Webometric evaluation of institutional repositories

[1]Alastair G Smith

[1] *alastair.smith@vuw.ac.nz*
School of Information Management, Victoria University of Wellington, New Zealand

Abstract

This research project investigated (a) whether there is a relationship between the Web Impact Factor (WIF) of an institutional repository, and conventional measures of research quality of the institution; and (b) whether a relatively new search engine, Blekko (http://blekko.com/) was a viable tool for webometric investigation.

Blekko was used to count links made to the institutional repositories of Australasian universities. The WIF was calculated as the number of inlinks to the repository, divided by the number of documents at the repository.

At the time of the research, Blekko's coverage of many of the institutional repositories was inconsistent, so only a subset of the repositories was used in the analysis, and the results are inconclusive. However there appeared to be a small correlation between the WIF of a repository and the QS World Ranking; and for Australian institutions, the WIF of a repository and the institution's ERA score.

Introduction

Web Impact Factors (WIFs) have been a subject of webometric investigation since they were proposed by Almind and Ingwersen (1997). In particular they have been used to evaluate the research impact of research institutions such as universities, generally by measuring the inlinks to the domain of the institution, and dividing by a measure of the amount of material at the domain, for example the number of web pages in the domain. An issue with this approach is that many of the inlinks are to institutional web pages that are not research focussed, and the resulting WIF is not representative of the research impact of the institution. Many universities now maintain institutional repositories, servers containing copies of the research output of the institution. In theory, links to the institutional repository should be more indicative of the research impact of the institution than links to the institutional website as a whole. The current research project investigates web linking to institutional repositories, by calculating WIFs and comparing these with some conventional measures of research impact.

Over the years, a number of tools have been used to measure the number of inlinks to a domain. Generally these have been commercial search engines. However the search engines used in the past (such as Alta Vista, Yahoo Site Explorer, and Google) have either ceased to operate, or no longer provide the information required. At the time this research project was carried out (April 2012) a search engine, Blekko (http://blekko.com/) provided inlink information, so the current research offers an opportunity to test whether Blekko is a viable tool for webometric research.

Literature Review

The Web Impact Factor (WIF) was proposed by Almind and Ingwerson (1997), as a measure analogous to the journal impact factor used to evaluate conventional journals. The impact factor was calculated by dividing the number of links, or citations, made to a source (a website or a journal), by the number of linkable objects at the source (pages at a website, or articles in a journal).

It was possible to collect link data relatively easily through commercial search engines (Alta Vista provided useful tools for this) and from crawlers specifically designed for webometric

research. The concept of the WIF was taken up and used in a number of other studies. Smith and Thelwall (2002), for example, used both a specifically designed crawler and commercial search engines to compare WIFs for Australasian universities with conventional measures of research excellence, finding a general correlation.

Aguillo *et al* (2006) used a range of measures derived from commercial search engines to produce web rankings for a large set of institutions, and found that these cybermetric measures could complement traditional measures such as those based on the Science Citation Index, particularly for technologically oriented institutions, and institutions in developing countries.

With the increasing use of institutional repositories to store and make available the research output of institutions, there is interest in evaluating the impact of institutional repositories. Shukla and Poluru (2012) analysed the web presence of Indian state universities, and found that open access institutional repositories were helpful in increasing the visibility of institutions on the Web.

Scholze (2007) pointed out that the trend to place research in open access institutional repositories made more data available for statistical analysis and evaluation of research, and discussed some ways in which citation and usage data for repositories could be collected and analysed.

Zuccala *et al* (2007) used web link analysis and server log files to study the impact and usage of an institutional repository, the National electronic Library for Health, gaining insights into who used the repository, and how they reached it.

Eccles, Thelwall and Meyer (2012) used link data to evaluate several digital resources in the humanities. They collected link data from Yahoo!, and compared target repositories with "comparator" sites, discovering some methodological issues, such as repositories changing their URLs, but also gaining some insights that could inform promotion and marketing of repositories.

Links made to research information on the Web are not exactly analogous to citations. Smith (2011) found that a significant number of links to institutional repositories were made from non-research sites such as Wikipedia. Sato and Itsumura (2011), who analysed links to the Kyoto University institutional repository, also found that links were made from non-academic sources such as Wikipedia and personal web pages. This indicates that a WIF may be measuring a different kind of research impact than a conventional citation based metric.

Research Questions

The current research project examined the use of WIFs to evaluate an organisation's research impact, by considering links to the institutional repository. Arguably links made to the institutional repository are more indicative of the organisation's research impact than links made to the organisation as a whole. To test the effectiveness of these measures, the WIF based measures were compared with some conventional measures of research quality, such as ISI citation measures and New Zealand's PBRF research quality scores. As an aside, these conventional measures are not without controversy. For example Chambers, Gardner and Smith (2010) argued that the methodology of creating the scores affects the way in which scientists carry out and publish their research.

Much of the previous webometric research has been based on search engines such as Alta Vista and Yahoo! Site Explorer which are no longer available for this purpose. At the time of the research was carried out, a relatively new search engine, Blekko ((http://blekko.com/) was the only public source of linking information, so the research project was an opportunity to test this search engine as a source of webometric data.

In summary, the research questions addressed by the project were:

1. Do webometric impact factors for institutional repositories correlate with other measures of research quality and impact of the institutions?
2. Is Blekko a viable tool for measurement of inlinks to an institutional repository?

Methodology

To test the viability of calculating WIFs for institutional repositories, the main institutional repositories of universities in Australia and New Zealand were selected, using the database ROAR (http://roar.eprints.org/). The sample comprised 39 institutional repositories, 31 from Australia and 8 from New Zealand.

Some institutions had more than one institutional repository listed in ROAR, but in practice there was just one major repository, and other listings were either for an older URL that is now redirected to the current repository, or a small departmental repository. For this research project, only the major repository URL was used.

The Blekko search engine provides options for exploring the links into a domain, and the pages at the domain, using the "/seo" tag. For example the query

```
ir.canterbury.ac.nz /seo
```

gives

- The number of pages that Blekko has indexed at the domain (the University of Canterbury institutional repository).
- The number of links to pages in the domain – this includes internal links within the domain. Blekko provides a list of originating domains for links, in theory allowing the number of internal links to be determined, and thus the number of external inlinks.

For this research project, several versions of the WIF were calculated.

- WIF(B): the number of inlinks (links to pages in the domain, less the number of internal links), divided by the number of pages identified by Blekko.
- WIF(R): the number of inlinks (links to pages in the domain, less the number of internal links), divided by the number of records listed in ROAR for the institutional repository.
- WIF(BR): the total number of inlinks reported by Blekko, including internal links, divided by the number of records listed in ROAR for the institutional repository.

The merits of these different versions of the WIF are discussed in the results section.

These WIFs were compared with conventional published measures of research impact. These were:

- The average citations/paper for research published by the institution, as measured by ISI/Thomson Reuters in their *Essential Science Indicator*s

(http://thomsonreuters.com/products_services/science/science_products/a-z/essential_science_indicators/).

- The score achieved by the institution in the QS world rankings of universities (http://www.topuniversities.com/university-rankings/world-university-rankings/2011). Note that the available scores were for the top 400 ranked universities, so not all institutions had a score.
- For Australian institutions, a score derived (Hare & Trounson 2011) from the 2010 ERA research assessment carried out by the Australian Research Council.
- For New Zealand institutions, the score achieved in the 2006 Performance Based Research Fund (PBRF) research assessment (Tertiary Education Commission 2006).

A potential methodological issue in studying links to material in institutional repositories is the use of a persistent URL format in links, of the form for example of http://hdl.handle.net/10063/331, rather than an institution specific URL for example http://researcharchive.vuw.ac.nz//handle/10063/331. While linking using persistent URLs is good practice, it creates problems for webometric analysis, since it is not easy to search for persistent URL links to a specific repository. Fortunately for webometric research, if not for persistence of links, most users creating links tend to use the institution specific URL rather than the persistent URL.

Results

Interpretation of the results is easier if the second research question, "how did Blekko perform as a webometric tool?" is addressed first. Several practical issues emerged during the research.

The number of pages at the repository reported by Blekko was on average 21% of the number of records in the repository reported by ROAR. This proportion varied significantly between institutions, from 0.2% at Deakin University, to 96% at University of Sydney. In some cases Blekko indexed more pages at the repository than the number of documents registered at ROAR. This may be due to indexing multiple pages associated with a repository item, for example a short and a long version of the record. The variation in coverage implies that Blekko indexing of the institutional repositories is inconsistent, being fairly comprehensive at some institutions, but only indexing a sample of sites at other institutions. This meant that WIF(B), which used the number of pages identified by Blekko, is probably not a good measure for evaluation.

The number of internal links was not available for all the repositories. In practice, Blekko lists a selection of the linking domains. In many cases this did not include the links made from the domain itself, so that the number of internal links could not be calculated. This meant that WIF(B) and WIF(R) could not be calculated for many repositories. It is also hard to evaluate what kind of "internal" links should be discounted in the case of institutional repositories. Traditionally, in calculating the WIF, links within the website were considered to be purely navigational, and not a measure of the impact of the website, and so were discounted. However links between documents in an institutional repository may be legitimate citations, which should be included as evidence of research impact. Conversely, links from other pages at the institution into the institutional repository may be simply navigational, or may be genuine measures of impact. This means that in the case of institutional repositories, the distinction between "external" inlinks and "internal" inlinks may not be useful. This means that the WIF(BR) may be the most realistic measure for the current research project.

That leads us to the first research question. "To what extent did the WIFs calculated in this research correlate with the conventional measures of research impact for the institutions?"

A range of potential correlations were investigated, using scatter graphs and the Excel CORREL measure. In view of the limitations noted above, data was analysed for institutions where Blekko provided significant data: i.e the number of internal links could be identified in the Blekko search output, and more than 1000 inlinks had been detected. This meant that 5 NZ institutions and 8 Australian institutions were included in the final analysis. The WIFs and research evaluation measures are included in the appendix.

Figure 1 compares the WIF(BR) with the average citations/paper from the ISI *Essential Science Indicators.* The Excel CORREL score is -0.145. This indicates that if anything there is a possible slight negative correlation.

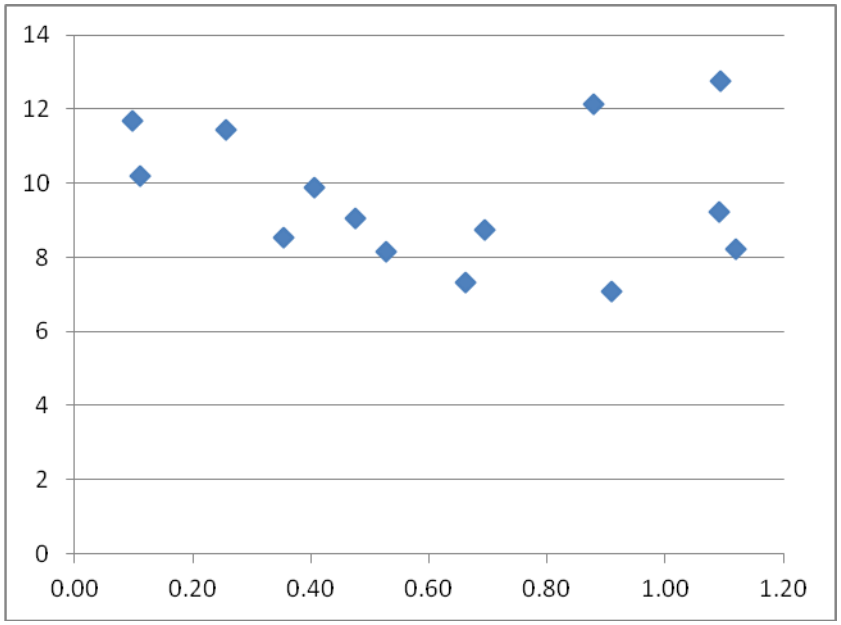

Figure 1. WIF(BR) (x-axis) compared with ISI citations per paper (y-axis).

Figure 2 compares the WIF(BR) with the QS World Rankings Score. This indicates that WIF(BR) had a small degree of correlation with the QS score, with some outliers, as indicated by the trend line (Excel CORREL score 0.282).

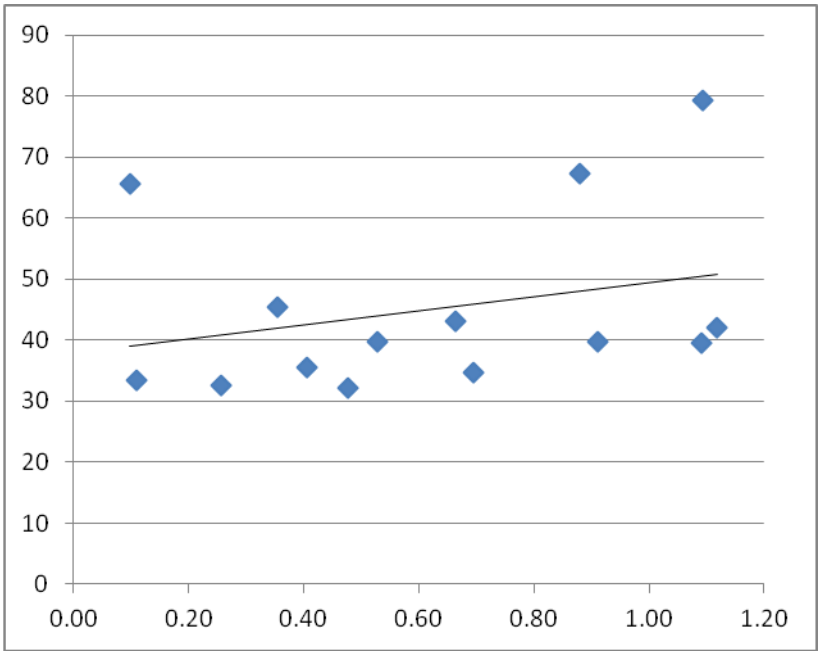

Figure 2. WIF(BR) (x-axis) compared with QS World Rankings score (y-axis).

For Australian institutions, WIF(BR) had a small degree of correlation with the ERA score for the institution, as shown in Figure 3 (Excel CORREL score 0.196).

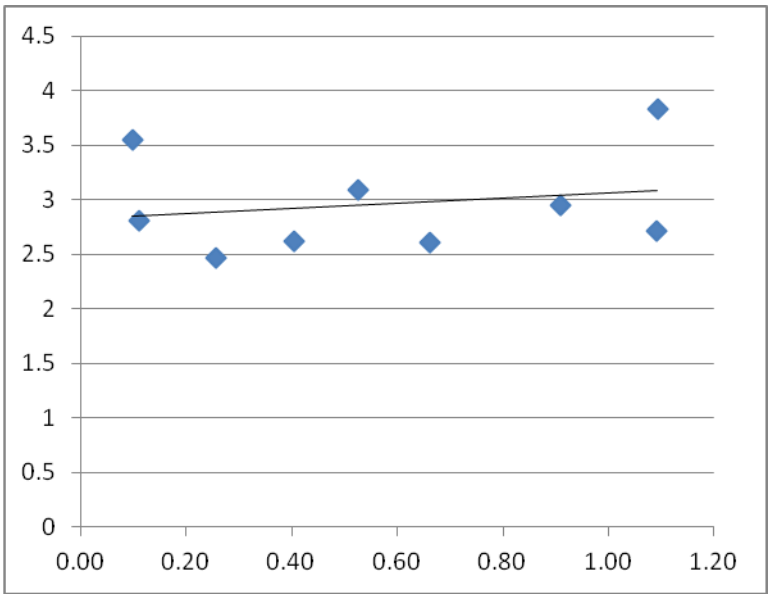

Figure 3. WIF(BR) (x-axis) compared with ERA score (y-axis), for Australian institutions.

For the small number of NZ institutions for which Blekko gave useful data, there did not seem to be any correlation between the WIF(BR) and the PBRF Quality Score (Excel CORREL 0.002). This is illustrated in Figure 4.

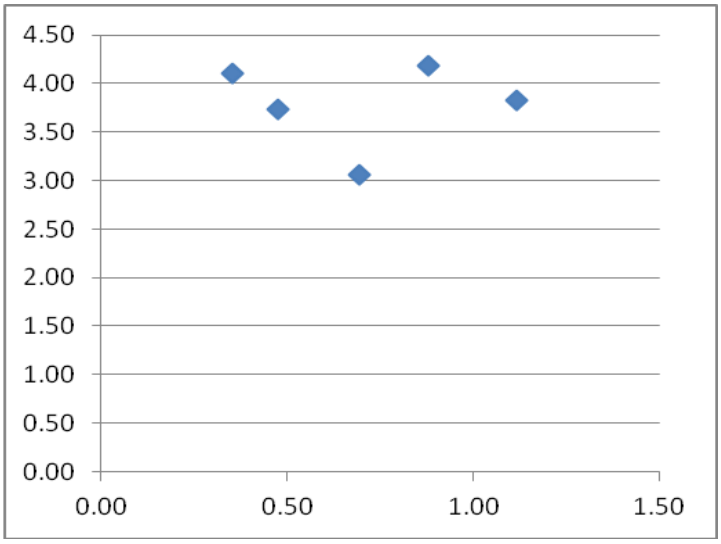

Figure 4. WIF(BR) (x-axis) compared with PBRF Quality Score (y-axis), for NZ institutions.

Discussion

The low level of coverage of the institutional repositories indicates that Blekko currently has limitations as a webometric tool. In the current study, there appeared to be significant variations in coverage of the repositories, and for that reason only a subset of the institutions was used in the analysis. It is possible that the utility of Blekko for webometric research will improve if the search engine robot achieves a greater depth of indexing of the sites, and more consistency of coverage. The lack of information about internal links is also a drawback, since for many institutions, it was not possible to determine the number of external versus internal inlinks. The limitations of Blekko are unfortunate, since it currently appears to be the only search engine that provides link data for whole domains.

In view of the limitations of Blekko's coverage of links to institutional repositories, and the small number of institutions for which reasonable coverage was achieved, it is hard to put much weight on the results. However there appeared to be a potential relationship between the WIF of the institutional repository, and two conventional measures of research quality: the QS World Rankings score and the ERA score. Interestingly there was if anything a negative relationship between the WIF and the citation based measure, the ISI citations/paper.

It is also possible that web links to institutional repositories, because they come from different types of sources, are measuring a different type of research impact from the conventional measures. Previous research (Smith 2011, Sato and Itsumura 2011) indicates that significant links are made from sites such as Wikipedia to institutional repositories, so the web links to institutional repositories may indicate the impact of the institution's research on the general non-academic audience, as opposed to the narrow research community. WIFs for institutional repositories may complement conventional measures of research impact, rather than being an alternative.

As yet, many institutional repositories only contain a small percentage of the institution's research (Mustatea 2008) so for most institutions the material in the institutional repository is not representative of the research output as a whole. It might be useful to investigate WIFs for repositories that have a high coverage of their institution's output.

While this exploratory research is inconclusive, it indicates that further webometric evaluation of institutional repositories is warranted, particularly if higher quality tools become available.

References

Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology*, *57*(10), 1296-1306.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics". *Journal of Documentation*, 53(4), 404–426.

Chambers, G. K., Gardner, J. P. A., & Smith, A. G. (2010). Publish and perish: A new look at bibliometric statistics in the PBRF age. *New Zealand Science Review, 67*(4), 106-113.

Eccles, K. E., Thelwall, M., & Meyer, E. T. (2012). Measuring the web impact of digitised scholarly resources *Journal of Documentation*, 68(4), 512-526.

Hare, J., & Trounson, A. (2011). Excellence in research for Australia lays bare research myths. *The Australian*, 2 February 2011. http://www.theaustralian.com.au/higher-education/excellence-in-research-for-australia-lays-bare-research-myths/story-e6frgcjx-1225998312883

Mustatea, N. (2008). *To what extent is material in institutional repository representative of an institution's research output?* Master of Library and Information Studies research project, Victoria University of Wellington.

Sato, S., & Itsumura, H. (2011). How do people use open access papers in non-academic activities? A link analysis of papers deposited in institutional repositories. *Library, Information and Media Studies, 9*(1), 51-64.

Scholze, F. (2007). Measuring research impact in an open access environment. *Liber Quarterly: The Journal of European Research Libraries*, *17*(1-4), 220-232.

Shukla, S. H., & Poluru, L. (2012). Webometric analysis and indicators of selected Indian state universities. *Information Studies*, *18*(2), 79-104.

Smith, A. G., & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363–380.

Smith, A. G. (2011). Wikipedia and institutional repositories: An academic symbiosis? *Proceedings of the ISSI 2011 Conference, Durban, South Africa, 4-7 July 2011*. http://www.vuw.ac.nz/staff/alastair_smith/publns/SmithAG2011_ISSI_paper.pdf

Tertiary Education Commission (2006) *TEO results – All TEOs*. http://www.tec.govt.nz/Resource-Centre/Reports/PBRF-TEO-results---all-TEOs/

Zuccala, A., Thelwall, M., Oppenheim, C., & Dhiensa, R. (2007). Web intelligence analyses of digital libraries: A case study of the national electronic library for health (NeLH). *Journal of Documentation*, 63(4), 558-89.

Appendix: Table 1. Web Impact Factors and other measures of research quality.

Institution	*WIF(BR)*	*ISI citation/paper*	*QS score*	*ERA Score (Australia)*	*PBRF (NZ)*
University of Sydney	1.09	12.77	79.3	3.83	
University of Adelaide	0.10	11.69	65.7	3.55	
Queensland University of Technology	0.53	8.16	39.8	3.09	
University of Technology Sydney	0.91	7.09	39.7	2.95	

University of Tasmania	0.11	10.19	33.5	2.81	
University of Wollongong	1.09	9.23	39.6	2.71	
La Trobe University	0.40	9.87	35.5	2.62	
RMIT	0.66	7.33	43.1	2.61	
James Cook University	0.26	11.45	32.6	2.47	
University of Auckland	0.88	12.12	67.3		4.19
University of Canterbury	0.35	8.54	45.5		4.10
Victoria University of Wellington	1.12	8.23	42.1		3.83
Massey University	0.69	8.75	34.6		3.06
University of Waikato	0.48	9.04	32.2		3.73