

Linking to Institutional Repositories from the general Web

Alastair G Smith

alastair.smith@vuw.ac.nz

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

Abstract

The project investigated the kind of links made from the general web to Institutional Repositories (IRs). Yahoo Site Explorer was used to collect URLs of pages linking to the larger IRs in New Zealand; as well as three well established overseas IRs. Samples of these links were classified, for example as to whether they were formal citations, informal links, or subject directory links. Formal links comprised only 1.85% of the links, however informal links (largely from blogs and Wikipedia) comprised 18.2%. Subject directory links comprised almost 50% of links. These results indicate that IRs have a useful role in making research information available to the general public, even though they do not appear to be highly cited within the research literature.

Introduction

Institutional repositories (IRs) are becoming an important form of research publishing. Many universities around the world are creating IRs and encouraging the deposit of publications by their staff. These publications may also have appeared in conventional research media such as journals and conference proceedings, but institutions are encouraging deposit in IRs for a number of reasons, including: open access to research, preservation of research outputs, and showcasing an institution's research output, and facilitating communication between researchers. Are IRs in fact contributing to research communication, and is this reflected in, for example, citation impact?

IRs are one form of open access (OA). Research on OA has focussed on whether the wider availability of articles through open increases citation impact, although conclusions from this research are mixed. The citation impact of articles in four disciplines at varying stages of OA adoption (philosophy, political science, electrical and electronic engineering, and mathematics) indicated that articles freely available over the Internet had greater impact (Antelman, 2004). A study of the ISI citation rates for research publications in ecology, applied mathematics, sociology, and economics found that while there appeared to be a clear citation advantage for OA publications, this advantage varied between disciplines (Norris, Oppenheim, & Rowland, 2008). However an analysis of articles in physiology indicated that while OA publishing reached more readers than subscription publishing, there was not a clear citation advantage (Davis, Lewenstein, Simon, Booth, & Connolly, 2008). A study of condensed matter research indicated that any citation advantage of OA may be due to earlier availability, and that there is not a long term citation advantage (Moed, 2007).

Many IRs are primarily based on theses (Electronic Theses and Dissertations) however Lariviere and colleagues found no evidence that these have a positive affect on the impact of the theses (Lariviere, Zuccala, & Archambault, 2008).

Since IRs by definition are websites, it would seem logical to apply webometric methods to investigate them. However there has been relatively little webometric exploration of IRs. Zuccala and colleagues combined a LEXIURL study of IRs with interviews of IR managers, suggesting that the linking analysis provided by LEXIURL could be a useful management tool (Zuccala, Oppenheim, & Dhiensa, 2008).

How can IRs be investigated webometrically or bibliometrically? In practice, identifying citations to IRs from conventional research publications is difficult. Existing citation searching tools such as Scopus and the ISI citation indexes are not very useful in identifying documents that cite a specific institutional repository, because the format of citations stored in the database doesn't easily allow searches for the particular URLs which identify institutional repositories. Google Scholar crawls most IRs, and allows links to a specific document in an IR to be identified, but does not appear to allow searches for all links to an IR.

An additional problem is the Handle System of digital object identifiers (<http://www.handle.net/>). Repositories encourage linking to a persistent URL, in the format <http://hdl.handle.net/xxxx/yyyy> where xxxx identifies the specific repository, and yyyy identifies the particular document. So a document may have a URL reflecting the archive, for example <http://researcharchive.vuw.ac.nz/handle/10063/331> but references are encouraged to the persistent URL <http://hdl.handle.net/10063/331>. This approach is sensible, since while administrative changes may result in changes to the IRs URL, the handle URL will remain the same. However the existence of multiple URLs poses problems for link studies.

In any case IRs often contain documents that have been published elsewhere, and it seems that citations are often made to the conventionally published version, rather than the version in the IR, even if document was originally sourced from the IR. As an aside, the citing of a different version from the one consulted may be a concern, since in some cases the version in the IR is different from the conventionally published version (for example, it may be a pre-review version).

The current research project investigates what kind of links are made from the general web to IRs.

Methodology

This research investigated the kinds of links made from the general Internet to a selection of IRs, using Yahoo Site Explorer (YSE). The decision to investigate linking from the general Internet was taken due to the problems discussed above with search tools that covered the "scholarly" web; but also because it seemed useful to see how IRs influence the wider Internet community.

The IRs in this exploratory study were chosen because of the author's interest in the development of IRs in his own country, and also because the inclusion of NZ IRs might provide insights through the author's contextual knowledge. So the sample group included the larger IRs in New Zealand as well as three well established overseas IRs. ANU is one of the largest IRs in Australasia; QUT and Soton are examples of IRs that have achieved a high level of coverage of their research output.

Table 1: IRs studied

Abbrev	Institution	Items	URL
VUW	Victoria University of Wellington	367	http://researcharchive.vuw.ac.nz/
UAuck	University of Auckland	2371	http://researchspace.auckland.ac.nz/
UCanty	University of Canterbury	1529	http://ir.canterbury.ac.nz/
UOtago	University of Otago	677	http://eprints.otago.ac.nz/
LiU	Lincoln University	535	http://researcharchive.lincoln.ac.nz/
WaiU	Waikato University	460	http://waikato.researchgateway.ac.nz/
AUT	Auckland University of Technology	361	http://aut.researchgateway.ac.nz
ANU	Australian National University	46823	http://dspace.anu.edu.au/
QUT	Queensland University of Technology	11324	http://eprints.qut.edu.au/
Soton	University of Southampton	39596	http://eprints.soton.ac.uk/

Information about the number of items is taken from the Registry of Open Access Repositories (ROAR) (<http://roar.eprints.org/>).

YSE was chosen because at the time of the research (September 2008) it offered the best facilities for searching for inlinks to a particular site. Google's link search only identified links to a specific page, rather than all links to a website; Windows Live's link search (Thelwall, 2008) was not available at this time.

YSE provides a list of sites that link to a given website (for example, links to all pages with URLs beginning <http://researcharchive.vuw.ac.nz>). For the current research, the option to exclude links from the IR subdomain was taken (for example, links between different pages within the <http://researcharchive.vuw.ac.nz> site were excluded. However in practice some internal links were captured by YSE, since internal links are often use the handle URL (e.g. <http://hdl.handle.net/10063/...>), which YSE does not recognise as an internal link, and these had to be excluded manually. Arguably links between different documents in the same IR are legitimate citations, and should be counted; however few if any of the internal links found in this study were of this nature; most were navigation links within the repository.

YSE allows the list of linking pages to be downloaded as a tab delimited file, which was loaded into an Excel spreadsheet for analysis. YSE only allows the first 1000 linking pages to be downloaded. Communication with Yahoo indicated that results from YSE are presented in no particular order, so it was assumed that these 1000 linking pages are indicative of all the pages that link to the IR. However this assumption is a potential limitation in the results.

In Excel, the list was cleaned by removing internal links, links from spam pages, broken links and duplicate links, and a random sample of 100 links classified.

The links were classified according to a scheme (Table 2) modelled on one used for analysing links to research articles in open access journals (Kousha & Thelwall, 2007).

Table 2: Classification scheme for links made to IRs

Code	Type of link	Note
1a	Formal Research impact	Formally cited research in journals, conference proceedings, etc (not counted in 1a1 or 1a2)
1a1	Research or technical report	Citation from a research or technical report
1a2	Online magazine article	Citation from a formally published online magazine
1b	Informal Research impact	Linking from general web sources: online reading lists, etc
1b1	Link from blog or discussion posting	Some blogs are hosted by an online magazine, so in some cases it was difficult to distinguish from a magazine citation.
1b2	Link from Wikipedia or other online reference resource.	Links from multiple versions of the same Wikipedia article were counted as one link
1b3	Page using image from IR	Some IRs (for example ANU) contain images, which may be included directly into external web pages.
2	Self Publicity	Link from a person's home page to the person's publications in an IR.
3a	Links from General Web directories	Links to specific documents in the IR from general directories of Internet resources
3b	Links from Subject directories	Links to specific documents in the IR from specialised subject directories
3c	Links from Directories of IRs	Links to the IR as a whole from directories of repositories, for example ROAR
3e	Links from a virtual repository	Some repositories (for example Picture Australia) make direct links to objects held in other repositories (e.g ANU)

A sample of links were classified by another observer, confirming the overall reliability of the classification process. However it is clear that in the evolving Internet environment, there are grey areas in classification – for example, distinguishing between a blog and an online magazine can be difficult. Where non-English pages were the source of the link, contextual information and Google's translation facilities were used to determine the nature of the link.

It was also clear that there was a high degree of duplication of pages. Wikipedia articles, for example, occur in different language versions on the Wikipedia website, and are often copied in their entirety to other sites. Where possible, these duplicates were identified and removed.

Results

To get a picture of the kind of links made to IRs from the general Web, the proportion of links in each category were calculated as a proportion of the real, non-internal links reported by YSE and aggregated over all the samples. This is shown in Table 3.

Table 3: types of links made to documents in IRs

Type of link	Percentage of real, non-internal links
1a Formal Research impact	0.74%
1a1 Research or technical report	0.74%
1a2 Online magazine article	0.37%
1b Informal Research impact	1.78%
1b1 blog	7.32%
1b2 wikipedia or other online reference resource	7.54%
1b3 page using image from IR	1.55%
2 Self Publicity	3.77%
3a General Navigation, directories etc	1.04%
3b Subject specific navigation	48.74%
3c Directory of IRs	15.01%
3e links from a virtual repository	11.39%
Formal research links	1.85%
Informal research links	18.20%
Directory links	49.78%

As well as the percentages of links made in the specific categories, the percentage of links made under broader categories was calculated:

- Formal research links: this is the sum of 1a, 1a1, and 1a2, and are the links that could be regarded as equivalent to citations in a conventional print environment. As can be seen this is a fairly small proportion of the links made to IRs.
- Informal research links: the sum of 1b, 1b1, 1b2, 1b3. These are links made from Web sources that do not have direct equivalents in the conventional publishing environment. It is interesting that blogs and online reference sources such as Wikipedia are a significant source of links to IRs.
- Directory links: this is the total of 3a, 3b. These are links made from directories to documents in IRs because of their information value. It is interesting that despite the ubiquity of search engines, subject directories are still an effective way to locate information in IRs.

The kinds of links made to each individual IR are reported in Table 4. These figures should be treated with caution, since this research is exploratory and as noted above, there are limitations to the sampling procedure used. However these results give an indication of the range of types of links made to IRs. The numbers of links are given as a proportion of the total links reported by YSE:

- Formal links: sum of 1a, 1a1, and 1a2
- Formal and informal links: sum of 1a, 1a1, 1a2, 1b, 1b1, 1b2, 1b3
- Significant links (specific “citations” and listings of documents in subject directories – i.e. all links relating to IR documents because of their information value): 1a, 1a1, 1a2, 1b, 1b1, 1b2, 1b3, 3a, 3b, 3c

Table 4: links and impact factors for each IR

	Formal links	Formal and informal links	Significant links	Research Impact Factor	Subject Impact Factor
VUW	0.00%	3.52%	5.53%	0.038	0.060
UAuck	1.00%	7.00%	12.00%	0.030	0.052
UCanty	0.00%	0.36%	7.00%	0.001	0.025
UOtago	0.61%	3.33%	54.85%	0.049	0.802
LiU	0.00%	6.00%	6.00%	0.022	0.022
WaiU	0.00%	2.00%	3.00%	0.007	0.010
AUT	0.00%	11.00%	17.00%	0.066	0.101
ANU	0.90%	13.60%	19.20%	0.025	0.035
QUT	3.00%	42.00%	65.00%	0.203	0.314
Soton	6.00%	18.00%	77.00%	0.131	0.559

The impact factor has been widely used to gauge the overall impact of journals and websites. For this research, two impact factors were calculated:

- Research impact factor: the formal and informal research links (1a, 1a1, 1a2, 1b, 1b1, 1b2, 1b3) divided by the number of documents reported at the IR by ROAR.
- Subject impact factor: all links made to IR documents because of their information value (1a, 1a1, 1a2, 1b, 1b1, 1b2, 1b3, 3a, 3b), divided by the number of documents reported at the IR by ROAR.

It can be seen that few links are formal citations. It is notable that the larger, more mature IRs (ANU, QUT, Soton) have higher proportions of formal, informal, and significant links. The high impact factors for QUT and Soton may indicate the influence of mandatory deposit of publications at these institutions (Swan & Carr, 2008)

In less mature IRs, impact can be influenced by a few well linked articles. For example, a report of the UOtago implementation of the ePrint server (Stanger & McGregor, 2006) has been widely cited in the open access community, contributing to a relatively high citation rate for this IR.

The substantial numbers of informal links made to documents in IRs indicates that IRs provide an easy way for the general public to access original research. Two examples of this:

- An LiU study, of the carbon produced in producing and transporting food to European consumers, received numerous links from blogs discussing the topic of carbon emissions, as well as from Wikipedia articles.
- A UOtago study of bias in football refereeing was linked from sports blogs.

The number of links made to IRs from Wikipedia is also of interest. The Wikipedia community is placing importance on citing evidence for assertions in Wikipedia articles, and this seems to be reflected in links made to research available on IRs. This is another good example of how IRs make research material available to the general public, and how Wikipedia (despite its faults) provides an entry point to sources of information on a topic.

Conclusions

This research found few formal citations to documents in IRs, with only 1.85% of links being formal. This may be due to the nature of the general web that YSE indexes. A study done in a database that indexes the research web, such as Google Scholar, might reveal a higher proportion of formal citations. A possible approach would be to take a sample of documents in IRs and search in, for example, Google Scholar, for citations to these documents. However exploratory checks of this approach indicate that many documents in IRs do not show citations in Google Scholar, so that the sample would need to be large in order to produce useful results.

It is also possible that as IRs mature, there will be more conventional citations to them, and this will be reflected in the Scopus and ISI citation databases. A study of links to open access social science journals found that 19% of links were formal (Kousha & Thelwall, 2007), and possibly in future IRs may achieve the same level of formal impact.

The research also indicates that IRs that achieve a good coverage of the institutions output, such as QUT and University of Southampton, also have high impact factors.

Based on the evidence of this exploratory study, the contribution of IRs may be in the availability of research material to the general public, rather than to research communities. This was shown by the relatively high proportion (almost 15%) of links made from blogs, and from Wikipedia and similar reference sources. As one observer has written “The development of institutional repositories has opened the path to the mass availability of peer-reviewed scholarly information and the extension of information democracy to the academic domain”(White, 2008) .

Walt Crawford, writing about library and information studies literature, says “To a great extent, the formal literature now serves as history, explication, formal results of formal research studies, and background. The action is in the informal literature”(Crawford, 2008). This is probably true of many research areas now. An example of the growth of informal literature is research blogs. In a number of research areas, researchers are communicating through blogs, for example, the Webometrics Blog (<http://webometrics.blogspot.com/>). A study of the links in 15 academic blogs indicated that 12.5% of links were to a paper by another researcher (for example in an IR), and that blogs were part of the research communication process (Luzón, 2009). This trend could lead to a research communication environment in which research blogs and institutional repositories work together in synergy.

References

- Antelman, K. (2004). Do open-access articles have a greater research impact? *College and Research Libraries*, 65(5), 372-382.
- Crawford, W. (2008). Thinking About Library Literature. *Online*, 32(6), 58-60.
- Kousha, K., & Thelwall, M. (2007a). How is science cited on the Web? A classification of google unique Web citations. *Journal of the American Society for Information Science and Technology*, 58(11), 1631-1644.
- Kousha, K., & Thelwall, M. (2007b). The Web impact of open access social science research. *Library & Information Science Research*, 29(4), 495-507.
- 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.
- Luzón, M. J. (2009). Scholarly hyperwriting: The function of links in academic weblogs. *Journal of the American Society for Information Science and Technology*, 60(1), 75-89.

- Moed, H. F. (2007). The effect of "open access" on citation impact: an analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13), 2047-2054.
- 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.
- Stanger, N., & McGregor, G. (2006). *Hitting the Ground Running: Building New Zealand's First Publicly Available Institutional Repository* (Information Science Discussion Paper 07). Dunedin, NZ: University of Otago. Retrieved 24 January 2009 from <http://eprints.otago.ac.nz/274/1/dp2006-07.pdf>
- Swan, A., & Carr, L. (2008). Institutions, Their Repositories and the Web. *Serials Review*, 34(1), 31-35.
- Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study windows live *Journal of the American Society for Information Science & Technology*, 59(1), 38 - 50.
- White, B. (2008, 21-24 April 2008). *Minding our ps and qs: Issues of property, provenance, quantity and quality in institutional repositories*. Paper presented at the IATUL 2008, AUT University, Auckland, New Zealand. Retrieved 24 January 2009 from <http://digitalcommons.massey.ac.nz/bitstream/10179/645/>
- Zuccala, A., Oppenheim, C., & Dhiensa, R. (2008). Managing and evaluating digital repositories. *Information Research*, 13(1). Retrieved 24 January 2009 from <http://informationr.net/ir/13-1/paper333.html>