Substance without Citation: Evaluating the Online Impact of Grey Literature[1]

David Wilkinson, Pardeep Sud, Mike Thelwall
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.
E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470, Fax: +44 1902 321478
Keywords: Grey literature, web impact, webometrics.

Abstract
Individuals and organisations producing information or knowledge for others sometimes need to be able to provide evidence of the value of their work in the same way that scientists may use journal Impact Factors and citations to indicate the value of their papers. There are many cases, however, when organisations are charged with producing reports but have no real way of measuring their impact, including when they are distributed free, do not attract academic citations and their sales cannot be tracked. Here, the Web Impact Report (WIRe) is proposed as a novel solution for this problem. A WIRe consists of a range of web-derived statistics about the frequency and geographic location of online mentions of an organisation's reports. WIRe data is typically derived from commercial search engines. This article defines the component parts of a WIRe and describes how to collect and analyse the necessary data. The process is illustrated with a comparison of the web impact of the reports of a large UK organisation. Although a formal evaluation was not conducted, the results suggest that WIRes can indicate different levels of web impact between reports and can reveal the type of online impact that the reports have.

Introduction

Performance metrics are widely used in organisations. In industry, for example, there are numerous metrics of three kinds: input, output and process (Geisler, 2000) although for most businesses the ultimate metrics are perhaps profitability and predicted future profitability. For organisations that produce information, knowledge or cultural artefacts, however, profit is not always relevant or the most important indicator of success. For example publishers may count profits as of primary importance, but authors may value total sales or the receipt of literary awards above author royalties alone. Similarly, artists may rely upon revenues from sales of their work but regard exhibitions in prestigious galleries or prizes as their primary goals. In academia, profit seems rarely to be a primary consideration because researchers typically give away their work to conferences and journals without charge. Nevertheless, in recent times, governments in many countries have increasingly asked for evidence of value for money in terms of societal impact (Bornmann, 2013).

There seem to be currently three key types of indicators for researchers' work: income generation, citations, and peer review. In Australia, scientists are primarily judged and rewarded by their ability to attract external funding (Butler, 2003). In The Netherlands and the Post-2007 UK Research Excellence Framework, researchers in some areas of science are judged partly on the rate at which their work attracts citations (Bence & Oppenheim, 2004; Moed, 2005). Citations, or citation-related

[1] This is a pre-final version of: Wilkinson, D., Sud, P., & Thelwall, M. (in press). Substance without citation: Evaluating the online impact of grey literature. *Scientometrics*.

journal Impact Factors, have also been used by other governments (e.g., China, Spain, Finland). In New Zealand and, for a considerable time, the UK, peer review has been used to rate researchers in order to target government money at the most successful (e.g., http://www.ref.ac.uk/). Sometimes, however, knowledge workers are asked to produce outputs for the public domain in non-academic formats that cannot be effectively measured using citation counts. This information could be in the form of a free online or paper report, for instance (Jeffery, 2000). Examples include most public service reports, such as those identifying and advocating healthy lifestyles, those recommending new business strategies or styles, and those producing background information of wide value, such as the Oxford Internet Institute's Internet usage surveys (e.g., Dutton & Elsper, 2007). The reports may be produced by individual researchers or small groups of researchers as part of a funded research project or by specialist organisations on a contract basis, such as *empirica GmbH* or *Idea Consult*, or may be produced by government organisations as part of a wider remit, for example the National Endowment for Science, Technology and the Arts (NESTA) business innovation strategy publications. Although white papers seem not to be valued and there does not seem to be any published research about how to assess them (other than using traditional citation analysis for research-related white papers), they are significant in some contexts. For example, a survey of 141 marketing managers in large UK computer service companies found that 89% read the grey literature (mainly from the web) but only 2% read marketing journals (Bennett, 2007).

The reports described above may be targeted at a dispersed audience (e.g., the general public, entrepreneurs) for which it is not easy to find direct indicators of uptake or impact. To illustrate this, it would not be reasonable to evaluate a report advocating healthy lifestyles by whether healthy living became more widespread after the report because of the multiple other influences simultaneously operating (e.g., unhealthy food advertising campaigns). It can also be expensive to evaluate behaviour changes in the public and whilst controlled experiments or surveys are often used for this (Lefebvre & Flora, 1988), these can be inappropriate, particularly for reports aimed at generating awareness or advocating non-personal changes. The same is true for all except the largest advertising campaigns aimed at changing public behaviour (i.e. social marketing). In addition, advocacy publications can be useful as part of a long term strategy even if they have no measureable policy impacts (Baumgartner, 2007).

In this article the Web Impact Report (WIRe) is proposed as an indicator to help evaluate the impact of reports for which the audience is primarily non-academic. This does not include academic preprints (Gentil-Beccot, Mele, Brooks, 2010), for which citations are probably adequate. The underlying assumption is that counting the number of times a report is mentioned on the web may give useful, albeit partial, information about its impact. An important advantage of WIRes is that they are relatively inexpensive and hence can be conducted periodically as part of an on-going monitoring process. For example a WIRe could be conducted annually to identify the most successful publications produced by an organisation during the previous year. Assuming that an organisation's reports are not designed to have an online impact, a WIRe will be an indirect indicator and most valuable when used in conjunction with a range of other indicators, such as newspaper mentions.

The idea of measuring web impact is not new: the term Web Impact Factor was coined by Peter Ingwersen (1998) for metrics based upon counts of links to a collection of web pages. In addition, other authors have evaluated the impact of academic journal articles through various online measurements (Kousha & Thelwall,

2007; Vaughan & Shaw, 2003), the impact (loosely speaking) of authors online (Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998) and the spread of an issue online (Thelwall, Vann, & Fairclough, 2006). Nevertheless, no previous research has set out and evaluated a method for estimating the impact of non-academic reports (e.g., white papers, online magazines or newsletters, promotional leaflets) using online methods. This task is very similar to web issue analysis except that there is a need for a comparative approach in order to deliver a more convincing impact evaluation, and that less information about the context of the online invocations of the report or campaign is needed.

Web Impact Reports

The goal of a Web Impact Report (WIRe) is to evaluate the web reflection of a collection of documents, whether published online or offline. This is achieved by identifying and characterising in various ways the online mentions of these documents in three different ways: Web location, topic, and genre. Some of these characterisations can be conducted automatically, whereas others require a human classifier. The proposed steps are listed below together with explanations and justifications.

1. Lists and counts of web pages citing the documents

The most obvious way of constructing an online indicator for a document is to count the number of web pages that mention it. The simplest way to do this is to construct a query that matches the document and record Google's hit count estimate for this query. The query could be a phrase search for the document title if it was unique enough to never occur in other contexts. For non-unique titles, compound queries must be constructed so that virtually all matches are mentions of the document (i.e., a high precision query). These compound queries would start with the document title and include some other information likely to be mentioned almost always alongside the document, such as its authors or the name of the organisation publishing it. In some cases this is not possible – particularly for documents with short titles, without attributable authors and published by organisations with widely used names (e.g., NESTA's "Hidden innovation" report).

This Google hit count estimate is a quick and simple way of getting very approximate impact estimation for a collection of documents but the reliability and information content of the results can be substantially improved. Instead of hit count estimates, complete lists of matching URLs can be obtained. This is an improvement because the hit count estimates can be unreliable (Thelwall, 2008; Uyar, 2009) and because additional information can be extracted from URL lists, as described below. Full URL lists can be extracted from the results pages manually but this can be time-consuming so the use of software like Webometric Analyst (http://lexiurl.wlv.ac.uk) is recommended to automate this although this program is only able to use Bing. A problem arises if there are more than 1,000 results because search engines never return any results after the 1,000th. The "query splitting" technique has been designed to resolve this issue by automatically constructing new queries to retrieve additional results (Thelwall, 2008). This is available in Webometric Analyst. Individual search engines cover only a fraction of the web and so it is recommended to combine the results of at least two search engines in order to get a more comprehensive list. This may not be practical because it requires manual extraction of results for search engines other than Bing that do not allow automatic querying.

Once lists of URLs of pages matching the document title queries have been constructed as above then these URL lists can be processed to give more robust statistics than URL counts. More specifically, it is better to report counts of matching web sites than counts of matching web pages (Thelwall, 2009). This is because there are many reasons why a single action could result in a document being mentioned on multiple pages within a web site. For example, the document could be mentioned in a report that is available in both PDF and HTML formats, or the document could be mentioned in a single blog entry but the entry could be accessible by an individual post URL and a URL for a blog archive page containing all posts for a single month. There are two ways in which the web site of a URL can be easily and automatically identified: by the full domain name or by the server-level domain name ending (e.g., wlv.ac.uk for the University of Wolverhampton because all of its domain names end in .wlv.ac.uk). The latter, called site-level in webometrics terminology (Thelwall & Wilkinson, 2008), is only recommended in cases where URLs come from many different domain names but these domain names originate within a few large web sites. This is most likely to occur if a document is somehow adopted by a large organisation, like a university, and mentioned frequently on its constituent mini-web sites (e.g., for departments). Hence, the most appropriate indicator of online impact for a document is usually the number of unique domain names in the list of URLs of matching pages.

Some additional information can also be extracted from the domain names of the URLs of matching pages; their top-level domains (TLDs). For TLDs associated with countries, this can give a useful indicator of the international spread of a collection of documents. This is not a strong indicator because very common TLDs like com, net and org give no indication of the origins of the domain name. Nevertheless, if TLDs are extracted from long lists of domain names then it seems reasonable to infer that a document mentioned in a greater range of country code TLDs is more likely to have a wider international impact.

2. Content analysis

Whilst the counts of web sites and TLDs discussed above give impact indicators that are simple and easy to compare between documents, they give no idea about why the documents were mentioned online or about how they were used. To fill this gap, a random sample of pages should be visited, read and classified so that the main reasons can be summarised and reported. The sample should be taken with a maximum of one per domain name because the main impact measure is based on counts of domain names. The classification scheme should be constructed inductively, perhaps starting with an initial list of relevant classes and expanding the list to incorporate new unanticipated contexts. The classes and genres should be chosen to reflect the objectives of the investigation as well as the types of pages found online (Neuendorf, 2002; Wilkinson, Harries, Thelwall, & Price, 2003).

3. Research citation index

The web pages that mention a document may vary in importance. For example, a page could be a catalogue for an online book shop, which has little value in terms of indicating interest. In contrast, another mention might be a citation in an academic journal article or a carefully-prepared important government report, both of which may have considerable value for indicating impact. As a result, it is useful to separate out and give special treatment to the citing or mentioning documents that have the highest value. There is a simple way to identify a sample documents that are likely to

have higher value than average: their file format. Important documents are often published online in PDF format and academic documents are sometimes also posted as word processor files. Hence, it would be useful to identify a collection of documents that is likely to have a higher proportion of high quality content than general web pages. For this, instead of manually filtering the full list of matching URLs from part 1 above, new type-specific searches for Word, PDF, open document format or other types can be conducted first to construct a list of high value documents (e.g., adding filetype:pdf, filetype:doc, filetype:docx or filetype:odf to each original search). If the organisation believes that other high value types of document may tend to be published online in HTML format then it may be possible to conduct additional searches for these by adding additional keywords to the basic searches for them (e.g., "syllabus" to focus on academic course descriptions) to narrow down he results to appropriate types of document.

The document type searches will produce a list of URLs of citing documents and a count of these URLs could be reported as impact indicators but it would be better to conduct additional filtering first to improve the results. This can be achieved by constructing a list of the key information from each one, such as its title and authors, and then using this list to eliminate duplicates. This approach, although manual and time-consuming, is necessary because a big organisation may publish several important citing documents on their web site so it would be desirable to count each document separately. In addition, important reports are sometimes reposted elsewhere on the web and so it is useful to eliminate the reposted copies from the statistics. Finally, and perhaps most importantly, the *research citation index* produced as a result of this is a list of the most important documents citing the set studied, so the complete list is of value for the document owners to browse in order to quickly get an idea of the main sources of impact of their work.

Case study: NESTA

This section describes a WIRe commissioned by NESTA, which funds innovative ideas, researches innovation and attempts to promote innovation in the UK. One of its strategies to tie the latter two together is to publish a series of reports. These include commissioned short "provocations", which are pamphlets that discuss an idea and seek to be provocative. For example, "Beginning at the Beginning" by Anthony Sargent and Katherine Zeserson was designed to "examine and challenge the traditional place of creativity in UK society". NESTA also produces more substantial research reports which are published in the form of glossy pamphlets. One example is "Total Innovation", which, "examines why harnessing the hidden innovation in high-technology sectors is crucial to retaining the UK's innovation edge". All these pamphlets are published and distributed free as printed documents and/or are made available on the NESTA web site, which also offers a short summary of each one.

NESTA requested twice-yearly web impact reports on its documents (and its web site) to assess their impact. Without webometrics, the only practical way to assess the impact of the documents would be to count the number of times they were mentioned in newspaper articles (e.g., via LexisNexis searches, Cronin & Shaw, 2002). Since media coverage tends to coincide with publication launches, web reports gave the potential to track impact over the longer term, to get some idea about how the reports were received and to find out about how the ideas were being used.

The first NESTA WIRe was based upon twenty documents published in 2006-7. Searches were constructed for all of these and submitted to Google, Yahoo! (now owned by Microsoft and merged with Bing) and Live Search (now Bing) via

Webometric Analyst. The results were combined and summarised, also using Webometric Analyst. Since the queries were not always able to generate a high proportion of correct matches for document mentions, the additional step was taken to check a random sample of up to 50 URLs from different domains and to count the number of correct matches. The proportion of correct matches thus found was then used as a correction factor to multiply the original figures and hence to give an estimate for the total number of correct matches, as shown in Table 1. This checking process was not straightforward in the case of one report, *Hidden Innovation*, because NESTA held a conference with the same name and it was not always clear whether a mention of the term referred to the document or the conference.

Table 1. Anonymised results from search engine queries for 20 NESTA documents (March, 2008).

Report	URLs matching a query for the report	Web sites (domains) of URLs matching a query for the report	Incorrect matches in a random sample of up to 50 web sites	Estimated number of web sites correctly mentioning the report
1	488	137	20%	**110**
2	44	25	5%	**24**
3	3	1	0%	**1**
4	23	5	0%	**5**
5	16	6	20%	**5**
6	65	33	3%	**32**
7	99	42	17%	**35**
8	8	3	0%	**3**
9	75	25	4%	**24**
10	2502	285	46%	**154**
11	35	7	17%	**6**
12	48	16	6%	**15**
13	353	75	32%	**51**
14	77	16	20%	**13**
15	72	18	22%	**14**
16	26	8	20%	**6**
17	123	33	33%	**22**
18	306	116	86%	**17**
19	503	107	57%	**46**
20	554	67	86%	**10**

Table 1 shows a wide divergence in the apparent online impact of the NESTA documents. In addition to the core web site count statistic (by domain name) and the TLD count statistic, the number of URLs, websites (by domain level ending) and second or top-level domains (STLDs) are also reported since these give some extra information. A private mini-web site was also constructed to give NESTA full access to complete lists of URLs, sites, domains, STLDs and TLDs in case they wanted to check the results or to visit some of the pages.

A content analysis was conducted by a single researcher (the last author). This showed the importance of blogs and government sources. Few pages were critical of NESTA and so the impact is mainly positive. The classification was not cross-checked by a second researcher due to cost constraints; it was a relatively expensive

part of the report because of the human time needed. The following results were obtained.

- Government (37%) Government departments and government-funded organisations.
- Press or blogs (24%) Online newspapers and online versions of offline newspapers, unless the newspaper is specific to a company or affiliated to an academic organisation (e.g., regional research forum). Includes all blogs, whether written by journalists, professionals or the general public.
- Academic source (18%) University or other similar academic institution, including research-only government and non-profit research institutes.
- Non-profit (11%) Any other organisation.
- Industry (8%) Commercial organisations.

Figure 1 breaks down the sources of mentions for each document from the classification exercise. The spread of types of sources of link is quite even across most of the different reports, with press or blog attention appearing in all, and most attracting some kind of university interest.

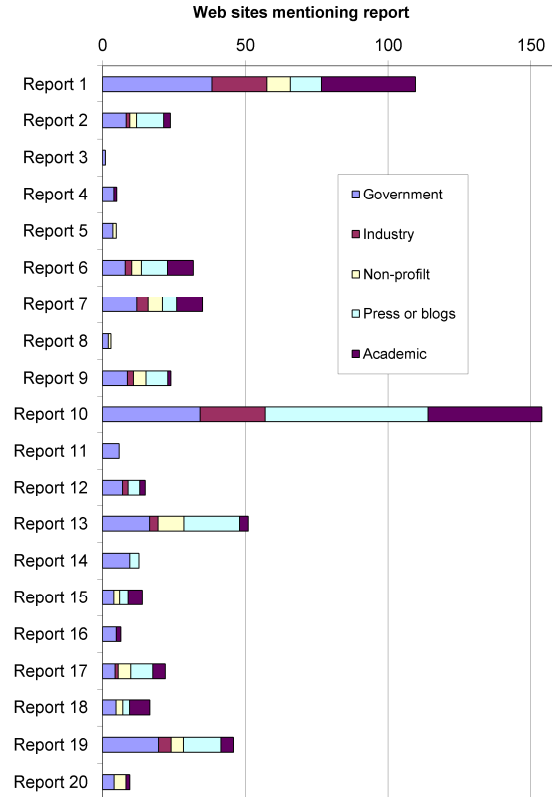

Figure 1. Anonymised sources of mentions of NESTA documents (predicted from the 324 classified), documents listed in chronological order of publication.

The final main aspect of the report, dubbed the NESTA Citation Index, was a list of Word and PDF documents mentioning any of the 20 NESTA documents. The rationale behind this decision was that NESTA wanted evidence of the intellectual impact of their documents but typical web pages carried simple mentions rather than detailed analyses and the Web of Science gave too few citations to be useful, perhaps because of the normal time delay for citation counts. An investigation of web pages mentioning a sample of NESTA documents found that the most detailed discussions tended to be found in PDF or Microsoft Word documents. This included white papers from various organisations, MA and PhD theses, and technical reports. A restriction

to searching for just Word and PDF documents therefore probably captured most of the more substantial discussions of NESTA reports and reduced the total number of matches so that they could all be manually checked for relevance. In retrospect, the searches should also have included other document formats, such as ODF, however.

The PDF and Word counts are summarised in Figure 2, together with a classification of document type. This classification was necessary due to the presence of online PDF regional newsletters that either listed a report or briefly discussed it but did not analyse it to the same extent that most other documents did. The main NESTA citation index, a complete listing of all different PDF and Word citing documents, was placed online in a private web page.

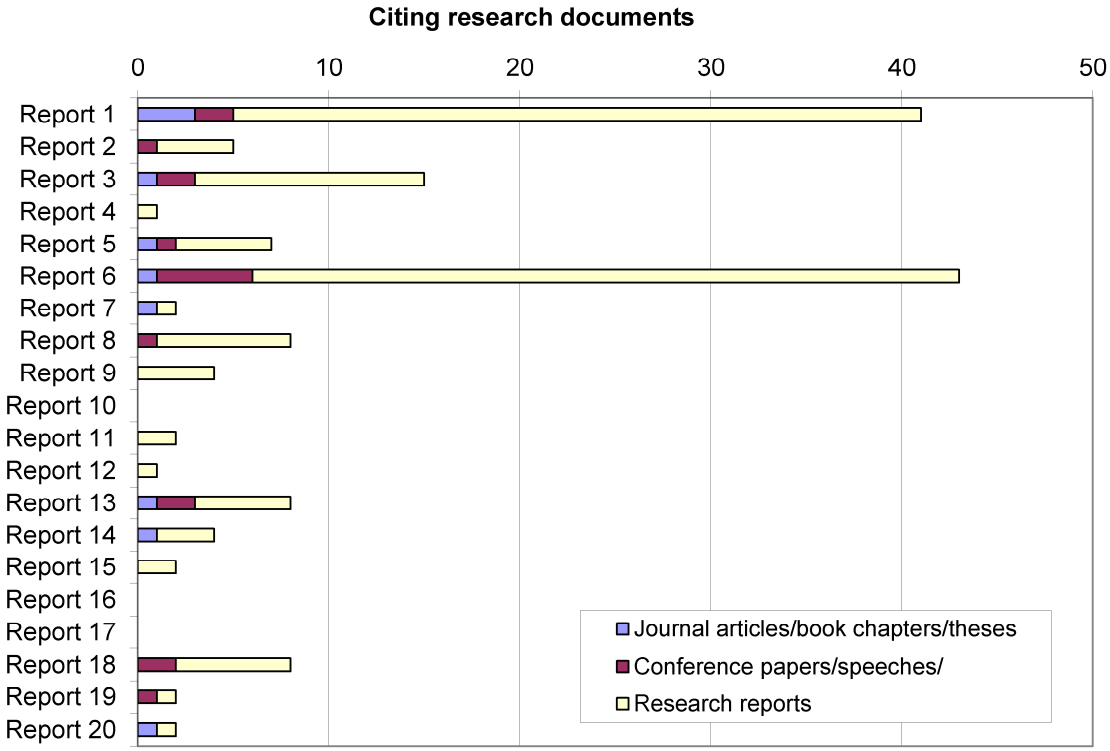

Figure 2. Anonymised research documents, white papers and presentations citing the NESTA reports – excluding all NESTA-authored publications and presentations.

The NESTA citation index was useful to demonstrate substantial impact for the NESTA documents. In contrast to the data reported in Figure 2, only three citations were found to NESTA documents via searches in Google Scholar and the Thomson-Reuters Web of Science. This demonstrates the value of the WIRe for this type of document because NESTA did not want to wait several years to get more reasonable academic citation counts – due to publication delays, several years would be needed – and because during the study it became clear that the natural home for NESTA citations was in the grey literature rather than in academic publications.

Conclusion

The Web Impact Report is an attempt to apply webometric techniques to measure the impact of a type of document that previously seems to have been evaluated using only press mentions. This takes advantage of both the wide range of types of material on the web and the increasing tendency to post PDF versions of research reports, white papers and similar documents online. The absence of an online equivalent of the Science Citation Index necessitated the use of heuristics to identify web pages citing

or mentioning the documents analysed and considerable human effort to check the results and to classify the matches. Overall, the results were able to find new information in terms of the discovery of a range of contexts in which NESTA documents were mentioned. This information provides useful feedback to NESTA authors and managers. It is not possible to be sure whether the impact figures are reasonable or reliable impact estimators, however, because there is no established way to assess the impact of these kinds of document and so it remains a management decision as to whether the results are regarded as being important or not. A corollary of this is that it is difficult to provide convincing evidence of the value of a WIRe, which is a limitation of the current paper. In particular, this research has not formally evaluated WIRes in any specific context with any qualitative or quantitative measure, however, and so their validity in any context is unknown.

In conclusion, the WIRe is a new type of evaluation that scientometricians can conduct as a service to large organisations that publish reports and documents online and who seek evidence about the impact of these publications.

Acknowledgement

This research is part of the FP7 EU-funded project ACUMEN on assessing Web indicators in research evaluation.

References

Baumgartner, F. (2007). Punctuated equilibrium theory and environmental policy. In R. Repetto (Ed.), *Punctuated equilibrium models and environmental policy* (pp. 97-116). New Haven: Yale University Press.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. *Journal of Information Science, 30*(4), 347-368.

Bennett, R. (2007). Sources and use of marketing information by marketing managers. *Journal of Documentation*, 63(5), 702-726.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, 64(2), 217-233

Butler, L. (2003). Explaining Australia's increased share of ISI publications - The effects of a funding formula based on publication counts. *Research Policy, 32*(1), 143-155.

Cronin, B., & Shaw, D. (2002). Banking (on) different forms of symbolic capital. *Journal of the American Society for the Information Science, 53*(13), 1267-1270.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science, 49*(14), 1319-1328.

Dutton, W. H., & Elsper, E. J. (2007). *The Internet in Britain 2007*. Oxford: Oxford Internet Institute.

Gentil-Beccot, A. Mele, S, Brooks, T. (2010). Citing and reading behaviours in high-energy physics, *Scientometrics*, 84(2), 345-355.

Geisler, E. (2000). *The metrics of science and technology*. Westport, CT: Quorum Books.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation, 54*(2), 236-243.

Jeffery, K.G. (2000). An architecture for grey literature in a R&D context, *International Journal on Grey Literature*, 1(2), 64-72.
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 and Technology, 58*(7), 1055 -1065.
Lefebvre, R. C., & Flora, J. A. (1988). Social marketing and public health intervention. *Health Education Quarterly, 15*(3), 299-315.
Moed, H., F. (2005). *Citation analysis in research evaluation*. New York: Springer.
Neuendorf, K. (2002). *The content analysis guidebook.* London: Sage.
Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology, 59*(1), 38-50.
Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. New York: Morgan & Claypool.
Thelwall, M., Vann, K., & Fairclough, R. (2006). Web issue analysis: An Integrated Water Resource Management case study. *Journal of the American Society for Information Science & Technology, 57*(10), 1303-1314.
Thelwall, M., & Wilkinson, D. (2008). A generic lexical URL segmentation framework for counting links, colinks or URLs. *Library and Information Science Research, 30*(2), 94-101.
Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, 35, 4, 469-480.
Vaughan, L., & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology, 54*(14), 1313-1322.
Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29*(1), 49-56.