STAFF PAPER

NATIONAL LIBRARY OF AUSTRALIA

Exposing the deep web to increase access to library collections

A paper presented at the Ausweb 2005 Conference

Tony Boston

Abstract

The National Library of Australia is making digital copies of special collection materials available over the Internet. About 100,000 collection items including pictures, maps, sheet music, manuscripts, and some books and serials have been made available online. This content is delivered dynamically from a database developed to manage the Library's digital collections. Since 2002 the Library has been exposing this content to Internet search engines to increase access to the material and provide multiple discovery pathways for Library users. This paper documents lessons learned in exposing the deep web and presents statistics on increased web usage focussing particularly on the Library's Pictures Collection. Application of technologies which can be used to share deep web content such as the Open Archives Initiative Protocol for Metadata Harvesting are also explored.

Keywords

Deep web; Search engine harvesting; Metadata harvesting; Digital collections; Digital libraries

Introduction

Many organisations are now delivering material onto the web served from backend databases via dynamic URLs. Such content is often not reachable by following hyperlinks as it may only be found via web search forms and hence is not indexed by Internet search engines such as Google. Such content has come to be known as the "deep" or "hidden" web [1] as it is not available through popular search engines and hence is not easily found by web users.

The magnitude of this problem has been highlighted in many studies. In 2001 the size of the deep web was estimated to be about 400 times that of the searchable web with over 7,500 terabytes of information [2]. The majority of content from cultural institutions as well as scientific, academic and government institutions is now delivered dynamically

from databases. This implies that a substantial amount of high quality and detailed content across a broad range of subject areas is not accessible via Internet search engines.

However, search engines remain the most popular way of finding information on the Internet. A 2003 New Zealand study of government online information [3] suggests that search engines are the most common method of finding government information on the Internet (41%), followed by going directly to the government department website (23%). This trend is even higher among young users, with over 60% of under 20 year olds and 52% of 20-29 year olds using search engines as their main pathway to government information resources.

The growth in traffic (data transfers) on the Internet is well documented. Annual rates of growth are typically between 50-150%. Data downloads by Australian Internet users rose by 60% between 2002 and 2003 and 130% between 2003 and 2004 (September quarters) [4]. Such general rises in usage are also reflected in the National Library's web site which has undergone an increase in page views of about 130% over the period July 2002 to May 2004. Figure 1 shows total National Library web site page views (excluding the Pictures Catalogue) for the period June 2002 to December 2004.

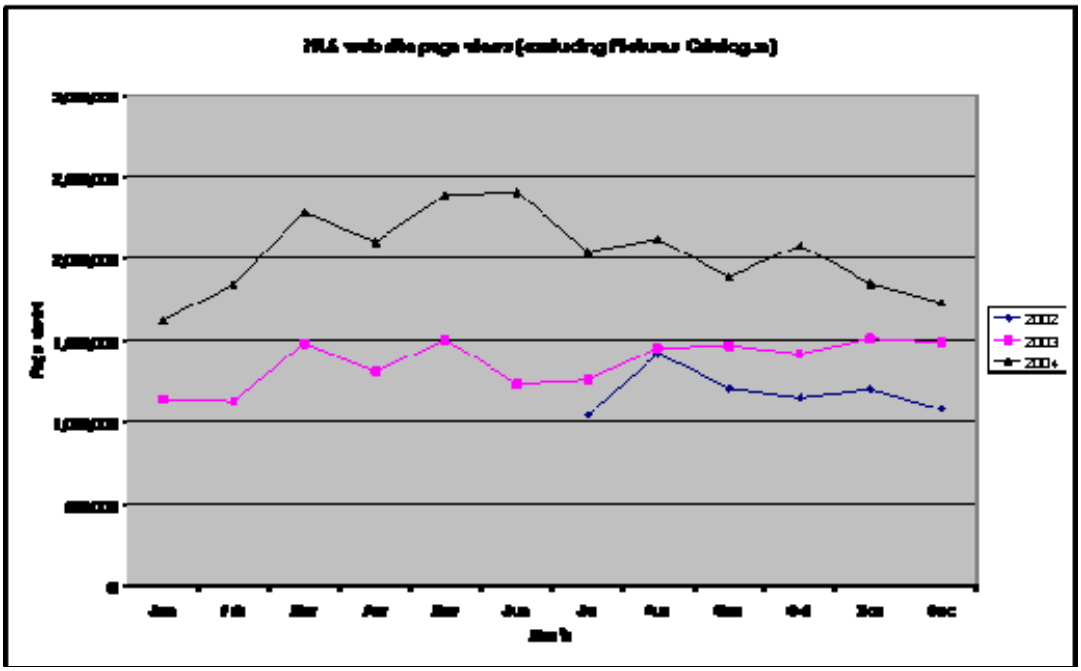

Figure 1: NLA web site page views from July 2002 to December 2004

This paper documents an increase in usage of the National Library's digital collections that is attributed directly to indexing by Internet search engines over and above the general increase in Internet usage seen in Figure 1. Statistics used are from web server access logs. Such logs are of limited use [5] and need to be interpreted with caution, but can be used to indicate broad trends in usage and source of accesses. All web server logs have been pre-processed to eliminate accesses due to search engine harvesting.

Exposing the deep web is highly relevant to libraries which traditionally attempt to make their collections as easily accessible to as wide a range of users as possible. Initiatives such as Google's plan to scan the full text of out of copyright items from major US and

European libraries as well as extracts from in copyright material [6] will increase the imperative for information providers to ensure their collections are "search engine accessible". And there is a growing expectation, especially among younger Internet users, that the high quality collections held by libraries and other cultural institutions should be accessible through Internet search engines.

Exposing the deep web

The National Library of Australia began digitising its Pictures Collection in 1996. Since 2001 the Library has increased the volume and coverage of material being digitised to also include maps, sheet music, manuscripts and some books and serials. To deliver online access to digitised items from the Pictures Collection, in 2001 the Library developed the Pictures Catalogue [7], which provides access to about 100,000 digitised paintings, drawings, prints and photographs. Users can search the catalogue by creator, title, subject and other attributes and display the metadata describing the item along with a JPEG image of the item. Figure 2 shows a photograph by Harold Cazneaux titled *18 footers racing Sydney Harbour* from the Pictures Catalogue. Note that the Library's digital collections are not full text, but include images of the collection item and descriptive metadata which is indexed for discovery purposes.

Figure 2: Cazneaux, Harold, 1878-1953. *18 footers racing Sydney Harbour*. 1921. http://nla.gov.au/nla.pic-an7831529-5

In November 2002, the Library began exposing its digital collections to Internet search engines like Google by creating lists of URLs for search engines to harvest [8]. Each URL resolves to a dynamically generated page as in Figure 2 that includes metadata describing a collection item along with an image of the item. The URLs for each digital collection item include a persistent identifier assigned to the item and make use of the Library's resolver

service to provide a stable web address for long term access [9]. The persistent identifier is stamped on the image within a standard footer and the persistent URLs for the image page and the image itself can be used for citation and reference purposes.

A separate list of URLs was created for each of the Library's digital collections, namely pictures, maps, sheet music, manuscripts, and books and serials. Each collection includes thousands of items which are listed on a series of web pages, each containing 100 links, that resolve to the collection items. These pages use the robot directive "noindex, follow" to direct search engine harvesters to follow links to content but not index the list pages. The URL lists themselves are also dynamically generated with new content automatically added to the list as new items are digitised and made available on the Internet. These lists are connected to the Library's digital collections infrastructure page [10] to provide a starting point for search engine harvesters. In Figure 3, the indexing pathway is represented in the top section of the diagram. Search engine harvesters follow the URL lists to index pages from the Pictures Catalogue database. The access pathway for library users is shown in the lower part of the diagram. After indexing by Internet search engines, an additional path is supported: one via the Library's web site and Pictures Catalogue search service and the new path through Internet search engines that access the digital collections via the Library's resolver service.

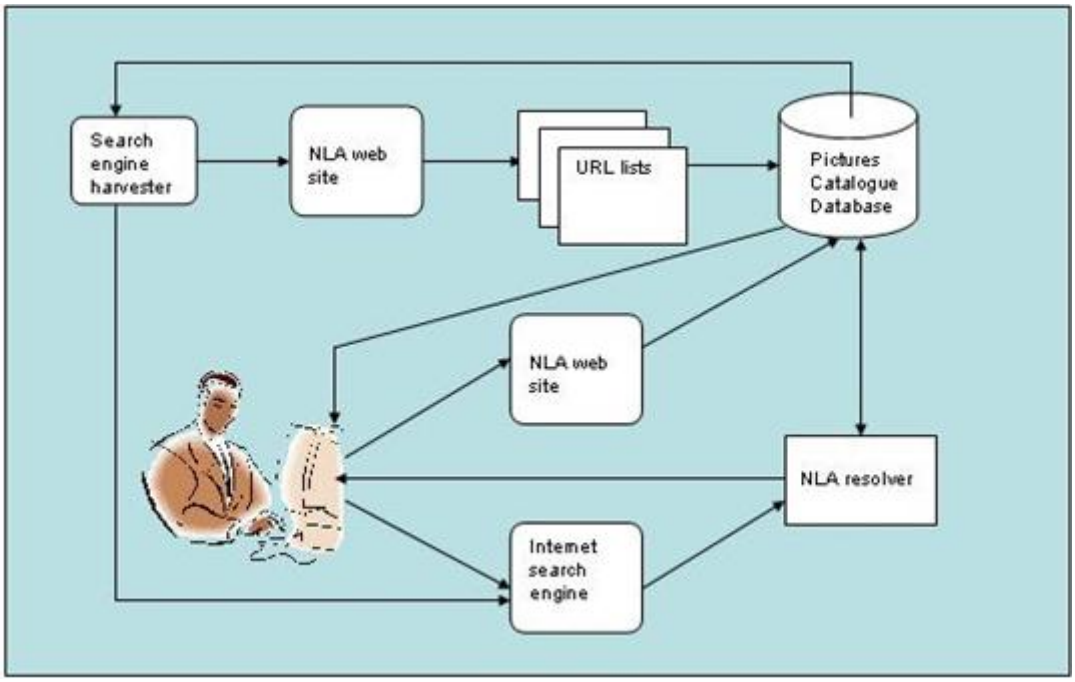

Figure 3: Search engine indexing and user access paths to the Pictures Catalogue

The URL lists were linked into the Library's web site and seed pages were submitted to the most popular search engines to encourage harvesting of these pages. Within one month of the release of these URL lists in November 2002, the Library was recording a marked increase in usage of its digital collections. Figure 4 shows page views of the Pictures Catalogue from July 2002 to December 2004. The increase in usage over the period July 2002 to May 2004 was 370% almost three times the 130% experienced by the NLA web site in the same period. An increase in usage can be seen in December 2002 which

correlates with increased access to the Library's digital collections from users of Internet search engines.

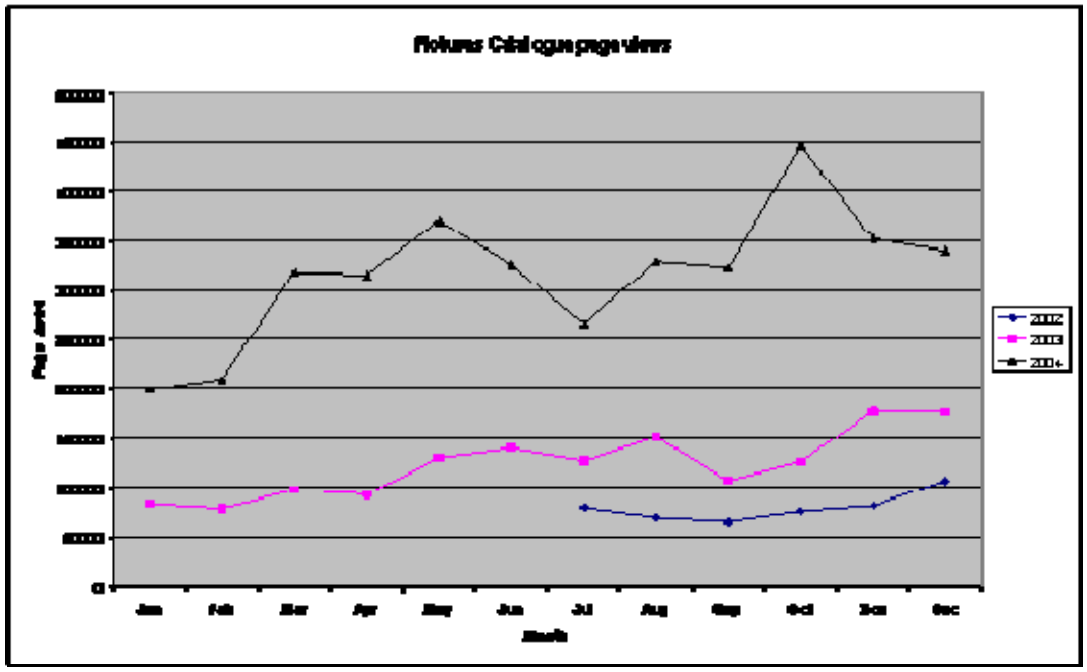

Figure 4: Pictures Catalogue usage from January 2002 to December 2004

Exposing the deep web to Internet search engines provides a new access path for users. Users of the National Library's web site can access the Pictures Collection via the Pictures Catalogue. Users of Internet search engines access these pictures via the Library's resolver service. Analysis of web server access logs can distinguish these two forms of access. Figure 5 shows the change in the number of Pictures Catalogue referrals for the period January 2002 to August 2003. During this period access to the Pictures Catalogue from the National Library's web site shows little change with accesses in 2002 and 2003 showing a remarkably parallel trend. In contrast the number of accesses via the resolver service shows a large increase from December 2002 as the Library's digital collections, including pictures, were exposed for harvesting by Internet search engines.

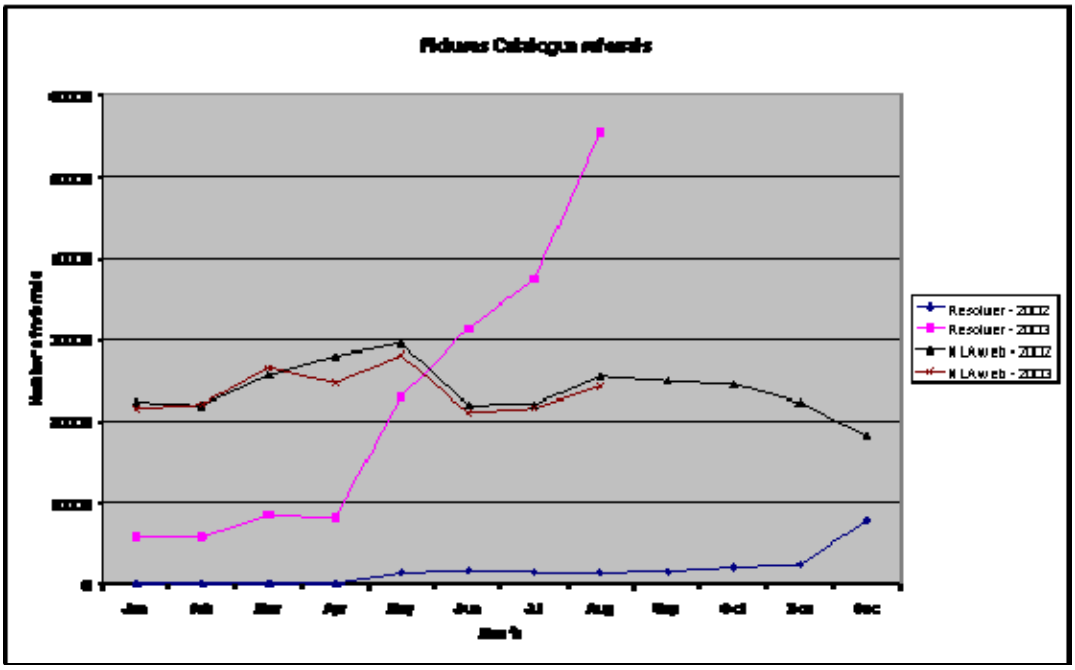

Figure 5: Pictures Catalogue referrals for January 2002 to August 2003

More pathways, more users

The National Library is committed to providing access to its collections through as many pathways as possible. In general terms there are now three major discovery pathways to the Library's collections: via the Library's catalogue; via federated/specialised discovery services such as Libraries Australia, PictureAustralia or MusicAustralia; and since late 2002, via publicly-accessible Internet search engines.

The statistics in Figure 5 suggest that the increased usage of the Pictures Catalogue is a result of increased access via the Library's resolver service. Analysis of referral data in resolver service web logs provides more information on the origin of these additional accesses. In December 2002 referrals from Google and Yahoo! increased from about 1-2% to about 20% and 10% respectively. Google is the major source of referrals in the resolver web access logs from December 2002 and made up 30-40% of accesses by mid 2003. Other major sources of referrals include search engines such as AOL Search and NineMSN as well as the National Library's catalogue and web site and discovery services such as PictureAustralia, Australia Dancing and the Australian Government's Culture and Recreation Portal.

Figure 6 shows the number of referrals originating from Google in the Library's resolver web logs for the years 2002 to 2004. A marked increase in referrals from Google (less visible due to the log scale) was seen in December 2002 from 300 to 4,000 referrals per month. This upward trend continued in 2003 rising to a peak of 67,000 referrals per month in August 2003 before stabilising at between 40-70,000 referrals per month in late 2003 through to November 2004.

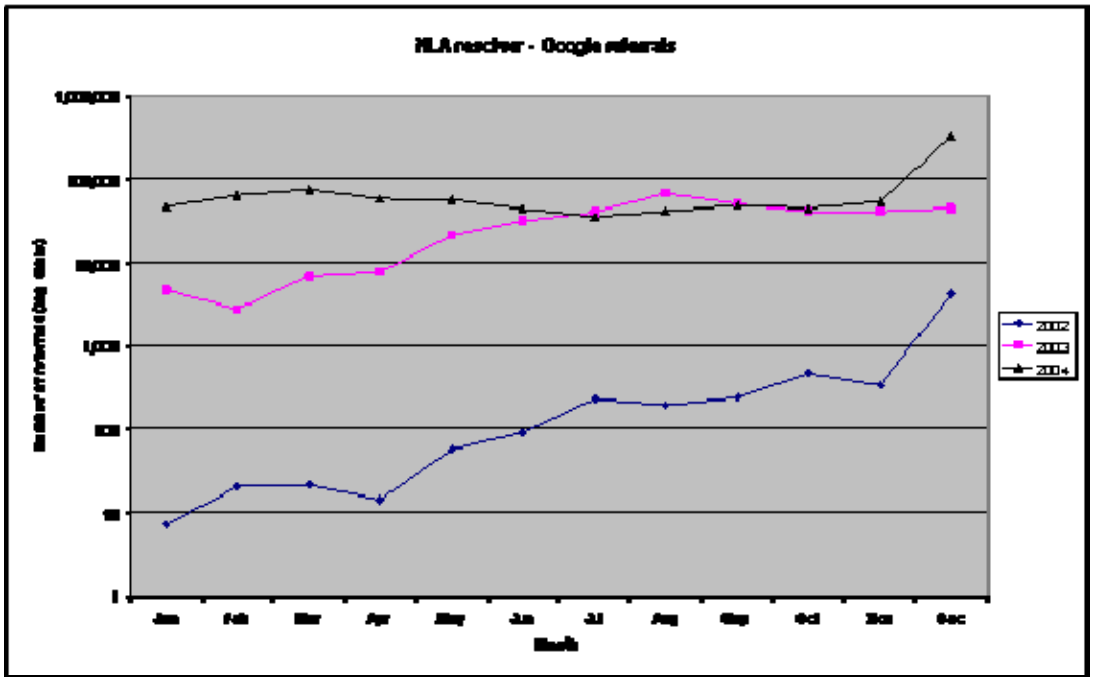

Figure 6: Google referrals in NLA resolver logs for 2002 to 2004

By the end of 2004, most of the larger Internet search engines had indexed some or all of the Library's digital collections content including Google, Yahoo!, AOL Search, Nine MSN, Ask Jeeves, AllTheWeb, Teoma, AltaVista, GigaBlast, LookSmart and Lycos. In making this content available to search engines, a number of lessons have been learnt as detailed below:

1. Implement persistent URLs. URLs to dynamically generated content need to be persistent. The National Library creates a simple persistent URL for its digital collections using a resolver service. An example URL which resolves to a page describing a Harold Cazneaux image of 18 footers on Sydney Harbour (Figure 2) is: http://nla.gov.au/nla.pic-an7831529-5. The resolver is essentially a re-direction service which enables the Library to change the technology used to deliver its digital collections without breaking the published persistent URL.

2. Ensure that the content to be harvested is not blocked via a robots.txt entry on your web site. Harvesting of areas of a web site can be enabled or disabled for different search engines by naming them in the robots.txt file via the user-agent field. The user-agent "GoogleBot" is the Google search engine harvester. Robot (search engine harvester) directives embedded in HTML pages such as noindex, follow can be used to control harvesting and which pages are indexed. Robots.txt exclusions can be used to block harvesters from following additional links once pages to be indexed have been captured. This prevents harvesting getting out of control and dynamic content being harvested more than once.

3. Ensure that the content to be harvested is accessible by following HTML links from the home page or a prominent page on your web site. The more prominent the link(s) to the content you want to be harvested, the more likely it is to be harvested. For content indexed by Google, more links from more pages measured by their

PageRank™ algorithm will lead to a higher ranking in results and inclusion of as much content as possible, measured via hypertext-matching analysis, will also increase a page's position in results [11]. PageRank™ and an analysis of page content seem to contribute about equally to a page's position in Google's results.

4. Create URL lists in a tree structure (hierarchy) of links where the pages to be indexed are the leaf nodes rather than simply as linked lists. A tree structure means that the harvester only has to go a few levels deep to get all of the content and this can be done very quickly. A hierarchy of N levels of A links allows fast access to A^N pages, so for example 3 levels of 100 links gives fast access to 100^3 or 1,000,000 pages. Linked lists where a harvester has to follow a long chain of "Next" links may not be able to be captured in full in the limited time available to the search engine harvester.

5. Some harvesters have a limit as to the number of links they will follow on one page. The Google harvester will only follow a maximum of 250 links per page.

6. Some database driven sites require cookies. Some search engine harvesters (eg GoogleBot) will not accept cookies. It is best to avoid cookies and session-ids when developing pages to be harvested. Session-ids can cause problems where the same content is given a different URL (via the session-id) which can lead to incomplete indexing or duplicate pages in some search engine indexes.

7. Search engine harvesting and indexing are independent steps. A lag of several weeks can occur between when pages are harvested and when they appear in the Internet search engine.

8. Search engine harvesters can place a considerable load on a web server with multiple harvesters accessing content at the same time.

The National Library has been successful in getting its digital collections content indexed by Internet search engines. However, it has not as yet been able to get many of its images indexed by Internet image services such as Google Image Search [12]. It is not yet clear why the Library's digital collection images have not been included in Internet image services. Google are currently working on improved mechanisms for harvesting images which it is hoped will solve this issue at least for Google Image Search in 2005.

OAI harvesting

The Open Archives Initiative [13] Protocol for Metadata Harvesting (OAI-PMH) is designed to enable transport of structured metadata from data repositories across the Internet to build federated services. In August 2003 the National Library released an OAI interface [14] to its digital collections. Through this interface service providers can harvest the National Library's digital collections' metadata to build federated services. About 100,000 metadata records are available through this interface. In September 2003 the Library's digital collections were harvested by OAIster [15] a federated discovery service developed by the University of Michigan that provides access to over 5 million previously difficult-to-access, "academically-oriented digital resources" from 444 contributing institutions.

In November 2003 an OAI interface to the National Library's PictureAustralia service [16] was used to add over 750,000 records to OAIster. PictureAustralia provides access to images of Australian people, places and events from about 40 cultural (including libraries, museums, galleries, archives and historical societies), scientific, government and academic institutions.

In March 2004, OAIster signed an agreement with Yahoo! to provide its metadata to Yahoo! Search to improve access to previously difficult-to-locate online scholarly resources [17]. At the National Library an immediate increase in the number of accesses to our digital collections was noted from the Yahoo! search engine. Figure 4 shows a marked increase in page views of the Pictures Catalogue in March 2004. Referrals in the resolver logs originating from Yahoo! Search increased from 18,000 in February 2004 to 40,000 in March 2004 as shown in Figure 7. This increase of 117% is one of the largest monthly increases in referrals from an Internet search engine and is much larger than the 17% increase in Google referrals seen from February to March 2004.

Since May 2004 referrals to the Library's resolver from Yahoo! have overtaken those from Google which, at least in part, may be attributable to increased visibility of the National Library's digital collections within search results through the contribution of OAIster content to Yahoo!.

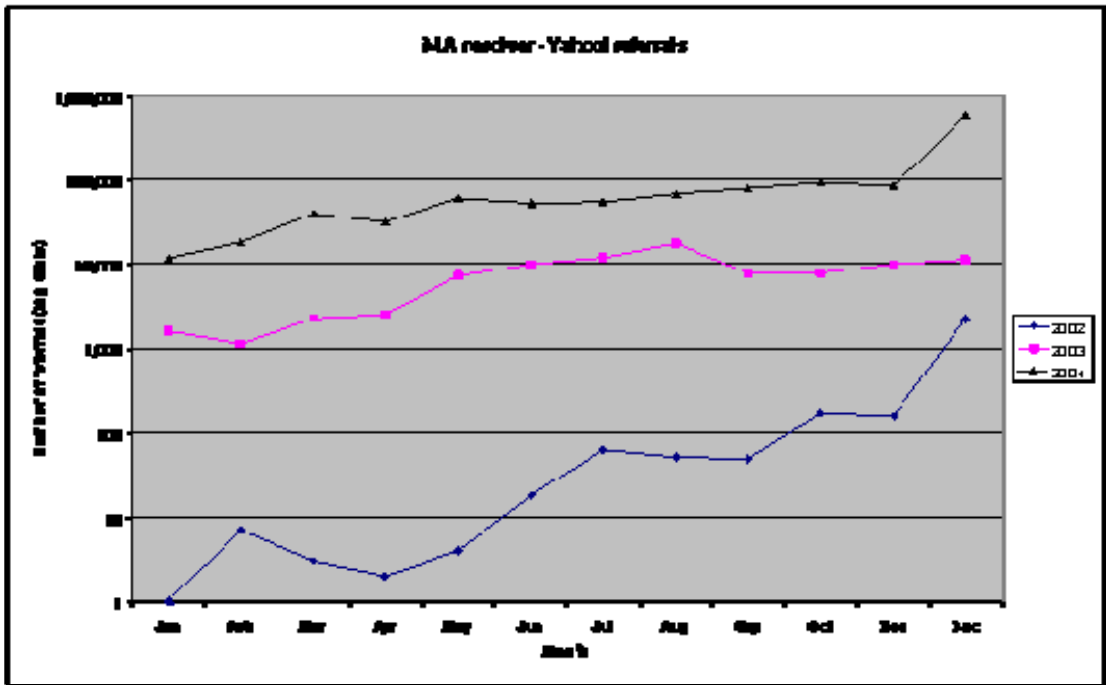

Figure 7: Yahoo! referrals in NLA resolver logs for 2002 to 2004

An alternative approach to use of URL lists for seeding search engines is use of the DP9 [18] an OAI gateway service for web crawlers. This service, developed by the Old Dominion University , provides a persistent URL for OAI repository records and converts these requests to an OAI query against the appropriate repository. A DP9 gateway is available for each OAI repository registered on the Open Archives Initiative web site. The National Library has not implemented a DP9 gateway as the already developed URL lists have served the same purpose.

Future directions

It is hoped that eventually URL lists or DP9 gateways to OAI services will not be required to seed Internet search engines as search engine harvesters will begin to directly support the OAI-PMH. Recently Google has begun experimenting with OAI harvesting, initially via "static" OAI site lists [19] and has begun doing full OAI harvests of selected contributors. Google has experimented with harvesting the National Library's digital collections using the OAI-PMH. The advantage of using OAI is its support for "incremental harvesting", so that after the initial harvest, metadata about only the new, changed or deleted records needs to be harvested. In February 2005, Google harvested the Library's digital collections with these results:

Start: 2005-02-09T22:09:04Z Finish: 2005-02-10T03:45:10Z records: 97662
Start: 2005-02-10T22:09:14Z Finish: 2005-02-10T22:09:44Z records: 62
Start: 2005-02-11T10:09:22Z Finish: 2005-02-11T10:11:41Z records: 194
Start: 2005-02-11T16:09:22Z Finish: 2005-02-11T16:09:32Z records: 0
Start: 2005-02-12T22:09:32Z Finish: 2005-02-12T22:09:41Z records: 0

The initial repository harvest on 9 February 2005 took about 5 1/2 hours, but the incremental harvests are much faster as only a small number of records are in scope.

It is hoped that other search engine vendors will follow Google's lead in this area and implement support for the OAI-PMH. This would increase the efficiency of harvesting for sites wishing to seed search engines with structured metadata.

OCLC's Open WorldCat program [20] provides an exemplar of what is possible when Library service providers cooperate with Internet search engines. Through this program OCLC has made several million bibliographic records available through both Google and Yahoo!. Library records are branded with a "Find it in a Library" tag before the item title and OCLC has created a Library subset within the browser Yahoo! toolbar [21]. When users find a record of interest in Google or Yahoo! search results, by following the link they are transferred to the WorldCat database. Within WorldCat the user can enter their local library and find libraries in their area that hold the item in question.

The National Library is currently redeveloping its Kinetica service. In December 2004 Libraries Australia, a new web search interface to the Australian National Bibliographic Database (ANBD) and several overseas databases, was released [22]. The ANBD contains about 13 million bibliographic records for over 38 million items held in over 1,100 Australian libraries. Libraries Australia provides the ability to not only 'find' descriptions of items in library collections, but numerous options to 'get' these items. Searching is easy and once you find what you want, you can choose to use the item in your local library, purchase a copy through the Library's copying service or buy it from an online bookseller. Many items are also available immediately online. This service is currently available through Australian libraries. In 2006, it will no longer be necessary to visit a library to use the service. It will be made free of charge to all Australians with access to the Internet. This will be of particular benefit to people in regional or remote areas who do not have easy access to a library.

In tandem with this initiative to open up access to the ANBD, the National Library would like to explore exposing its bibliographic records to Internet search engines. A model of

national and regional union catalogues exposing content in a coordinated way to Internet search engines to improve the find and get experience for library users is envisaged. The OCLC WorldCat program provides a model of what is possible in this area. In 2005 the National Library will be exploring cooperative approaches with international union catalogues to seamlessly provide library users with the ability to "find globally, get locally".

Conclusion

Today's users have come to expect instant and simple access to information resources through use of Internet search engines. By exposing the deep web, users can now easily find National Library digital collection items in most popular search engines.

However, it is equally important that high quality, subject or audience specific content is not lost in a "one size fits all" search experience. A need will always exist for specialised gateways to such content, as Internet search engines are realising through initiatives such as Google Scholar [23], a gateway to research literature.

The National Library of Australia has increased the visibility of its digital collections by making them available for harvesting by Internet search engines. A demonstrated increase in access to these persistent, citable collections has been the result to the benefit of library users.

References

[1] S. Raghavan & H. Garcia-Molina (2001). *Crawling the hidden web.* Proceedings of the Twenty-seventh International Conference on Very Large Databases, p. 129 – 138. ISBN:1-55860-804-4. http://www.dia.uniroma3.it/~vldbproc/017_129.pdf

[2] M. K. Bergmann (2001). *The Deep Web: Surfacing Hidden Value*. The Journal of Electronic Publishing, vol. 7, no. 1. http://www.press.umich.edu/jep/07-01/bergman.html

[3] V. Parr & M. Yamine (2003). *Government Online, a national perspective 2003 - New Zealand*. Taylor Nelson Sofres consultants. http://www.e-government.govt.nz/docs/go-survey-2003/

[4] Australian Bureau of Statistics (2005). *Internet Activity, Australia*. http://www.abs.gov.au/Ausstats/abs@.nsf/0/6445f12663006b83ca256a150079564d?OpenDocument

[5] J. P. Goldberg (1995). *On interpreting access statistics.* http://www.goldmark.org/netrants/webstats/

[6] Google. *Google Print: Google checks out Library Books.* http://print.google.com/googleprint/library.html

[7] National Library of Australia. *Pictures Catalogue.* http://www.nla.gov.au/catalogue/pictures/

[8] D. Campbell (2003). *Simply Seeding Search Engines*. AusWeb 03, Ninth Australian World Wide Web Conference, Gold Coast, July 2003. http://ausweb.scu.edu.au/aw03/papers/campbell/paper.html

[9] T. Boston & N. Nguyen (2002). *A practical approach to ensuring the persistence of digital collections at the National Library of Australia*. http://www.nla.gov.au/nla/staffpaper/2002/boston2.html

[10] National Library of Australia . *Digital Collections Infrastructure*: *Access and Delivery*. http://www.nla.gov.au/digicoll/infrastructure.html#delivery

[11] Google. *Google Corporate Information: Technology*. http://www.google.com/corporate/tech.html

[12] Google. *Google Image Search*. http://images.google.com/

[13] Open Archives. *Open Archives Initiative*. http://www.openarchives.org/

[14] National Library of Australia. *National Library of Australia Digital Object Repository*. http://www.nla.gov.au/digicoll/oai/

[15] University of Michigan Digital Library Production Service. *OAIster*. http://oaister.umdl.umich.edu/o/oaister/

[16] National Library of Australia . PictureAustralia. http://www.pictureaustralia.org/

[17] University of Michigan News Service (2004). *U-M expands access to hidden electronic resources with OAIster*. http://www.umich.edu/news/?Releases/2004/Mar04/r031004

[18] Old Dominion University Digital Library Group. *DP9- An OAI Gateway Service for Web Crawlers*. http://arc.cs.odu.edu:8080/dp9/index.jsp

[19] Open Archives Initiative (2002). *Specification for an OAI Static Repository and an OAI Static Repository Gateway*. http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm

[20] OCLC. *Open WorldCat program*. http://www.oclc.org/worldcat/open/default.htm

[21] OCLC. *Yahoo! Toolbar with WorldCat searching of library materials.* http://www.oclc.org/toolbar/default.htm

[22] National Library of Australia . *Libraries Australia*. http://librariesaustralia.nla.gov.au/

[23] Google. *Google Scholar*. http://scholar.google.com/