Submission to the National Library of New Zealand

on

the 2010 Web Harvest

8 February 2010

1. General

1.1 The mission of InternetNZ (Internet New Zealand Inc) is to protect and promote the Internet for New Zealand. We advocate the ongoing development of an open and uncaptureable Internet, available to all New Zealanders. InternetNZ is non-partisan and is an advocate for Internet, and related telecommunications, public and technical policy issues on behalf of the Internet Community in New Zealand - both users and the Industry as a whole.

1.2 InternetNZ welcomes this opportunity to provide comments on the New Zealand Web Harvest 2010 options paper.

1.3 InternetNZ supports the National Library's long-term mission to preserve New Zealand's online heritage. An increasing amount of information now resides solely on the Internet, much of it with important social and cultural value. It is essential that this information be preserved for future generations.

2. Notification

2.1 The October 2008 web harvest rankled many in New Zealand's technical community who were concerned at the lack of formal notification and the way in which the harvest was conducted.

2.2 InternetNZ is pleased that the National Library will be taking steps to improve communications with the wider technical community (including advising site hosts/owners and network operators) before archiving any information that resides on New Zealand's public Internet.

2.3 The communications channels outlined in the options paper are necessary, and should all be utilised so that webhosts are better able to plan and respond to the April 2010 and future harvests.

2.4 Particular use should be made of social networking/media and mailing lists to publicise details of the harvest, including the National Library's chosen robots.txt policy and harvester location, as widely as possible.

2.5 InternetNZ suggests that the National Library also communicate with technology media in advance of the harvest. This will help spread news of the harvest to those who don't participate in social media or mailing lists.

2.6 InternetNZ is happy with the proposed four to six-week formal notification period.

3. Robots.txt

3.1 Ignoring robots.txt entries is considered bad form in the web community. In many cases, robots.txt is applied for perfectly valid reasons i.e. to obscure URLs such as administration screens and website test pages not meant to be publicly accessible.

3.2 Blanket disregard of robots.txt could adversely affect the harvest. Some sites, for instance, are understood to implement countermeasures against unknown crawlers, believing them to be malicious. This may include the banning or blocking of IP addresses, a situation the National Library is likely to want to avoid.

3.3 InternetNZ suggests that all visible web content be harvested but that the National Library honour robots.txt entries that are specifically directed at its crawler (option 2.4). We recognise the limitations of this approach if website owners are arbitrary in their use of robots.txt, so if option 2.4 is untenable then option 2.3 is a suitable fall-back position.

3.4 Both these options strike a balance between the concerns of content owners who wish to keep some information private and the National Library's mission to capture as much content as possible. Coupled with a wide-ranging notification regime they should allay any serious concerns content owners have with respect to robots.txt.

4. Location of harvester

4.1 Given 73 percent of information harvested in 2008 came from sites hosted within New Zealand, and 26 percent from sites hosted overseas[1], InternetNZ prefers that the harvest be managed from within New Zealand (option 3.2).

4.2 An offshore harvest will force those majority content owners who host in New Zealand to pay for international bandwidth to return content. International traffic in New Zealand often costs content owners real money, especially smaller organisations that serve up large amounts of content.

4.3 InternetNZ also notes that some content will only be available if the harvest equipment is located within New Zealand. An offshore-initiated harvest will miss this potentially valuable content.

4.4 If the National Library does elect to conduct the harvest from offshore then InternetNZ suggests it advise the technical community of the IP ranges the harvester is operating in. This means that bandwidth-sparse operators can rate limit the harvest, lessening its bandwidth and financial impact.

5. Other issues

5.1 InternetNZ suggests that the National Library configure its harvester to also check IPv6 websites. The pool of remaining IPv4 addresses is nearing exhaustion and in future an increasing number of websites will be IPv6-only.

5.2 InternetNZ thanks the National Library for the chance to contribute its advice and suggestions on the Web Harvest options paper.

Regards,

[1] http://www.natlib.govt.nz/about-us/current-initiatives/web-harvest-2010

Vikram Kumar
CEO

For further information please contact:
Jordan Carter, Policy Director
+64 4 495 2118, jordan@internetnz.net.nz