New Zealand information on the Internet: the power to find the knowledge

Smith, Alastair G.

School of Information Management, Victoria University of Wellington, Wellington, New Zealand

Paper for presentation at LIANZA 2011, 30 October- 2 November 2011, Wellington, New Zealand.

Introduction

In a world of apparently ubiquitous information, does knowledge still equal power? Whatever the answer to this question, we will not have power unless we can retrieve our knowledge. Despite the advances of the last decades, issues remain in finding information on the Web relating to Aotearoa. These include: the efficiency with which the global search engines index the NZ web space, searching for macronised words, the quality of Wikipedia information about NZ, and the availability of open access NZ research.

Using Web Search engines for NZ information

While NZ has some homegrown search engines, most people will start a search for NZ information with one of the major global search engines, since these have good coverage of NZ information (which may be held at sites outside of the .nz domain) and sophisticated searching and ranking features. However estimated recall rates are in the order of 50% and search engine crawlers do not reach all of the .nz webspace (Smith 2004).

In order to investigate this further, I decided to investigate the extent to which the major search engines index NZ web pages that might provide useful information.

A sample of pages were taken from sites linked in the Te Puna Web Directory. This means that the sample contained pages that were likely to have information content relevant to NZ. Pages were checked to see that they were still live. Then a unique phrase from the page, without punctuation that might confuse the search engine results, was searched on both of the main search engines:

- Google
- Bing

Note that Yahoo!, another commonly used search engine, now uses Bing for its search results (http://searchengineland.com/yahoos-transition-to-bing-organic-results-complete-49228).

The result indicated that Google has good coverage of the NZ web space – 98% of web pages were indexed. Note that this isn't inconsistent with the observation that recall rates for a particular query are of the order of 50%. The pages might not necessarily be found in a search for the topic of the page, due to their ranking, choice of keywords, etc.

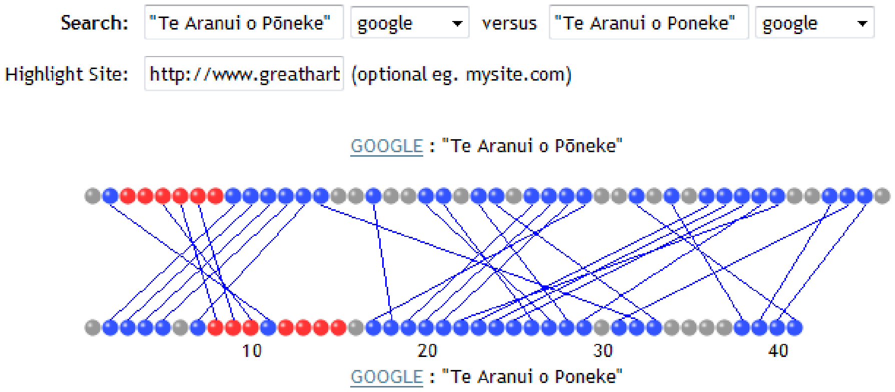
Bing on the other hand, appears to have have less coverage of the NZ web space. Only 67% of the NZ pages were found in the Bing database.

The lessons for librarians here is that while Google in particular has good coverage of NZ material, we shouldn't expect all material relevant to a NZ query to be found in a search. While the coverage of Bing was not as good as Google, it is still valuable to do searches on multiple search engines, databases, directories etc to maximise recall.

The case of the missing macron

If you are searching for a term in Te Reo, do you include the macron? There is significant information relating to Māori on the web, but use of the macron in Te Reo is inconsistent. It can be ignored (Maori), included (Māori), replaced by an umlaut (Mäori), or represented by a double letter (Maaori). Search engines index these forms differently. This can significantly affect retrieval – the web page for the cycling and walking route around Wellington harbour, Great Harbour Way/ Te Aranui O Pōneke, has a lower ranking on Google if the macron is omitted in a search on the name of the path.

A tool that illustrates this is the Thumbshots ranking tool.



Search: "Te Aranui o Pōneke" google versus "Te Aranui o Poneke" google

Highlight Site: http://www.greathart (optional eg. mysite.com)

GOOGLE : "Te Aranui o Pōneke"

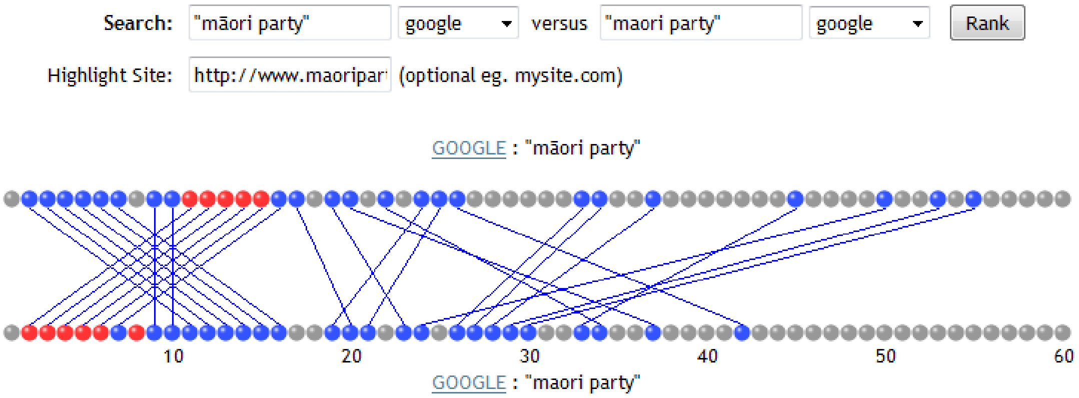
GOOGLE : "Te Aranui o Poneke"

This image shows the ranking of the Te Aranui O Pōneke site using the Thumbshots ranking tool (http://www.thumbshots.com/Products/ThumbshotsImages/Ranking.aspx). It compares two Google searches, one, represented by the upper row of dots, for “"Te Aranui o Pōneke" with the macron, and the lower row for “"Te Aranui o Poneke" without the macron. Higher ranked sites (those that appear at the top of the search result) appear on the left. Blue dots indicate pages that are found by both search formulations, and lines link the same page. Grey dots indicate sites that are found by only one formulation. Red dots are pages at the Te Aranui O Pōneke website. So it's clear that the ranking is very different for the two formulations. Most users only look at the ten items that appear in the default first page of Google results. However if the top (left hand end) 10 items are examined, only three pages from the website are found with the non-macron formulation. None of them, incidentally, are the main page – which is ranked top by the macron formulation, but not found at all by the non-macron formulation. It appears that the only pages found by the non-macron formulation are the ones where “Poneke” appears in the automatically generated URL, where the macron has been dropped. Note also that fewer overall pages are found with the non-macron forumulation.

A similar result occurs for Bing, although less pronounced since a smaller number of hits is found.

In practice, sites tend to use a mixture of the macronised form and the non-macronised form, so that pages are likely to be found using either formulation. But the ranking can be different. For example a search on the resource rich Māori Battalion website (http://www.28maoribattalion.org.nz/) using gets a different order of results depending on whether the macron or non-macron formulation is used.

Particularly topical is a search for the website of the Māori Party, which in fact does not use macrons in the text on its home page (although a macron appears in the logo). In the search shown below, the Party website doesn't make it to the default first page of a search on "māori party", although it is highly ranked for “maori party”. (In fact the rankings have now changed – both formulations bring the Party website to the top, perhaps because of links from sites using the macronised version of the term “māori”).



So the lesson for librarians here is to be aware of macrons in Te Reo terms, and be prepared to search on both macron and non-macron versions when using Web search engines.

Interestingly, common library OPACs seem to be macron agnostic – searches produce the same results whether or not the macron is used in the search terms.

Wikipedia and NZ information

There have been a number of studies that compare Wikipedia information with information in "authoritative" sources, generally indicating that Wikipedia has a level of accuracy in the same order of magnitude as the authoritative sources. For example Giles found that for scientific topics, Wikipedia had a similar level of accuracy to the Encyclopedia Britannica (Giles 2005).

There is no reason to suppose that the case would be different with NZ information, and it's not too hard to find examples of errors for NZ information in both Wikipedia and the conventional authoritative sources. For example: the respected *Columbia Encyclopedia*, which the doyen of reference work evaluators, Bill Katz, says "easily shoves all competitors to one side" (Katz 2002 p.252), lists the height of Mount Cook as 3735m, in both the print 6th edition and the various online versions. Unfortunately this is too low – Wikipedia (along with Te Ara, DOC, and other sources), gives the height as 3754m. This mistake, incidentally, is quite curious – it doesn't even correspond to the 3764m height that Mt Cook had before the 1991 summit collapse.

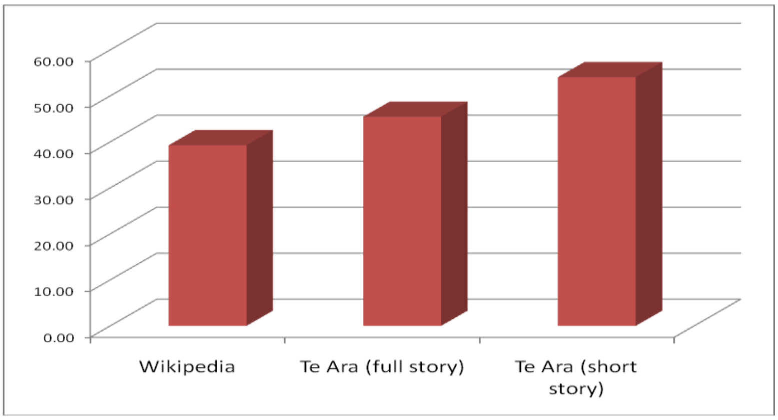
Recently as part of the INFO523 information sources and services course I asked students to update or correct a Wikipedia article on a NZ topic that they have knowledge of. As well as familiarising students with the nature of Wikipedia, this exercise has yielded a couple of interesting insights. Firstly, only about 5% of students find something that is actually an error that they can correct – the vast majority add or update information. Secondly, because they are unfamiliar with Wikipedia conventions, some students make small errors in formatting etc when they edit the article. Interestingly, these errors are corrected very quickly – sometimes within minutes, and always within days. This indicates that at least for NZ topics, Wikipedia has evolved an informal but robust checking and correcting process that contributes to a high level of accuracy.

However accuracy is only one aspect of the utility of an information source. Readability is also important. I decided to compare the readability of NZ information in Wikipedia with an obvious authoritative alternative, the online *Te Ara Encyclopedia of New Zealand*.

I compared a sample of paragraphs selected at random from Wikipedia articles about NZ, and Te Ara articles. Te Ara has two levels of articles – the "short story" a "quick, easy summary" and the full story, for more in depth coverage. I compared the readability of paragraphs from both short stories and full stories. I did not sample articles from the 1966 Encyclopedia, which is also on the Te Ara site.

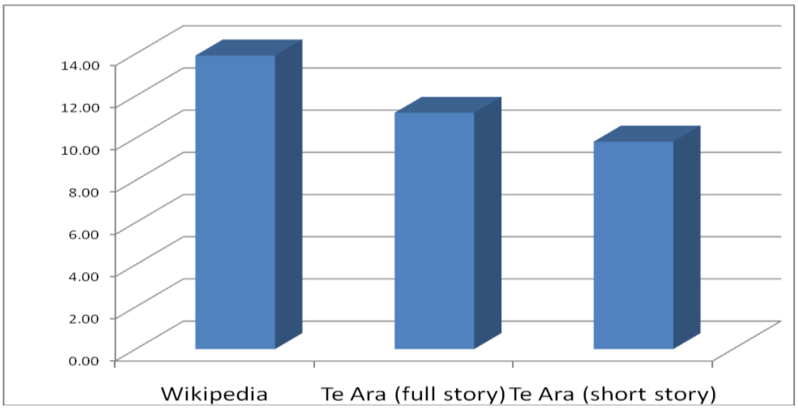
To measure the readability, I used the Flesch reading ease (higher numbers indicate easier to read), and the Flesch Kincaid grade level (US school grade level, lower number easier to read). Both of these were measured by Microsoft Word.

For the Flesch reading ease, the average Wikipedia article scored 39, the full story in Te Ara 45, and the short story 54.



As a comparison, *Time Magazine* is said to have an index of about 52, and the *Harvard Law Review* an index in the low 30s. So the readability of Wikipedia articles about NZ are closer to that of a legal publication than to that of a popular news magazine.

For the Flesch-Kincaid grade score, there is a similar result. The average Wikipedia article scored 14, the full story in Te Ara 11, and the short story 10.



At a grade score of 14, the average Wikipedia article about NZ is in theory not comprehensible for any level of school student, whereas the average full Te Ara story is understandable by NCEA students, and the Te Ara short story should be understandable by most secondary school students.

The lesson here for librarians is that we should not shy away from recommending Wikipedia as a source – it does have useful NZ information, particularly on topics relating to popular culture - but we should also make people aware of the limitations. Wikipedia articles vary in quality, it can pay to check the history to see if a fact has been the subject of recent changes, and that other sources such as Te Ara may be more comprehensible.

New Zealand Open Access Information

NZ research is being made available through open access. However NZ institutional repositories have not been successful in getting high rates of deposit of their institutional output. Nicole Mustatea found that institutions without mandatory deposit have very low rates of deposit. The two NZ repositories she surveyed had in the order of 1% of the institutions publications (Mustatea 2008). This means that much of the research output of NZ institutions is only available through subscription services, which reduces its visibility, as research indicates that open access documents are more frequently cited in most disciplines (Norris 2008).

NZ online journals and conference proceedings are not well indexed by global searching tools such as Google Scholar, partly due to a lack of sophistication in structuring the publications for search engine crawlers, and also due to link rot, which makes the publications inaccessible.

A case study will illustrate the point. The NZ library profession's online journal, the *New Zealand Library and Information Management Journal* is available online without subscription. So far so good.

In an article in *NZLIMJ*, Ailsa Parker (Parker 2007) surveyed online citations made from NZ journals published one to four years previously. 30% of these citations were not accessible. This is a useful contribution to our knowledge, but although Ailsa's paper is online at the LIANZA website, a Google Scholar search does not find it.

Why? Because *NZ Library and Information Management Journal* was published online as a single PDF file for each issue, so Google only indexes the first page or so of the text. You can find the article online via Google Scholar, but only because it has also been made available on the institutional repository at Whitireia, where Ailsa worked.

The lesson here for NZ librarians is that we should encourage our institutions to embrace open access through the development of institutional repositories, preferably with mandatory deposit, and with provision for persistent URLs to address the link rot problem that Ailsa has identified. We can also hope that the new version of *NZLIMJ* will be structured to make it more easily harvested by Internet search engines.

An added benefit for institutions that provide open access is that it helps make research available to the wider community. I recently surveyed links between Wikipedia and NZ and some other institutional repositories, and found that Wikipedia was a significant source of citations to institutional repositories, indicating the repositories role in making research available to the wider community served by Wikipedia (Smith 2011).

Conclusion

Searchers for NZ information who use common Web tools such as global search engines and Wikipedia should be aware of the specific issues relating to NZ information, such as macronised Te Reo terms. Librarians have a role in making NZ's research information available, in particular through ensuring that research appears in open access repositories that are accessible to search engines. Librarians have the power to ensure the availability of New Zealand's knowledge.

References

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070), 900-901.

Katz, W. A. (2002). *Introduction to reference work* (8th ed.). Boston: McGraw-Hill.

Mustatea, N. (2008). *To what extent is material in institutional repository representative of an institution's research output?* Report submitted to the School of Information Management, Victoria University of Wellington in partial fulfilment of the requirements for the degree of Master of Library and Information Studies.

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, 59(12), 1963-1972.

Parker, A. (2007). Link rot: How the inaccessibility of electronic citations affects the quality of new zealand scholarly literature. *New Zealand Library & Information Management Journal*, 50(2), 172-192.

Smith, A. G. (2011). Wikipedia and institutional repositories: An academic symbiosis? *Proceedings of the ISSI 2011 Conference, Durban, South Africa*.

Smith, A. G. (2004). Searching for NZ information in the virtual library. *Electronic Library*, 22(6), 492-497.