Wikipedia and institutional repositories: an academic symbiosis?

Alastair G. Smith

alastair.smith@vuw.ac.nz

School of Information Management, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand

Abstract

A study of citations from Wikipedia articles to documents in institutional repositories showed that although the number of citations was small in relation to the number of documents (citations made to the repositories were 0.35% of the number of documents in the repositories) institutional repositories were a useful source of research information to support Wikipedia articles. 35% of citations were for background information, and 65% were citations supporting specific points, for example: scientific or historical facts, expression of a consensus, attribution of an idea, a convenient summary, the source of a classification, or to give an academic cast to a popular culture article. The types of documents cited reflect the range of academic publishing: 22% of citations were to PhD Theses, 15% to Master level theses, 21% to journal articles, 17% to conference papers, and 11% to technical or working papers. Although the materials in the repositories were overwhelmingly in English, 35% of citations were made from non-English versions of Wikipedia, indicating that institutional repositories play a role in making research available across national and language barriers. Although Wikipedia has been viewed with suspicion by academia the study indicates a potential symbiosis between Wikipedia and academic research in institutional repositories.

Introduction

Institutional repositories have become a popular method for universities and other research institutions to make their research available. There are, however, doubts about whether the information in these repositories are actually being used. Over a similar time period, Wikipedia has become a popular source of information. A small but significant number of links made to documents in institutional repositories are made from Wikipedia articles (Smith, 2009). This suggests that one of the roles of institutional repositories is to make research information available to the non-academic Internet community that builds and uses Wikipedia. The current study investigates in more detail how Wikipedia articles cite documents in institutional repositories, and suggests that rather than being a threat to academia, Wikipedia can have a symbiotic relationship with institutional repositories which promotes the wider availability and visibility of academic research.

There has been an ongoing debate about the quality of Wikipedia (Rector, 2008). However Don Fallis has pointed out that the question is not whether Wikipedia is more reliable than other sources, but whether it is more reliable than the sources that people would use if it did not exist, i.e. other free online sources (Fallis, 2008). In practice, Wikipedia is used by a majority of tertiary students as one of the sources for their course related research (Head & Eisenberg, 2010).

In order to bolster its credibility, Wikipedia places some emphasis on supporting assertions through citations, and this has to some extent been successful. A study of outbound links in Wikipedia articles to articles in scientific journals found good agreement with the citation patterns of scientific literature (Nielsen, 2007). A comparison of Wikipedia citations in articles on brain science with articles in an expert compiled encyclopedia indicated that the Wikipedia citations were of a similar quality, although experts tended to cite more highly ranked journals (Stankus & Spiegel, 2010). However a study of citations in Wikipedia history articles found many claims that were not verified through citations, but pointed out that this highlights the bottling up of information behind subscription barriers (Luyt & Tan, 2010); institutional repositories may be a way to provide reliable sources in a way that can easily be found and used for citation in Wikipedia.

This project examines a sample of links made between institutional repositories and Wikipedia to learn why the links have been made, and what lessons can be learned for the design and management of institutional repositories.
The research questions addressed are:

To what extent are documents in institutional repositories cited from Wikipedia?
What are the reasons for citing institutional repository documents from Wikipedia?
Are citations made from different language versions of Wikipedia to institutional repositories in English?
What types of institutional repository documents are cited from Wikipedia?
What are the discipline areas of institutional repository documents cited from Wikipedia?

Methodology

A convenience sample of links was gathered by using Yahoo Site Explorer (http://siteexplorer.search.yahoo.com/) to create a list of links to a number of Australasian institutional repositories. LexiURL was considered as an alternative, but appeared to have problems operating through the university's firewall.
The options chosen in Yahoo Site Explorer were:

- Inlinks
- Except from this subdomain (which excludes links made within the repository, which are generally navigation links)
- To entire site (i.e. to the individual documents, rather than just to the top level of the repository)

The URLs of the linking sites were then downloaded to a spreadsheet, and filtered to select links from wikipedia.org.
Each site was viewed, and searched for the link(s) to the target document in the institutional repository. For each link, the following was recorded:

- The language version of the Wikipedia article
- The text of the wikipedia reference, e.g. "David Alan Wardle Raman Scattering in Optical Fibres, thesis Doctor of Philosophy in Physics The University of Auckland , January 1999 page 22"
- The reason for making the link e.g. reference/background information, citing in support of specific point etc
- If the link was made to support a particular point, the text that referred to the IR document, e.g. "It grows to around 30 millimetres (1.2 in) shell width.[1]"
- the type of institutional repository document, e.g. thesis, journal article, etc
- The academic discipline that the institutional repository document fell into, e.g. Pure Science, Arts, etc.

Some issues with this methodology need to be recognised.
Yahoo Site Explorer was used by asking it to look for links to the URL of the institutional repository, for example http://researcharchive.vuw.ac.nz for Victoria University of Wellington. However the recommended format for citing documents in institutional repositories is a persistent, repository-independent format of the form http://hdl.handle.net/10063/.... Unfortunately Yahoo Site Explorer does not appear to allow a search for links made to a group of documents using this format. Fortunately for this study, it appears that most users ignore the advice, and make their links using the URL of the repository, so the methodology harvested a significant number of links. However it is possible that this method underestimated the number of links made to institutional repositories.

Yahoo Site Explorer only allows 1000 links to be downloaded. This means that for institutional repositories that had more than 1000 inlinks, the number of links from Wikipedia is unknown. It is assumed that the links are presented in random order, but it has not been possible to confirm this. However in the sample of 17 institutional repositories, only 6 had more than 1000 links found by Yahoo Site Explorer.
Some links were made from administrative or discussion areas of Wikipedia. These were not included in the analysis.
Due to these limitations, the quantitative results need to be treated with caution, and the significance of the study lies in the insights that it gives into how Wikipedia and institutional repositories interact.

Results

First, to what extent are documents in institutional repositories cited from Wikipedia? Table 1 summarises the institutional repositories and the links made to them

Table 1. Institutional repositories and inlinks.

Institution	*Repository URL*	*Total inlinks*	*Sample*	*Wiki links*	*Estimated Wiki links*	*Number Records*
Auckland University Tech	aut.researchgateway.ac.nz	1305	1000	3	4	963
Auckland University	http://researchspace.auckland.ac.nz	3206	1000	16	51	4457
Canterbury University	http://ir.canterbury.ac.nz	1563	1000	26	41	4856
CPIT	http://repository.cpit.ac.nz	70	70	0	0	128
Otago University	http://cardrona.eprints.otago.ac.nz	56	56	1	1	35
Otago University	http://eprints.otago.ac.nz	876	876	6	6	855
Otago University	http://eprintstetumu.otago.ac.nz	102	102	2	2	69
Otago University	http://ourarchive.otago.ac.nz/	70	70	0	0	181
Lincoln University	http://researcharchive.lincoln.ac.nz/	817	817	7	7	2805
Massey University	http://muir.massey.ac.nz/	808	808	7	7	1465
UNITEC	http://unitec.researchbank.ac.nz	40	40	0	0	230
Victoria University Wellington	http://researcharchive.vuw.ac.nz/	752	752	13	13	1357
Waikato University	http://researchcommons.waikato.ac.nz	2381	1000	21	50	3842
Wintec	http://researcharchive.wintec.ac.nz	73	73	0	0	337
Australian	http://dspace.anu.edu.au	1760	1000	44	77	43609

National University						
Monash University	http://arrow.monash.edu.au/	2879	1000	10	29	2528
Swinburn University	http://researchbank.swinburne.edu.au	624	624	9	9	17772
	Totals	17,382	10,288	165	297	85,489

The repositories are those of the main NZ tertiary institutions, plus three Australian institutions. While four repositories had no links from Wikipedia, they have been included in this table since the objective is to evaluate the extent of linking.
"Total inlinks" is the number of links made to the repository URL, as shown by Yahoo Site Explorer. "Sample" is the number of links downloaded, up to a maximum of 1000, as discussed in the methodology section. "Wikipedia links" is the number of links found by filtering on "wikipedia.org" (note that this includes links made from non-English versions of Wikipedia, for example de.wikipedia.org). "Estimated wikipedia links" is an estimate of the total number of links made to the repository from Wikipedia, by assuming that the sample from Yahoo Site Explorer is random, and multiplying the number of links found in the sample by the ratio of the total inlinks to the sample. For example, Canterbury University had a total of 1563 inlinks, but only 1000 were sampled. 26 wikipedia links were found in the sample, so the estimated Wikipedia links are

26*1563/1000= 41

The estimated wikipedia links enable a comparison with the total size of the repository: "number of records" is the total number of documents in the institutional repository, as reported by the Register of Open Access Repositories (ROAR, http://roar.eprints.org/).
Comparing these figures indicates that the number of links made from Wikipedia to the sampled repositories is 0.35% of the number of documents in the repositories, and that links made from Wikipedia are 1.6% of the total links made to the repositories. While these numbers are small, it is worth noting that in most document collections, only a small number of documents get a high degree of use and recognition, and institutional repositories may not be unusual in this regard.
Citations to primary sources (which many institutional repository documents are) are actually discouraged in the Wikipedia guidelines[1], since there is a danger of non-expert interpretation of research; secondary sources are preferred. However it was clear that many articles refer to primary sources, for a variety of reasons.
Looking at the reasons for citing from Wikipedia articles to institutional repository documents, 35% were background or further information (i.e. from the external links or references section of a Wikipedia article), while 65% were citations relating to specific points (i.e. listed in the notes section of the article, as a footnote from a specific point in the article text).
Background or further information citations were generally to documents that gave further information on an aspect of the Wikipedia article (for example several versions of the article on Thai timekeeping systems cited a paper by an ANU academic on the topic) but included a few cases where the institutional repository document was authored by the subject of the article, for example the article on a philosopher Daniel Ross includes a link to his PhD thesis in the Monash University repository. The article on the Tongan island of Kolonga includes a

list of theses completed by people from the island, including one held in the Waikato University repository.
Citations made to support specific points were made for varied reasons. Some examples of the broad range of reasons for citing institutional repository documents were:

- Scientific facts: for example, a PhD thesis on the distribution and ecology of the sulphur-crested cockatoo was cited as evidence that cockatoos had been introduced outside their natural range.
- Historical facts: for example, the metadata for a photograph of a railway station was cited as a source for the date of closure of the station.
- Expression of consensus: for example, a PhD thesis on the ecology of an invasive species of barberry was cited as support for the statement that the species was considered to be a threat to the indigenous NZ ecosystem.
- Attribution of an idea: for example a PhD thesis on the philosopher Michel Foucault was cited as evidence that he was associated with the development of the concept of "limit experience".
- A convenient summary: for example an economics working paper on the professionalisation of rugby was cited as a source for the history of rugby league in Australia.
- Source of a classification: for example an article in a geology journal was cited as the source of the classification of volcanic rocks in the central North Island of New Zealand.
- Giving an academic cast to a popular culture article. For example, the Tamil and Hindi articles on the actor Charlize Theron include a citation to an MA thesis on sex in women's advertising (which interestingly is not cited in the English version of the article).

In a number of cases the point being supported by the citation was not central to the cited repository document. For example an article about a British military unit cited an MA thesis on the development of African history as a discipline, since it happened to mention that the unit had been based at a particular college in Accra. This may be somewhat different from standard academic citation, where there is usually a close correspondence between the topic of the citing document, and the cited document. This phenomena may reflect the tendency to use online documents as sources regardless of appropriateness, and the tendency to find documents through keyword search engines, which may find mention of a minor point in a document that is about an unrelated topic.
The types of repository documents cited reflect the range of materials in institutional repositories. 22% of citations were to PhD Theses, 15% to Master level theses, 21% to journal articles, 17% to conference papers, and 11% to technical or working papers. The remainder were to unpublished documents, photographs, book chapters, book reviews and books (some repositories act as e-book publishers). In a number of cases, the journal articles cited were also available in the online archives of the journal, and the fact that the repository version was cited indicates the value of multiple online locations for a document.
The broad academic disciplines of the documents linked to were 35% pure science (largely biology), 12% applied science, 13% Social Science, and 40% arts (largely history). It is likely that this reflects the content of the repositories, although it would be interesting in a future study to test the representativeness of documents cited.
Although the materials in the repositories were overwhelmingly in English, there was significant citing from non-English versions of Wikipedia. 65% of citations were made from the English version of Wikipedia, 8% to the German version, and the remainder to a wide variety of language versions. South East Asian language versions were well represented,

possibly indicating that research by Australasian institutions is becoming less Eurocentric, and is relevant to their near neighbors.
An interesting aspect was that in a number of cases, a document would be referenced from non-English versions of a Wikipedia article, but not from the English version. For example a book review debunking the idea that Marco Polo did not travel to China is cited in a number of different language versions of the article on Marco Polo, but not in the English version. A possible explanation is that the contributors to the English version have a wider variety of sources to choose from, and don't feel restricted to online sources such as institutional repositories. If this is the case, it indicates the value of institutional repositories in making research available internationally.
There was some indication of a Pareto-like distribution, with a small number of documents being cited several times. In the sample, the most highly cited were the book review (noted above) debunking the idea that Marco Polo did not in fact travel to China (cited 11 times, mainly in different language versions of the same article), and a background paper for a New Zealand soil science conference which was cited 9 times, mainly as a source of the date of human settlement of New Zealand. 25 documents were cited more than once.

Conclusions

It has been suggested that institutional repositories are a way of making research information available to the general public (White, 2008). While a study of the citation impact of theses in institutional repositories found that there was no evidence that having the theses in a repository improved the citation impact (Lariviere, Zuccala, & Archambault, 2008) this study illustrates that theses in institutional repositories are being used as evidence for Wikipedia articles.
A recent study of the use patterns of Wikipedia contributors showed a preference for online sources (Huvila, 2010) and the current study seems to support this, showing that institutional repositories are a convenient source of research information for Wikipedia contributors. In fact adding links to Wikipedia has been used as a deliberate strategy by libraries to promote the use of their digital collections (Lally & Dunford, 2007).
Although the current study is based on small numbers, it indicates that institutional repositories are one way in which research information is being used to improve the reliability of Wikipedia information.
Although many citations were to theses, the significant number of citations to journal articles stored in institutional repositories indicates that this is a useful role of institutional repositories, even if the journals are also available online through other sources. A study of LIS literature indicated that journal articles are not being systematically archived in institutional repositories (Way, 2010), and the same is probably true of other disciplines. It would be valuable for institutional repositories to systematically include journal articles and conference papers published by members of the institution, even if these articles are available elsewhere.
The large number of citations from non-English versions of Wikipedia is a reminder that over 75% of Wikipedia is written in languages other than English, and these have their own communities of practice (Hara, Shachaf, & Hew, 2010). This result indicates that institutional repositories may be facilitating access to research for users in other language groups.
Wikipedia has been viewed with suspicion by academia: a survey of academics found that although many academics use Wikipedia, they see it as disrupting the model of academic communication (Eijkman, 2010). At the same time there have been suspicions of institutional repositories. Dorothea Salo suggests rather bluntly “[The institutional repository] is like a roach motel. Data goes in, but it doesn’t come out” (Salo 2008).

In contrast the current study indicates a potential symbiosis between Wikipedia and academic research in institutional repositories. This symbiosis may be providing a route by which data escapes from Salo's roach motel. In order to take advantage of this symbiosis academics should be systematically placing their research work in institutional repositories.

Note

1 http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources#Scholarship

References

Eijkman, H. (2010). Academics and Wikipedia: Reframing Web 2.0+ as a disruptor of traditional academic power-knowledge arrangements. *Campus-Wide Information Systems, 27*(3), 173-185.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science & Technology, 59*(10), 1662-1674.

Hara, N., Shachaf, P., & Hew, K. F. (2010). Cross-cultural analysis of the Wikipedia community. *Journal of the American Society for Information Science and Technology, 61*(10), 2097-2108.

Head, A. J., & Eisenberg, M. B. (2010). How today's college students use Wikipedia for course–related research. *First Monday, 15*(13)

Huvila, I. (2010). Where does the information come from? Information source use patterns in Wikipedia. *Information Research, 15*(3)

Lally, A. M., & Dunford, C. E. (2007). Using Wikipedia to extend digital collections. *D-Lib Magazine, 13*(5/6)

Lariviere, V., Zuccala, A., & Archambault, E. (2008). The declining scientific impact of theses: Implications for electronic thesis and dissertation repositories and graduate studies. *Scientometrics, 74*(1), 109-121.

Luyt, B., & Tan, D. (2010). Improving Wikipedia's credibility: references and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology, 61*(4), 715-722.

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *First Monday, 12*(8)

Rector, L. H. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review, 36*(1), 7-22.

Salo, D. (2008). Innkeeper at the Roach Motel. *Library Trends, 57*(2), 98-123.

Smith, A. G. (2009). Linking to institutional repositories from the general Web. *Proceedings of ISSI 2009,* 14-17 July 2009 Rio de Janeiro. 211-217.

Stankus, T., & Spiegel, S. E. (2010). Wikipedia, Scholarpedia, and references to journals in the brain and behavioral sciences: A comparison of cited sources and recommended readings in matching free online encyclopedia entries. *Science & Technology Libraries, 29*(3), 258-265.

Way, D. (2010). The open access availability of library and information science literature. *College & Research Libraries, 71*(4), 302-309.

White, B. (2008). Minding our ps and qs: Issues of property, provenance, quantity and quality in institutional repositories. *IATUL 2008,* 21-24 April 2008 AUT University, Auckland, New Zealand.