Web Based Impact Measures for Institutional Repositories

Alastair G Smith

[1] *alastair.smith@vuw.ac.nz*
Victoria University of Wellington/ Te Whare Wānanga o te Ūpoko o te Ika a Māui, School of Information Management, Rutherford House, 23 Lambton Quay, Wellington 6140, (New Zealand/Aotearoa)

Abstract
This study investigated webometric measures that could be used to evaluate the impact of institutional repositories, using Australasian university repositories as a case study. URL citation inlinks (occurrences of the repositories' URL in the text of web pages), as found through Google searches, were counted. As well as links from the general web, links made from other Australasian academic institutions and from Wikipedia were counted.

For repositories with significant deposit ratios, there appeared to be a small correlation between the URL citation inlinks from other Australasian academic institutions, and some conventional measures of research impact: the ISI citations/paper, the QS ranking score, and the ERA quality score.

Repositories with higher deposit ratios appeared to achieve more inlinks from other Australasian academic institutions, indicating the value of repositories encouraging high deposit rates of their institutions research output.

Institutions with repositories that had a high Wikipedia Web Impact Factor were not necessarily highly ranked in terms of inlinks from other tertiary institutions or ISI citations per paper. This indicates that repositories impact on the general web is different from their impact on the research community.

Conference Topic
Webometrics (Topic 7); Scientometric Indicators (Topic 1)

Introduction
Institutional Repositories are a common way for institutions to make their research outputs available. This study investigated opportunities for webometric evaluation of the research done by institutions through their repositories.

Most commonly webometric studies use inlink counts: links made from other websites to the site being studied. These are viewed as being analogous to citations in conventional publishing (Gairin, 1997). Either total inlink counts are used, or a Web impact factor (Almind & Ingwersen, 1997), analogous to the Journal Impact Factor for conventional publications. The Web Impact Factor is defined as the ratio of the inlink counts to a measure of the size of website, for example the number of pages.

There are a number of ways reported in the literature for counting the inlinks made to a website such as an institutional repository. Unfortunately several tools used in the past are no longer available. An early study of Web Impact Factors for Australasian universities (Smith & Thelwall, 2002) used Alta Vista to identify pages linking to the universities, but this tool is no longer available. Yahoo Site Explorer was used by many studies, for example to create a ranked list of world class universities (Ortega & Aguillo, 2009), but since Yahoo has been merged into Bing, the Site Explorer tool no longer provides useful link data. Thelwall and Sud (2011) reviewed alternative methods of estimating the online impact of organisations, including URL citation inlinks (discussed below) and organisational title mentions.

The current study used the technique of URL citation inlinks, proposed by Kousha and Thelwall (Kousha & Thelwall, 2007). This uses a search engine such as Google to locate in-text mentions of URLs associated with an institutional repository, which can be assumed to be links to documents at the repository.

There have been other webometric studies of institutional repositories. Placing research materials in repositories was found to increase the amount of data available for bibliometric analysis (Scholze, 2007). Zuccala *et al* (2007) studied an institutional repository by using web link analysis and server logs in order to investigate how users located and used the repository. An analysis of the web presence of Indian state universities (Shukla & Poluru, 2012) found that open access institutional repositories were helpful in increasing the visibility of institutions on the Web. A range of webometric measures have been used to create the Ranking Web of World Repositories (http://repositories.webometrics.info) (Aguilllo *et al* 2010) in order to support the use of repositories for scientific evaluation purposes.

A previous paper (Smith, 2012) found little correlation between impact measures of institutional repositories calculated from a new search engine Blekko (http://blekko.com/), and conventional research impact measures. The paper suggested that since links to institutional repositories are different in nature from conventional measures of research impact, measures for institutional repositories should be considered to be complementary to conventional measures, rather than directly comparable.

In the current study, impact measures were calculated based on:

- Links from the general Web
- Links from academic domains, which might be considered to be more equivalent to conventional measures of research impact
- Links from Wikipedia, which might be considered to be more indicative of the impact that the repository has in making research available to the lay community. A previous paper (Smith, 2011) identified a significant number of links from Wikipedia to institutional repositories, and proposed that the value of institutional repositories may lie in making research available to the general Web community, rather than to the research community.

Research Questions

This study addressed the following questions:

1. What impact factor measures are appropriate for evaluating the impact of institutional repositories on the Web?
2. Do web based impact factors for institutional repositories correlate with conventional impact measures of the research output of institutions?
3. Do institutional repositories have a greater impact if a higher proportion of their research output is in the institutional repository?
4. Are there specific impact factors that reflect the different impact that institutional repositories have?

Methodology

This study investigated institutional repositories at tertiary institutions in Australasia (Australia and New Zealand). These were identified from ROAR (http://roar.eprints.org/). Repositories with less than 1000 items reported in ROAR were excluded, resulting in 39 institutional repositories being included in the study.

Google was searched with a formulation that identified pages that contained URL citation inlinks. The search excluded links from the institution itself (on the argument that these would be likely to be navigational links or self citations). Google was set not to record search history or use previous searches in interpreting the search formulation. This is important, since Google by default attempts to optimise the search on the basis of a users previous searching, which of course is counterproductive for webometric work. The data was collected in January 2013.

A typical formulation, for example for University of Auckland, was:
`"researchspace.auckland.ac.nz" -site:auckland.ac.nz`
This searched for web pages which included, in the text of the page, the basic URL of the archive, and excluded pages on the University of Auckland site.

Initially, only links from the institutional repository itself were excluded, for example:
`"researchspace.auckland.ac.nz" -site:researchspace.auckland.ac.nz`
However scanning the results indicated that this still included many pages at the institution which were either navigational in nature, or self citations (a staff member linking to their publications from their home page, for example), and it was decided that the formulation excluding all links from the institution was more appropriate.

In some cases an institutional repository had more than one URL, for example Australian National University required a formulation:
`"digitalcollections.anu.edu.au" OR "dspace.anu.edu.au" -site:anu.edu.au`
This searched for web pages that included, in the text of the page, either of the basic repository URLs, but excluded pages at the ANU site.

The Web Impact Factor for each repository was calculated by dividing the URL citation inlinks count by the number of documents in the institutional repository. The number of documents in the institutional repository was taken from ROAR.

In addition to the basic URL citation inlinks count, some counts were done of inlinks from specific types of domains.

An *Academic Institution Inlinks Count*, which might be more comparable to the citations made between research publications, was found by adding to the basic formulation a requirement that linking pages were in the Australasian academic domains, edu.au or ac.nz. So for example the formulation for University of Auckland became:
`"researchspace.auckland.ac.nz" -site:auckland.ac.nz site:edu.au OR site:ac.nz`
This of course is only identifying links from Australasian academic institutions, and a more global formulation would include all academic domains (.edu, .ac.uk, individual domains of European academic institutions which generally have second level domain, for example uni-muenchen.de, etc). However this would lead to an excessively complex search formulation and it was decided that links within Australasia would give sufficient indication of the viability of the concept of an educational inlinks count.

An Academic Inlinks Web Impact Factor was calculated from the academic inlinks count, by dividing the academic institution inlinks count by the number of documents at the repository.

To measure the impact of the institutional repositories on the general lay community, a *Wikipedia inlink count* was calculated by adding to the search formulation a requirement that

links were made from Wikipedia. So for example the formulation for University of Auckland became:

```
"researchspace.auckland.ac.nz" -site:auckland.ac.nz site:wikipedia.org
```

A Wikipedia Web Impact Factor was calculated from this by dividing the Wikipedia inlinks count by the number of documents at the repository. Due to the relatively small number of Wikipedia inlinks in relation to the number of documents in the repository, the Wikipedia Web Impact Factor was multiplied by 1000 to give a whole number.

Several conventional measures of research impact were identified and used as comparisons with the impact measures calculated in the study. These were:

- The number of citations/paper for each institution, taken from Thomson/ISI's *Essential Science Indicators*, part of the Web of Knowledge. The version used covered documents indexed by ISI in 2002-2012.
- The overall score from the QS world rankings of universities (http://www.topuniversities.com/university-rankings/world-university-rankings/2011). The QS rankings are a widely accepted measure of the quality of academic institutions worldwide. Only 26 institutions in the current study had QS ranking scores, since only the top 400 institutions worldwide were published.
- An average research excellence score derived (Hare & Trounson, 2012) from the 2010 ERA (Excellence in Research for Australia) research assessment carried out by the Australian Research Council. This of course was only available for the 29 Australian institutions in the study.

The study also looked at the extent to which an institutional repository contained a significant proportion of the research output of the institution. Mustatea (2008) found that many institutional repositories only contain a small proportion of the institution's publications when compared with the publications indexed in the ISI databases. In the current study, a ratio, the *Deposit Ratio*, was calculated. This was the ratio of documents deposited in the institutional repository, compared with the number of papers indexed by ISI in Essential Science Indicators. This is of course a crude measure, since there will be documents in the repository that would be not appropriate for indexing by ISI, and outputs indexed by ISI that may not be deposited in the repository for copyright or other reasons.

Results

The study addressed the first research question, "What impact factor measures are appropriate for evaluating the impact of institutional repositories on the Web?" by investigating a range of measures based on URL citation inlink counts to the repositories. The usefulness of these measures is addressed in the answers to the following research questions.

The second research question asked "Do web based impact factors for institutional repositories correlate with conventional impact measures of the research output of institutions?" No correlation was found between the different Web Impact Factors and the conventional measures of research impact. While it is disappointing that the Web Impact Factor of institutional repositories appears not to be a useful substitute for conventional measures of research impact, it is not surprising. As mentioned earlier, links to institutional repositories come from different sources, and are made for different reasons, than the academic citations on which conventional measures are partly based. Also, the documents in a repository may include materials not representative of the institution's research output, including for example student work and digitised material such as historical photographs. So a

Web Impact Factor calculated on the basis of the raw number of documents in the repository may not be a good measure of the inlinks per research output.

When total inlink count is considered, the picture changes a little. For the group of repositories as a whole there is no appreciable correlation between the total inlink count and the conventional measures. However if only repositories that had an Deposit Ratio of more than 1 (the ratio of the number of documents in the repository to the number of papers indexed by ISI was greater than 1) were considered, there appeared to be small but positive correlations between the total link count from educational institutions and the conventional measures: the ISI citations/paper, the QS score, and the ERA score.

Given the small numbers, the best indication of the relatively weak correlations are the scattergraph representations in Figures 1-3. For reference, the Pearson correlation coefficients are also included. Note that QS and ERA scores were not available for all institutions. Only repositories with a Deposit Ratio greater than 1 are included.

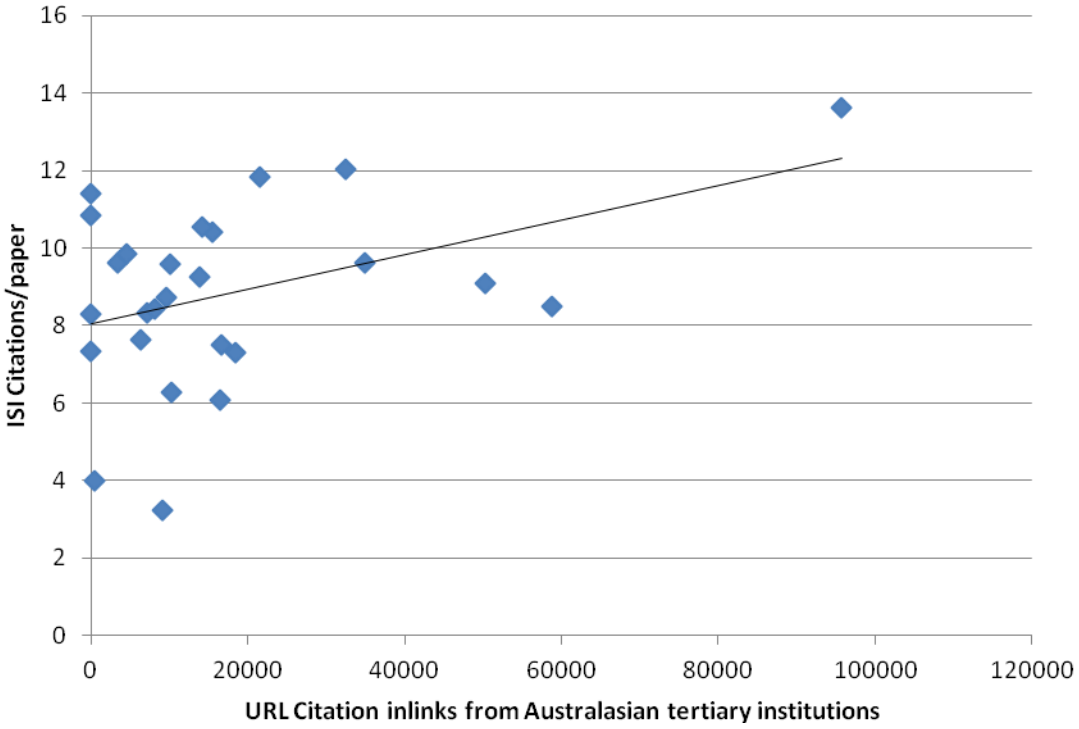

Figure 1. URL citation inlinks from Australasian tertiary institutions (Deposit Ratio>1) compared with ISI citations per paper (Pearson correlation coefficient 0.15).

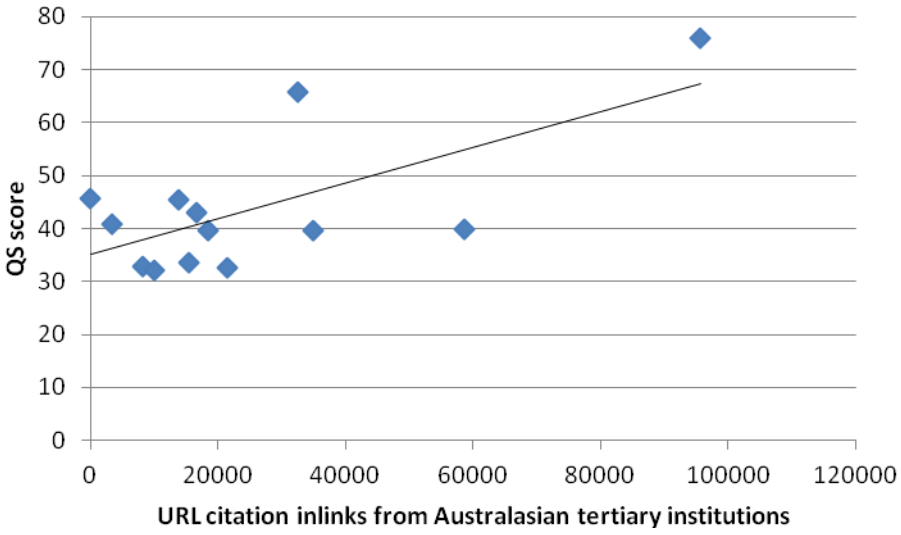

Figure 2. URL citation inlinks from Australasian tertiary institutions (Deposit Ratio>1) compared with QS score (Pearson correlation coefficient 0.14).

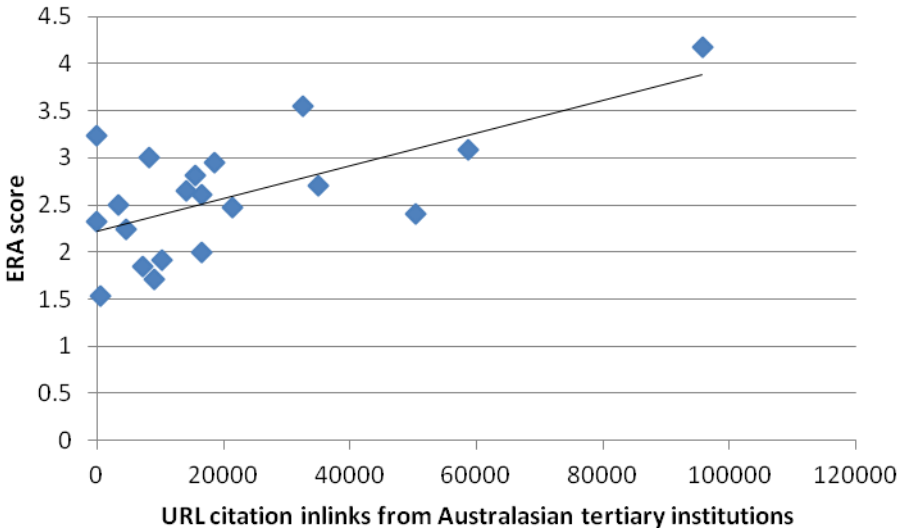

Figure 3. URL citation inlinks from Australasian tertiary institutions (Deposit Ratio>1) compared with ERA score (Pearson correlation coefficient 0.27).

This indicates that there may be value in calculating inlink counts that come from other research institutions, since these are likely to reflect the research value of the material in the repository. However the weak correlation means that further research using a larger set of repositories is needed, and that inlink counts for the institutional repository are unlikely to be a substitute for conventional measures of the research impact of institution as a whole.

Addressing the third research question, "Do institutional repositories have a greater impact if a higher proportion of their research output is in the institutional repository?" the study compared the Deposit Ratio of the repositories with URL citation inlink counts. This appeared to show a positive correlation. An indication of the relationship is shown graphically in Figure 4.

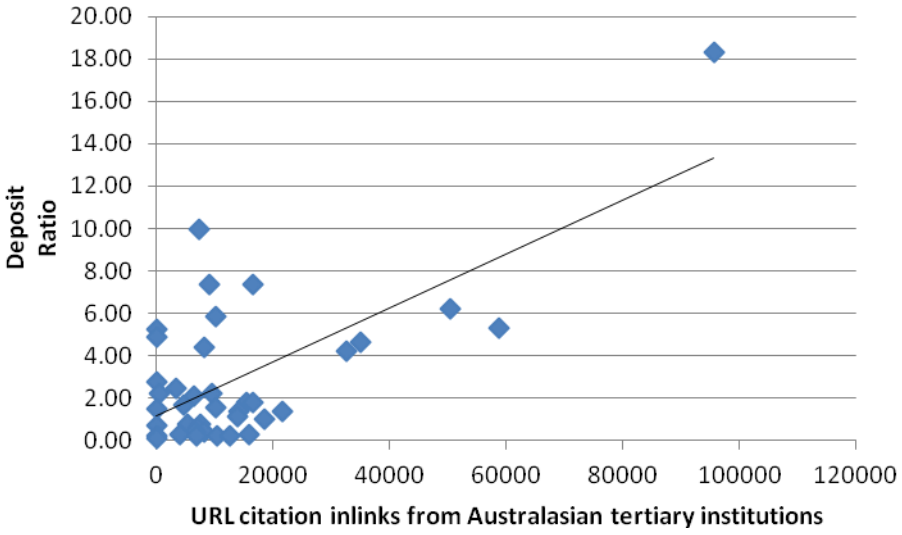

Figure 4. URL citation inlinks from Australasian tertiary institutions compared with Deposit Ratio of repository (Pearson correlation 0.68).

If this correlation is real, it appears to indicate that there is value in an institution maximising the number of research outputs in its repository, for example through mandatory deposit of publications. Higher deposit rates increase the visibility of the repository, and the chances of links being made from other research institutions.

In addressing the fourth research question, "Are there specific impact factors that reflect the different impact that institutional repositories have?" the study looked at the URL citation inlink count for links coming from Wikipedia, as a way of gauging the impact of the repositories on the general lay Web community. Neither the Wikipedia inlink count, nor the corresponding Wikipedia Web Impact Factor, correlated with the conventional measures of research impact. This is to be expected, since Wikipedia has a different purpose than research publishing. However the Wikipedia inlink count and the Wikipedia Web impact factor provide measures of the extent to which the repository is having an impact on the general lay Web community. Table 1 shows the top 10 institutional repositories by Wikipedia Web Impact Factor (multiplied by 1000 to bring the figure to an integral number). The Web Impact Factor has been chosen in this case because Wikipedia is likely to reference many of kinds of documents that are held in institutional repositories, for example photographs. For comparison, the table also shows these institutions' ranks by the inlink count from Australasian academic institutions and by their ISI Citations per paper.

Table 1. Top 10 institutional repositories by Wikipedia Web Impact Factor.

Institution	*Wikipedia WIF (x1000)*	*Academic Citation Inlink Rank*	*ISI Citation/paper Rank*
1. University of Sydney	60	14	5
2. University of Waikato	19	17	21
3. Victoria University of Wellington	15	24	27
4. Bond University	13	16	36
5. Flinders University	12	25	16

6. University of Technology Sydney	10	7	35
7. Massey University	9	15	23
8. University of Otago	5	30	4
9. University of Tasmania	5	11	15
10. University of Canterbury	5	13	22

This indicates that institutions that have repositories with a significant impact on the general Web may not be those that have high impact in the research community.

In this study, links from Wikipedia were investigated, but of course links from other kinds of Web sites could be measures of the impact of a repository on the general Web. For example, links from blogs, Twitter feeds, Facebook, etc could be investigated.

Conclusions

This exploratory study has investigated a range of webometric measures to evaluate the impact of institutional repositories. This is important given the resources that many research institutions are investing in their repositories.

It appears that the conventional web impact factor of institutional repositories does not correlate with conventional measures of research impact. This may be due to the number of documents in a repository not corresponding well with the conventional research output of an institution, as well as links being made to repositories for different reasons than citations are made to conventional publications. However there appears to be a small correlation between the number of links made to repositories from other academic institutions, and some conventional measures of research impact, for repositories with a high deposit ratio. This indicates that webometric measures based on links from research institutions to institutional repositories could be useful evaluation tools; particularly if the repositories achieve high deposit rates of the institutions research output. This also indicates that there may be scope for specialist web crawlers, such as the University of Wolverhampton's SocSciBot (http://socscibot.wlv.ac.uk/), to evaluate repositories for their research impact, since it appears that measures based on links from other research institutions, rather than the general web, are most valuable in terms of evaluating the research impact of the repository. There may also be value in structuring repositories in such a way that research material is differentiated from other material such as student work and digitised images.

The study also indicates that the research impact of a repository, as measured by the inlink counts from other tertiary institutions, may be enhanced by high deposit rates. While this is not surprising, it is a useful indicator to institutions that it is beneficial to encourage researchers to deposit their work in the repository.

The study also investigated measures of the repositories' impact on the general web. The specific example of links from Wikipedia was explored, showing that institutions whose repositories had a high impact in terms of their Wikipedia Web Impact Factor were not necessarily those that had a high conventional research impact. This reflects institutional repositories' value as a way of making research available to the general Web community, as well as to the research community.

This exploratory study is limited by being carried out on a limited number of institutions in a specific geographic area. Future research should see if the findings can be replicated over a

broader sample of repositories. There is also scope for studies of the repositories' impact on the general Web, looking at websites such as blogs, Twitter and Facebook.

References

Aguillo, I., Ortega, J., Fernández, M., & Utrilla, A. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics*, 82(3), 477-486.

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

Gairin, J. R. (1997). Valoracion del impacto de la informacion en internet: Altavista, el "citation index" de la red. impact assessment of information in the internet: Altavista, the citation index of the web. *Revista Espanola De Documentacion Cientifica*, 20(2), 175-81.

Hare, J., & Trounson, A. (2012). Excellence in research for Australia lays bare research myths. *The Australian*, (4/12/2012) Retrieved January 18 2013 from: http://www.theaustralian.com.au/higher-education/excellence-in-research-for-australia-lays-bare-research-myths/story-e6frgcjx-1225998312883

Kousha, K., & Thelwall, M. (2007). Google scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science & Technology*, 58(7), 1055-1065.

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.

Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, 45(2), 272-279.

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. (2011). Wikipedia and institutional repositories: An academic symbiosis? *Proceedings of the ISSI 2011 Conference*, Durban, South Africa. (pp. 794-800)

Smith, A. G. (2012). Webometric evaluation of institutional repositories. *Proceedings of the 8th International Conference on Webometrics Informetrics and Scientometrics (WIS) & 13th COLLNET 2012 Meeting, Seoul, Korea*. (pp. 722-729).

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

Thelwall, M., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organizations. *Journal of the American Society for Information Science and Technology, 62*(8), 1488-1497.

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.