the Domain Name Commissioner for a copy of the .nz zone file to build a complete seed list.
- Option 4.2: We **are** speaking to the commissioner about updating the appropriate .nz policies and procedures to alert registrants to the possibility of a harvest.
- Option 4.3: We **may** automatically scan the .com, .org and .net zone files to find hosts that are physically located in New Zealand and add them to the seed list.

### Depth and speed of harvest

In the 2008 harvest we performed a relatively shallow harvest, requesting a maximum of 50,000 URLs from government and education websites; and 10,000 URLs from other websites. Most websites responded with many fewer URLs (see Appendix). We waited a minimum of 5 seconds between requests to a specific host.

- Option 4.4: We **may** increase the maximum depth limits, causing much deeper harvests of large sites.
- Option 4.5: We **will** wait 5 seconds between requests to a specific host..

### Host-based queuing

The Heritrix harvester works best when it treats each host independently. However, in some cases a single server can host dozens or hundreds of virtual hosts. If several of these virtual hosts are harvested at the same time it can put substantial load on the server.

This situation arose two or three times in 2008, and when the webmaster contacted us we configured the harvester to reduce the load.

This year we will continue host-based harvesting, but will take some actions to mitigate its effects:

- Option 4.6: We **will** contact any webmaster that reported this problem in 2008 to prevent a recurrence.
- Option 4.7: We **may** perform an analysis of the domains to try and spot servers with many hosts, for example those that resolve to the same IP address.

## 6. Feedback

If you have feedback on the options presented in the document, or any other advice to the Library about the 2010 web harvest, please send it to **web-harvest-2010@natlib.govt.nz** by 9am Monday 8 February 2010.

You can give your feedback in person to Gordon Paynter at the New Zealand Network Operators Group Conference 2010 (<http://2010.nznog.org/>) in Hamilton on 28-29 January 2010.

## 7. Appendix: Details of the New Zealand Web Harvest 2008

This section includes information describing the October 2008 harvest, including statistics about the quantity of data downloaded, and about the type of data downloaded.

### 7.1. Overview of harvest size

This table provides an overview of the harvest size.

Statistic	October 2008 harvest
URLs requested	106,184,620
Data downloaded (bytes)	4,566,562,591,555
Hosts discovered	397,101
Average requests per host	267
Average data downloaded per host (bytes)	11,499,751

The number of hosts discovered excludes 215,890 invalid hosts (i.e. references to hosts that no longer exist or that never existed) and 247,926 invalid requests (mostly requests to invalid hosts). These invalid hosts and requests are excluded from all the statistics that follow.

### 7.2. Size of harvest by top-level domain

This analysis breaks down key statistics by top-level domain.

TLD	Hosts	Requests	Downloaded data (bytes)	Hosts (%)	Requests (%)	Bytes (%)
nz	363,279	96,258,398	4,138,066,218,127	91.5%	90.7%	90.6%
com	22,766	6,531,839	249,784,849,731	5.7%	6.2%	5.5%
au	2,484	936,087	35,054,353,545	0.6%	0.9%	0.8%
net	2,010	634,145	32,637,354,431	0.5%	0.6%	0.7%
org	1,641	728,240	31,542,328,343	0.4%	0.7%	0.7%
uk	667	118,741	6,102,303,861	0.2%	0.1%	0.1%
us	547	30,221	1,455,563,431	0.1%	0.0%	0.0%
biz	320	175,405	4,778,117,873	0.1%	0.2%	0.1%
IP Address	309	157,175	9,363,425,009	0.1%	0.1%	0.2%
info	288	186,839	6,412,521,919	0.1%	0.2%	0.1%
edu	243	3,962	273,604,108	0.1%	0.0%	0.0%
Other	2,547	423,568	51,091,951,177	0.6%	0.4%	1.1%
<b>Total</b>	<b>397,101</b>	<b>106,184,620</b>	<b>4,566,562,591,555</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

### 7.3. Size of harvest by New Zealand second-level domain

This analysis is restricted to data from hosts in the .nz top-level domain, and breaks down key statistics by second-level domain.

TLD	Hosts	Domains	Requests	Downloaded data (bytes)	Hosts (%)	Domains (%)	Requests (%)	Bytes (%)
ac	8,178	735	5,377,505	373,396,865,733	2%	1%	6%	9%
co	304,261	109,391	70,805,905	2,499,734,031,548	84%	87%	74%	60%
cri	87	10	53,182	4,422,762,951	0%	0%	0%	0%
geek	496	208	357,250	17,102,736,753	0%	0%	0%	0%
gen	984	336	422,625	24,134,943,612	0%	0%	0%	1%
govt	2,015	687	6,155,062	463,795,657,883	1%	1%	6%	11%
iwi	86	41	29,690	986,532,968	0%	0%	0%	0%
maori	240	116	236,725	8,340,406,702	0%	0%	0%	0%
mil	27	11	32,950	2,094,312,856	0%	0%	0%	0%
net	24,756	4,470	4,421,768	279,436,229,178	7%	4%	5%	7%
org	18,394	8,437	6,881,733	377,088,145,829	5%	7%	7%	9%
parliament	5	3	9,952	1,304,339,547	0%	0%	0%	0%
school	3,750	1,821	1,474,051	86,229,252,567	1%	1%	2%	2%
<b>Total</b>	<b>363,279</b>	<b>126,266</b>	<b>96,258,398</b>	<b>4,138,066,218,127</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

#### 7.4. Harvest statistics for New Zealand domains

This analysis is restricted to data from hosts in the .nz top-level domain, and shows the number of requests and data supplied.

<b>TLD</b>	<b>Domains</b>	<b>Hosts per domain</b>	<b>Requests per domain</b>	<b>Downloaded data per domain (bytes)</b>
ac	735	11.1	7,316	508,022,947
co	109,391	2.8	647	22,851,368
cri	10	8.7	5,318	442,276,295
geek	208	2.4	1,718	82,224,696
gen	336	2.9	1,258	71,830,189
govt	687	2.9	8,959	675,102,850
iwi	41	2.1	724	24,061,780
maori	116	2.1	2,041	71,900,058
mil	11	2.5	2,995	190,392,078
net	4,470	5.5	989	62,513,698
org	8,437	2.2	816	44,694,577
parliament	3	1.7	3,317	434,779,849
school	1,821	2.1	809	47,352,692
<b>Average</b>		<b>2.9</b>	<b>762</b>	<b>32,772,609</b>

### 7.5. Level of traffic per host

This table shows the number of requests made to each host. For example, 68% of hosts were asked for fewer than 10 pages.

Number of requests made	Number of hosts	Percent of hosts
1-9	270,598	68%
10-99	69,161	17%
100-999	40,619	10%
1000-9999	16,369	4%
10000-99999	354	0.1%
<b>Total</b>	<b>397,101</b>	<b>100%</b>

This table shows the quantity of data downloaded from each host. For example, 83% of hosts provided fewer than 1 megabyte of data.

Data downloaded	Number of hosts	Percent of hosts
< 1MB	322,951	81.3%
1 to 10MB	43,226	10.9%
10 to 100MB	22,082	5.6%
100 to 1000MB	8,365	2.1%
1 to 10 GB	455	0.1%
10 to 100 GB	22	0.006%
<b>Total</b>	<b>397,101</b>	<b>100%</b>

Note: 1MB = 1,000,000 bytes and 1GB = 1,000,000,000 bytes.

### 7.6. HTTP response codes

The table below lists the HTTP response codes most frequently encountered during the harvest.

HTTP response code	Number of HTTP requests	Percent of HTTP requests
200	93,602,392	87.9%
302	7,553,556	7.1%
404	2,727,667	2.6%
301	1,042,619	1.0%
400	478,832	0.4%
500	367,949	0.3%
403	366,412	0.3%
401	131,929	0.1%
303	121,203	0.1%
Other	54,969	0.1%
<b>Total</b>	<b>106,447,528</b>	<b>100%</b>

### 7.7. MIME media types

The table below lists the MIME types most frequently encountered during the harvest.

Mime type	Number of HTTP requests	Percent of HTTP requests
text/html	77,701,219	72.5%
image/jpeg	18,456,684	17.2%
image/gif	4,171,911	3.9%
application/pdf	1,521,757	1.4%
image/png	1,227,317	1.1%
text/dns	700,516	0.7%
text/plain	656,292	0.6%
Other	2,712,088	2.5%

### 7.8. Robots exclusion

This table quantifies the harvested material affected by robots.txt.

A robots.txt file was found on 115,374 hosts (29% of the total). For each URL, we analysed it against the relevant robots.txt file (as it was when it was harvested in October 2008) and determined whether access to the URL was allowed or disallowed by robots.txt.

In accordance with the protocol, if no robots.txt file was present we assumed that access was allowed.

Robots.txt result	Number of URLs	Percent of URLs
Access allowed	95,404,999	89.6%
Access disallowed	11,041,756	10.4%
Not checked	724	0.0%

No qualitative analysis of this material has been attempted.

Note: Statistics in this table are based on a different analysis than those above so the total number of URLs adds to a different figure because different methods were used to count URLs.