Bibliographic and Web Citations: What Is the Difference?

This paper has been accepted for publication in the *Journal of the American Society for Information Science and Technology*

Liwen Vaughan, Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, CANADA, N6A 5B7

E-mail: lvaughan@uwo.ca

Phone: 519.661.2111 x88499

Fax: 519.661.3506

Debora Shaw, School of Library and Information Science, Indiana University, 1320 E. 10th Street, Main Library 011, Bloomington, IN 47405-3907

E-mail: shawd@indiana.edu

Phone: 812.855.3261

Fax: 812.855.6166

Web citations have been proposed as comparable to, even replacements for, bibliographic citations, notably in assessing the academic impact of work in promotion and tenure decisions. We compared bibliographic and Web citations to articles in 46 journals in library and information science. For most journals (57%), Web citations correlated significantly with both bibliographic citations listed in the *Social Sciences Citation Index* and the ISI's Journal Impact Factor. Many of the Web citations represented intellectual impact, coming from other papers posted on the Web (30%) or from class readings lists (12%). Web citation counts were typically higher than bibliographic citation counts for the same article. Journals with more Web citations tended to have Web sites that provided tables of contents on the Web, while less cited journals did not have such publicity. The number of Web citations to journal articles increased from 1992 to 1997.

Introduction

In May 2002 Thomas Wilson induced a flurry of discussion with these words on the JESSE listserv on library and information science education:

> There have been a few mentions of Web citation searching possibly replacing citation indexing in time and I wondered how many people are now, as a matter of course, using counts of Web mentions in their cases for appointment, tenure or promotion.
>
> I looked at a couple of my own papers and counted the SSCI citations and then searched for mentions of the papers on the Web - the results left me wondering whether the reliance on citation indexing as a measure of performance is now past its sell by date.
>
> http://listserv.utk.edu/cgi-bin/wa?A2=ind0205&L=jesse&F=&S=&P=720

On June 19 Eugene Garfield responded, through the American Society for Information Science and Technology SIG Metrics discussion list. Garfield is President Emeritus of the Institute for Scientific Information (ISI), and creator of the SSCI (*Social Sciences Citation Index*) that Wilson mentioned. Garfield’s response reads in part:

> I suppose that some day the people who run google and other search engines will figure out a way to separate true research citations from mere mentions of names, but in the meantime it is really not defensible to compare information retrieval via WOS or STN or Dialog of ePsyche or whatever, to searches using google or other search engines over the Internet. That is why I don't waste my time trying to do so. http://listserv.utk.edu/cgi-bin/wa?A2=ind0206&L=sigmetrics&D=1&O=D&F=&S=&P=2512

Wilson’s initial posting is intriguing in that it situates the question of citation comparability in the context of individual faculty evaluation for appointment, tenure, or promotion. We briefly discuss this issue and related studies, then report an empirical study that compared bibliographic and Web citations to articles published in information and library science journals.

The Debate about Bibliographic vs. Web Citations

Early discussions of the Web in information science venues drew connections between bibliographic citations and Web hyperlinks. Larson (1996, p. 74) said succinctly, “the notion of citation is fundamental to both the scholarly enterprise and to hypertext networks where it provides the primary mechanism for connection and traversal of the information space (or ‘cyberspace’).” More recently Harnad and Carr (2000) provided the evocative description of bibliographic citation as “The mother of all hyperlinks.” Cronin, Snyder, Rosenbaum, Martinson, and Callahan (1998) derived 11 categories in which individuals were mentioned on the Web, which they termed “invocations.” They credited Gerry McKiernan’s Cited Sites Web page (http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm) with the first use (at least as early as 1996) of the neologism “sitation” to refer to a cited site; the term was also used by Rousseau (1997).

Although some have drawn the analogy between bibliographic citations and Web hyperlinks, others have suggested significant differences between the two. Egghe (2000) stated that Web hyperlinks are very different from citations in scientific papers, and van Rann (2001) considered Web hypertext links superficial. It should be noted that Web hyperlinking is different from Web citation. The former refers to hypertext links seen on Web pages while the latter refers to Web text citations or mentions of published papers. Some of the many studies on Web hyperlinks are reviewed below; very few studies have examined Web citations. The study reported in this paper compares the Web citations to papers published in LIS journals with bibliographic citations to these articles (as reported by ISI citation database).

Long-recognized lists of reasons for bibliographic citations (for example, Garfield, 1965) include various ways of giving credit to authors on whose work the current paper is based, as well as providing access to related work. Early speculation about motivations for Web links includes Rousseau’s (1997) postulation that links are generally made to help the reader locate additional information on a topic. Kim’s (2000) study of links from e-journal articles provides a sense of the authors’ motivations for hyperlinks, ranging from providing background to describing methods used. As Cronin et al. (1998, p. 1326) put it: “While traditional citation analysis can tell us a lot about the formal bases of intellectual influence, it quite naturally, tells us nothing about the many other modalities of influence which comprise the total impact of an individual’s ideas, thinking, and general professional presence.” The assumption behind Wilson’s posting, and in much of the subsequent commentary, is that an analysis of Web citations would help to document influence in this larger sphere.

Sloan (2001, online) recounted his experiences creating a “personal citation index” (including both bibliographic citations and Web citations) to gauge the impact of his work on “practitioners, researchers and LIS educators.” Citations were identified using the *Social Sciences Citation Index*, other commercial bibliographic databases, and various Web search engines. One of his papers, which was cited eight times in the ISI databases, had 76 citations when all sources had been consulted, excluding “sites that did not seem significant, or ... were largely redundant.” Sloan’s discussion mentioned the complexity of defining what is “international” on the Web. He also found many references from course syllabi and readings lists on the Web, and pondered whether such citations would mean more (“this is important for the next generation”) than a citation in a research article. On the other hand, inclusion in a readings list could also mean less than citation from a research article, as in “this work was not important to my research, but OK for students.” Just as Wilson did in his JESSE posting, Sloan contemplated whether this kind of personal citation index could be used to support a case for promotion and tenure; he also suggested it could help “neophyte academics” make the case that their work had had an impact.

Recent studies by Goodrum, McCain, Lawrence, and Giles (2001) and Zhao and Logan (2002) have compared citations from the Web-based citation indexing system Cite Seer/ResearchIndex with those from ISI databases. Both studies dealt with aspects of computer science, which were likely to produce Web-accessible primary publications. And both studies reported that ResearchIndex and ISI databases were complementary in their coverage of the field. Zhao and Logan (2002, pp. 467-469) provided an interesting discussion of the methodological advantages and disadvantages of Web publication citation analysis:

Advantages of ResearchIndex:

Contains more citing papers

Has a wider variety of document types

Contains more information about cited papers

May allow more automated data collection and analysis

Disadvantages of ResearchIndex

Many papers lack explicit dates of publication

Variety of referencing formats are not handled well by the automated indexing tools

Does not cover interdisciplinary aspect of the field as well

Difficult to parse detailed information on cited papers

Citation Measures in Evaluation

Garfield (1979, p. 240) described the consternation that erupted in response to Wade's (1975) article in *Science* describing citation analysis as a tool for science administrators. He noted that the concerns were about evaluation of individual scientists or academic departments, not the big-picture science policy issues. There has been a long-ranging debate on the desirability and validity of citation analysis for evaluations (for example, Anderson, 1991; Cronin, 1984; Cronin & Overfelt, 1994; MacRoberts & MacRoberts, 1989; Seglen, 1992; Taubes, 1993). Garfield's two-part essay on how to use citation analysis for faculty evaluations (Garfield, 1983a, 1983b) provided considerable context and encouraged caution in interpreting citation counts. He specifically noted the lag between publication and citation, with the consequence that assistant professors are considered for tenure "at a time when that citation data would be most useful, [but] it is not yet available" (Garfield, 1983a, p. 360). This concern foreshadowed Sloan's (2001) suggestion that Web citations could be especially valuable for new researchers seeking to demonstrate the impact of their work.

Careful studies showing the correlation of citation indicators with experts' assessments of the importance of scholarly articles (for example, Virgo, 1977) helped to confirm citation analysis as an acceptable and accepted measure for faculty evaluation. Cole (2000) estimated that one third of the tenure cases he observed as Provost at Columbia University included citation analysis in the assessment of the candidate. Many tenure and promotion guidelines posted on the Web include citation analysis as one measure candidates may include in their dossiers; major universities even instruct faculty members in how to search ISI databases for citation counts (for example, http://www.lib.ohio_state.edu/phyweb/citation.htm).

It took approximately a generation (20 years) for bibliographic citation analysis to achieve acceptability as a measure of academic impact. Some of the earliest proposals for extending recognition to Web text citations or Web hyperlinks came from Almind and Ingwersen (1997), Ingwersen (1998), and Cronin (1999), who anticipated the differing values tenure committees would place on citations from research papers or course syllabi. In a recent paper Wouters (2002) considered how Web-based indicators could be used in peer review and research assessment. Cronin (2001, p. 1) speculated on the potential new approaches to bibliometrics to "capture often liminal expressions of peer esteem, influence and approbation." He suggested that as "a more diverse publishing environment emerges, bibliometricians will have a much broader array of objects and artifacts to feed their accounts and analyses" (Cronin, 2001, p. 3).

Some intriguing comparisons between Web-based measurements and other evaluation measures have been conducted. Thomas and Willett (2000) found that Web link counts did not correlate with departmental rankings in the U.K. Research Assessment Exercise (RAE). The rankings have tended to correlate with citation analysis data, and Owen and Willett discussed the limitations of Web citing and searching that may impede obtaining complete counts. Thelwall (2001), however, found that Web Impact Factor measurements did correlate with RAE rankings. When universities, rather than departments, were analyzed researchers have found correlation at the departmental level (Chu, He, & Thelwall, 2002; Li, Thelwall, Musgrove, & Wilkinson, 2002). In addition to Sloan's (2001) account, Cronin and Shaw (2002) have also looked at Web mentions of individuals in comparison with citation measures. They found that citation counts and Web mentions the authors' names correlated quite well for highly cited researchers in information science. Lawrence (2001) found that for computer science journals and proceedings, high citation counts correlated with online availability. Vaughan and Thelwall (2003) found for both law and LIS journals, that the longer a journal Web site had been online, the more likely it was to receive links from other sites. They also found (Vaughan & Thelwall, 2002) that a journal's ISI impact factor correlated with the number of links to the journal's Web site.

Wilson’s innocent question about Web citation searching replacing citation indexing (or analysis) raises questions at the level of bibliometric theory as well as concerns at the level of Web practices and search engine implementation. It is certainly not clear that traditional bibliographic and Web citations are comparable, or even correlated. Thelwall’s (2002) recent work, for example, considers how a “Web Invocation Portfolio” could provide evidence of research impact. Empirical studies can assist investigations regarding possible extensions of citation analysis; these studies will need to consider both contextual and methodological issues as they attempt to replicate and extend bibliometric methods. To this end, we conducted a study of both the bibliographic citations and the Web citations to research articles in library and information science journals.

Methods

1997 Sample

We investigated similarities and differences between bibliographic and Web citations to research articles published in scholarly journals. We restricted our analysis to information and library science, using the journals in the ISI’s 2000 *Journal Citation Report* “Information Science and Library Science” subject category (the 2001 *Journal Citation Report* was not available at the time of the study). We removed non-research publications, such as *Library Journal*, and journals not written in English (to allow for the early predominance of English on the Web). We restricted our focus to full-length research articles, omitting brief communications, conference reports, editorials, book reviews, etc. In order to allow time for citations to be made, we searched articles published in 1997 for our initial analysis, which was conducted in 2002. The 1997 sample ultimately included 1209 research articles, from 46 journals in information and library science.

We searched for citations to each article in the *Social Sciences Citation Index* through the Web of Science, and recorded the number of citing articles (referred to as ISI citation below). We also searched for citations to each article on the Web, using the Google search engine, which covered more Web sites than any other search engine at the time of our study (Notess, 2002). Our Google search strategy was to enter the article title, as listed in the journal's table of contents, in quotation marks (i.e., phrase search in Google). If the article title was not sufficiently distinctive (e.g., "Evaluation of Public Libraries") to exclude possible false drops, we included the subtitle. If these two were still not distinctive, we added the author's (or authors') last name(s) or words from the title of the journal to the query. If Google reported that some results had been omitted, we selected the option to "repeat the search with the omitted results included." For each article we recorded the number of Web citations, after removing obvious false drops.

Because Google, a commercial search engine, served as the main source of data collection, we must mention the issues of coverage and reliability. Although Web search engine coverage is far from complete (Bar-Ilan, in press; Thelwall, 2001b), Google is consistently among the most popular (Sullivan, 2003) and most comprehensive (Bar-Ilan, in press). While early studies reported volatility of search engine performance and thus called into the question of the reliability of data collected from the engines (Rousseau, 1998/99; Snyder & Rosenbaum 1999), recent studies found search results to be fairly stable (Vaughan & Thelwall, 2003; Vaughan & Wu, in press), probably due to the improvement of search engine technology in recent years. A study conducted in the summer of 2002, around the time of data collection for this study, found Google to be the most stable of the several search engines examined (Vaughan, in press).

In order to investigate Web citations in more detail, we classified citing items by country, type of domain, and the source of the citation. Country and domain type were determined from the URL of the citing item. URLs for U.S. sites typically do not have a country designation; however, not all URLs without a country are from the United States. URLs that did not have clear country designation (e.g., .com) were classified as "country unknown." We were able to classify .edu sites as from the U.S., since that domain is used only in URLs from the United States. Domain types were classified into .com, .org, .edu, and "domain unknown." The sources of citing items were classified into the following categories:

journal: the journal, or the journal's publisher/sponsoring society lists it on its site. For example, a *JASIS* article was cited/listed in ASIST (sponsoring organization) or Wiley (publisher) Web site.

author: the author, co-author, or one of their employers lists the article; includes listing in author's CV, or on the department's Web site.

service: a Web bibliographic service lists the article, for example ResearchIndex (citeseer.nj.nec.com), DBLP bibliography, ASLIB current awareness service, IRList Digest

class: listed in a bibliography/reading list for a course (includes continuing education as well as university-based courses)

paper: cited in a paper that is posted on the Web (the vast majority were papers from conference proceedings or online versions of articles published in journals)

conference: cited in a conference announcement, report, or summary/description other: cited in another way (e.g., cited in a Web site on careers)

We test-classified some sample Web pages and developed our classification scheme based on the types of citations we encountered. Because the classification is fairly straightforward and does not require detailed analysis, the actual data collection on classification was done by two research assistants. We provided them with clear written instructions and encouraged them to consult us when there are any questions.

There were a total of 16,371 Web citations to all the 1997 LIS journal articles in the study. To reduce the classification task to manageable proportions we ranked journals by the number of Web citations per article and examined the top and bottom four journals in this ranking. For the top four journals, we took a systematic sample in the following way: every third Web citations to every third articles was selected and classified. We were able to classify each Web citation for each article in the bottom four journals, because these accounted for fewer than 300 Web citations.

1992 Sample

We found significant correlations between bibliographic and Web citations (discussed below), and wondered whether this relationship would hold for earlier years. We selected 1992 as the test year (allowing 10 years for citations to be made), and sampled the 15 journals with the highest Journal Impact Factor in the 2000 *Journal Citation Report*. Since two journals in this list were not published in 1992, we added the next two journals in the list. The resulting list of journals provided 554 research articles, which we again searched in the *Social Sciences Citation Index* through the Web of Science, recording the number of citations. We also searched for citations to each article on the Web, using the Google search engine with the procedures described above. We classified a systematic sample of one fourth of the Web citations to the three journals that had the most Web citations for both 1997 and 1992. Classification by country and domain type was not conducted for citations to 1992 articles because that classification had been found to be inconclusive for citations to 1997 articles (many sites fell into the “unknown” categories).

Findings

Correlation between Bibliographic Citation and Web Citation

Correlation tests were performed between bibliographic citation (ISI citation data) and Web citation (Google data) for each journal in the study and for both the 1992 and 1997 data. The Pearson correlation coefficient test was used if the frequency distributions for both bibliographic and Web citation data were not badly skewed. Otherwise, the Spearman correlation coefficient test was used. The results are summarized in Table 1 (1997) and Table 2 (1992). Statistically significant correlation coefficients are indicated by asterisks (* for significant at the 0.05 level and ** for 0.01 level). Descriptive statistics (mean, median, and standard deviation) are also provided. It is clear that Web citation counts are typically much higher than ISI citation counts. For example, the average number of Web citations to the 1997 *JASIS* (*Journal of the American Society for Information Science*) articles is 34, while ISI citations average only 5. Web citations to 1997 articles are typically higher than to 1992 articles, as confirmed by a paired t-test ($p<0.01$) for the 15 journals that have both 1992 and 1997 data.

TABLE 1. Correlation between Bibliographic Citation and Web Citation, 1997.

Journal Title	**Correlation coefficient**	**Number of articles**	**Mean, median, standard deviation of ISI citation**	**Mean, median, standard deviation of Web citation**
Aslib Proceedings	0.471**	37	1.22, 0, 2.583	5.57, 3, 7.381
Behavioral & Social Sciences Librarian	0	9	.22, 0, .67	6, 3, 8.87
Bulletin of the Medical Library Association	.435**	40	2.83, 2, 2.99	14.45, 7, 27.85
Canadian Journal of Information and Library Science	0.622	9	.89, 1, .782	3.78, 3, 3.801
College & Research Libraries	.667**	36	3.53, 2, 3.676	15.36, 11.5, 13.117
Database	0.286	25	.40, 0, .65	6.72, 5, 11.19
Electronic Library	.507**	25	1.32, 0, 1.63	8.2, 6, 9.22
Government Information Quarterly	.710**	20	2.2, 2, 2.98	6.6, 5, 5.13
Information & Management	.584**	22	1, 0, 1.72	10.05, 9, 6.74
Information Processing & Management	.471**	51	3.86, 2, 4.43	28.96, 23, 23.15
Information Society	0.35	22	2.86, 1.5, 4.09	33.86, 15.5, 46.93
Information Systems Journal	0.419	16	3.25, 3, 2.41	8.19, 6, 9.53
Information Systems Research	.697**	21	5.71, 3, 4.55	19.95, 19,10.41
Information Technology & Libraries	.644**	22	1.18, 0, 2.2	15.95, 10, 29.12
Interlending & Document Supply	-0.033	20	.70, 0, 1.03	4.85, 4.5, 2.94
International Journal of Geographical Information Science	.401*	39	4.05, 3, 4.12	21.85, 17, 14.27
International Journal of Information Management	.650**	31	2.19, 2, 2.21	7.26, 5, 7.47
Journal of Academic Librarianship	.581**	34	1.88, 1, 2.09	13.29, 10, 13.16

Journal of Documentation	.72**	26	5.92, 3, 6.54	16.27, 12.5, 13.28
Journal of Government Information	.432*	34	1.12, 1, 1.55	4.53, 2, 5.37
Journal of Health Communication	0.261	21	1.29, 0, 2.83	4.95, 3, 7.32
Journal of Information Ethics	-0.13	15	.20, 0, .414	11.80, 10, 5.240
Journal of Information Science	.651**	30	2.67, 1, 3.23	13, 9.5, 9.82
Journal of Information Technology	.476*	23	1.48, 1, .99	5.26, 5, 3.18
Journal of Librarianship and Information Science	0.228	17	.82, 0, 1.24	8.29, 4, 16.14
Journal of Scholarly Publishing	.799**	20	.65, 0, 1.565	4.10, 3, 4.025
Journal of the American Medical Informatics Association	.573**	54	11.7, 8.5, 10.63	18.28, 15.5, 14.12
Journal of the American Society for Information Science	.518**	89	5.03, 3, 6.707	34.42, 25, 28.631
Knowledge Organization	0.133	15	1, 0, 1.773	7.4, 6, 6.685
Law Library Journal	0.429	15	1.27, 1, 1.534	7.27, 7, 6.076
Library Acquisitions: Practice & Theory	0.36	24	1.17, 0, 1.9	6.25, 3, 8.38
Library & Information Science Research	0.477	16	3.88, 2.5, 4.8	9.38, 8, 6.63
Library Quarterly	0.184	12	5.08, 5.5, 2.97	15, 12.5, 7.01
Library Resources & Technical Services	0.328	15	1.8, 2, 1.7	7.73, 7, 5.79
Library Trends	.830**	48	2.54, 1, 3.4	16.69, 11, 15.3
Libri	-0.017	25	.80, 1, .82	.352, 3, 4.28
MIS Quarterly	0.233	21	10.76, 8, 8.41	32.33, 26, 24.04
Online	.554**	22	.95, 0, 1.94	4.86, 1, 8.11
Online & CDROM Review	.980**	12	2.83, 1, 3.93	9.33, 6, 9.09
Program	0.452	17	1.35, 1, 2	9.47, 8, 7.21
Reference & User Services Quarterly	.859*	7	1, 1, 1	12.57, 11, 8.73

Restaurator	0.233	15	3.2, 2, 2.833	2.6, 2, 2.293
Scientometrics	0.413	20	4.1, 2.5, 4.39	8.1, 5.5, 8.04
Social Science Computer Review	.811**	23	3.26, 2, 3.32	8.39, 7, 7.10
Social Science Information	.817**	28	1.32, 1, 2.04	5.04, 2, 6.73
Telecommunications Policy	.462**	68	1.81, 1, 1.81	7.1, 4.5, 7.51

TABLE 2. Correlation between Bibliographic Citation and Web Citation, 1992.

Journal Title	**Correlation coefficient**	**Number of articles**	**Mean, median, standard deviation of ISI citation**	**Mean, median, standard deviation of Web citation**
College & Research Libraries	.553**	35	4.77, 3, 4.66	5.34, 4, 4.93
Information & Management	0.225	51	5.33, 5, 4.45	3.78, 3, 4.1
Information Processing & Management	.534**	54	8.17, 5, 9.29	21.52, 14.5, 20.46
Information Systems Research	.735**	16	33.81, 12.5, 62.44	32.75, 15, 52.94
Information Technology & Libraries	.822**	14	4.71, 3, 4.03	11.79, 8.5, 18.76
Journal of Documentation	.809**	14	10.43, 7, 11.02	15.57, 13.5, 11.47
Journal of Information Science	.593**	47	2.55, 2, 2.91	4.32, 2, 6.09
Journal of the American Society for Information Science	.300*	65	8.77, 6, 13.86	23.88, 17, 24.87
Knowledge Organization	0.24	16	1.44, 1.5, 1.03	2.31, 1, 3.18
MIS Quarterly	.869**	25	26.28, 18, 26.08	30.32, 24, 24.99
Online	.603**	39	2.67, 2, 2.4	4.49, 3, 5.34
Restaurator	-0.023	13	1.54, 1, 2.15	5.31, 3, 7.15
Scientometrics	.485**	76	4.29, 3, 4.34	2.47, 1, 3.07

Social Science Computer Review	.481**	29	1.34, 0, 2.76	7.9, 1, 24.19
Telecommunications Policy	0.242	60	1.82, 1, 2.13	6.15, 4, 13.63

Among the correlating coefficients that are statistically significant, the highest is 0.98 (1997 *Online & CD-ROM Review*) and the lowest is 0.3 (1992 *JASIS*) with an average of around 0.6. For the 46 journals in the 1997 sample, 26 (56.5%) showed a significant correlation between bibliographic and Web citations. For the 15 journals in the 1992 sample, 11 (73.3%) have such a correlation. Among the four journals that do not have a significant correlation in 1992 data, two, *Restaurator* and *Knowledge Organization*, do not have a significant correlation in 1997 sample either. Both are published in non-English speaking countries: *Knowledge Organization* by the International Society for Knowledge Organization in Germany, *Restaurator* by Munksgaard in Denmark. This might be a factor in their relatively low Web citations given the predominance of English on the Web (Global Reach, 2002).

It is intriguing to examine journals that had no significant correlations in the 1997 sample. *Libri*, for example, is published in Germany, and the factors affecting *Knowledge Organization* and *Restaurator* may apply here as well. The *Canadian Journal of Information and Library Science*, *Behavioral & Social Science Librarian*, and *Library Quarterly* all have very small sample sizes (9, 9, and 12 respectively), which makes it very difficult to show a significant correlation. *Scientometrics* was found to be an outlier in a recent journal Web site study (Vaughan & Thelwall, 2002), and retained this image in our assessment as well. Four journals had ISI citation counts too low (median of zero, which means over 50% of the articles had no ISI citation at all) to correlate properly with Web citations: *Interlending & Document Supply*, *Database*, *Journal of*

Librarianship and Information Science, and *Journal of Information Ethics*. The lack of correlation for these four journals explains why the 1992 sample had a higher percent of journals with significant correlations between ISI and Web standings: the 15 journals in the 1992 sample were selected based on their higher Journal Impact Factor, which means higher ISI citations on average. We conclude that bibliographic and Web citations tend to correlate, although there are exceptions to this pattern as noted above.

Correlation between Journal Impact Factors and Average Web Citations

The above correlation was calculated for each journal using individual papers as the unit of data collection and data analysis. We next used individual journals as the unit of data analysis, investigating the relationship between the Journal Impact Factor (JIF) and average number of Web citations for the journal (total number of Web citations a journal received divided by the number of papers in that journal). We did this calculation for each of the 46 journals in the 1997 sample and correlated them with 1998 JIF and the 1999 JIF (since the JIF is based on citations to articles published in the two previous years, citations to 1997 articles will be factored into 1998 and 1999 JIFs). The Pearson correlation coefficients are 0.59 and 0.43 respectively for 1998 JIF and 1999 JIF. Although both are statistically significant at the 0.01 level, they are not very high.

Sources of Web Citations

Web citations were classified into the categories of journal, author, service, class, paper, conference, and other, as described in detail in the “Methods” section, above. Eight journals in the 1997 sample were studied: the four with the most Web citations (*Journal of the American Society for Information Science*, *Information Society*, *MIS Quarterly*, and *Information Processing &*

Management), and the four with the fewest Web citations (*Journal of Scholarly Publishing*, *Canadian Journals of Information and Library Science*, *Libri*, and *Restaurator*). The classification results are shown in Table 3.

TABLE 3. Sources of Web Citations to 1997 Articles.

Numbers in the bracket are percentages

Journal title	Journal	Author	Service	Class	Paper	Conference	Other	Total
Information Processing & Management	1 (0.7%)	10 (7.2%)	70 (50.4%)	10 (7.2%)	33 (23.7%)	1 (0.7%)	14 (10.1%)	139 (100%)
Information Society	6 (10.3%)	3 (5.2%)	14 (24.1%)	5 (8.6%)	15 (25.9)	5 (8.6%)	10 (17.2%)	58 (100%)
MIS Quarterly	4 (4.8%)	1 (1.2%)	0 (0%)	15 (18.1%)	51 (61.4%)	0 (0%)	12 (14.5%)	83 (100%)
Journal of the American Society for Information Science	10 (3.0%)	32 (9.6%)	106 (31.6%)	42 (12.5%)	78 (23.3%)	6 (1.8%)	61 (18.2%)	335 (100%)
Libri	0 (0%)	15 (17%)	27 (30.7%)	9 (10.2%)	24 (27.3%)	4 (4.5%)	9 (10.2%)	88 (100%)
Canadian Journal of Information and Library Science	0 (0%)	7 (21.2%)	5 (15.2%)	6 (18.2%)	12 (36.4%)	2 (6.1%)	1 (3%)	33 (100%)
Journal of Scholarly Publishing	0 (0%)	7 (9%)	31 (39.7%)	15 (19.2%)	23 (29.5%)	0 (0%)	2 (2.6%)	78 (100%)
Restaurator	0 (0%)	0 (0%)	10 (25%)	0 (0%)	24 (60%)	1 (2.5%)	5 (12.5%)	40 (100%)

Total	**21** (2.5%)	**75** (8.8%)	**263** (30.8%)	**102** (11.9%)	**260** (30.4%)	**19** (2.2%)	**114** (13.3%)	**854** (100%)

The two largest categories of Web citations were bibliographic service lists ("service" category, with 263 citations) and citations from a paper posted on the Web ("paper" category, 260 citations). We suggest that the service category count was inflated by DBLP (a computer science bibliography). DBLP has an entry under each author's name, so that a multiple authored paper would have multiple Web citations from DBLP. In addition, DBLP also has mirror sites, with the result that each DBLP citation would be counted multiple times by Google. Discounting the service citations for these reasons renders citations from Web-posted papers the most frequent type of Web citation. We suggest these are most directly comparable to traditional bibliographic citations.

Citations in the form of class reading lists ("class" category) represented approximately 12% of the cases. Authors' own Web listings of their papers ("author" category, 9%) outnumbered the listing by the journal ("journal" category, 3%). If we consider citations in the class and paper categories to be the best indicators of the impact of the cited article, then 42% (362/854) of Web citations represented some kind of impact. This percent rises to 49% (362/740) if we exclude the citations from "unknown" sources from the calculation.

Do the four most cited journals differ from the four least cited journals in the distribution of these citation categories? Data in Table 3 were aggregated into these two groups as shown in Table 4. A chi-square test showed a significant ($p<0.01$) difference between the two groups in terms of the cited category. There are two numbers in each cell of Table 4. The first is the observed count (data collected) and the second one is the expected count (from the Chi-square test). Comparing these two numbers, we see that the main difference is in the journal category. The least cited journals had no Web listing of their papers at all. However, their authors and Web bibliographic services did not under-represent the papers. Web citations in the journal category were typically in the form of a table of contents maintained on the Web by the journal. Some journals even had double lists of their articles. For example, *JASIS* was represented by the table of contents on the American Society for Information Science and Technology Web site and through the same information archived in the ASIS-L listserv, where it had been posted. This demonstrates the variability of Web citation data.

TABLE 4. Comparison between the Most Cited and the Least Cited Journals.

Note: The first number in each cell is the observed count and the second number the expected count

	Journal	**Author**	**Service**	**Class**	**Paper**	**Conference**	**Other**	**Total**
Most cited journals	21 15	46 54	190 189	72 73	177 187	12 14	97 82	615
Least cited journals	0 6	29 21	73 74	30 29	83 73	7 5	17 32	239
Total	21	75	263	102	260	19	114	854

Sources of citations to 1992 articles in the three most cited journals (*JASIS*, *Information Processing & Management*, and *MIS Quarterly*) were classified using the categories discussed above, with the results presented in Table 5. As with the 1997 data, most citations came from the "service" and "paper" categories. To determine if the distribution of citation categories of 1992 data differed from that of 1997 data, we summarized 1992 and 1997 data for these three journals and contrasted them in Table 6. A Chi-square test was performed on Table 6 data and the result shows a significant ($p<0.01$) difference between 1992 and 1997 data in the citation categories. Comparing the observed count (the first number in each cell of Table 6) with the expected count (the second number in each cell), we found that citations to 1992 articles fell proportionally more into the "paper" and less into the "class" category. It would appear that class reading lists are less likely to include 10-year-old articles, while these articles are still cited by other papers.

TABLE 5. Sources of Web Citation to 1992 Articles.

	Journal	**Author**	**Service**	**Class**	**Paper**	**Conference**	**Other**	**Total**
Information Processing & Management	0	5	23	5	33	0	4	70
Journal of the American Society for Information Science	1	7	44	3	19	0	13	87
MIS Quarterly	1	0	5	2	32	1	0	41
Total	**2**	**12**	**72**	**10**	**84**	**1**	**17**	**198**

TABLE 6. Comparison of Web Citation Categories, 1997 vs. 1992.

Year	**Journal**	**Author**	**Service**	**Class**	**Paper**	**Conference**	**Other**	**Total**
1992 (observed count)	2	12	72	10	84	1	17	198

(expected count)	4	14	65	20	65	2	27	
1997 (observed count)	15	43	176	67	162	7	87	
(expected count)	13	41	183	57	181	6	77	557
Total	**17**	**55**	**248**	**77**	**246**	**8**	**104**	**755**

Web Citation Classification by Country

Results of Web citation classification by country for the 1997 data are shown in Table 7. Of the 849 Web citations classified by country, only 1 was from Africa, 12 from New Zealand and Australia, and 27 from Asia. If Web citations can be seen as a measure of impact, then these journal articles have little impact through the Web outside North America and Europe, reflecting the general pattern of relatively lower Web penetration and use beyond these areas (Global Reach, 2002). Nearly a quarter of the citing URLs could not be assigned to a country (typically those with .com or .int domains). From those that could be classified, it appears that European Web citations outnumbered those from the U.S. We note, however, that U.S. sites accounted for only 28% of the citations, much lower than the 47% ascribed to the U.S. by the OCLC Web Characterization Project (2002). Due to the uncertainty in the data (large number of “unknowns”), no inferential statistical tests were performed to compare journals in citations by country. After experiencing this uncertainty in classifying citations to 1997 articles, we decided not to proceed with country classification for citations to 1992 articles.

TABLE 7. Web Citations by Country.

Journal	**U.S.**	**Canada**	**Europe**	**Asia**	**Africa**	**Australia/ NZ**	**Country unknown**	**Total**

Information Processing & Management	15	3	72	2	0	1	48	141
Information Society	16	15	3	1	0	1	20	56
MIS Quarterly	31	3	14	1	0	6	27	82
Journal of the American Society for Information Science	91	8	136	20	0	4	74	333
Libri	30	15	27	3	1	0	12	88
Canadian Journal of Information and Library Science	7	17	4	0	0	0	5	33
Journal of Scholarly Publishing	39	5	18	0	0	0	15	77
Restaurator	8	0	22	0	0	0	9	39
Total	**237**	**66**	**296**	**27**	**1**	**12**	**210**	**849**

Web Citations by Types of Domains

Table 8 presents the types of domains from which journal articles were cited. Many URLs did not have clear designation of the type of organization hosting the site, which resulted in 49% (416 of 854) of the sites falling into the "domain unknown" category. Due to this uncertainty, no inferential statistical analysis was performed to compare journals by types of domains citing them. However, the data in Table 8 still provide some useful information. Of the 438 citations with known domains, 57% are from .edu sites, suggesting that most Web citations come from college and university sources. Journals that have lower Web citation counts (the last four in Table 8) seem to have relatively fewer citations from .com and more from .edu sites. It should be noted that the data in Table 8 are based on citations to 1997 articles. Classification by domain type was not done for citations to 1992 articles because the large number of "unknown" sites renders the analysis inconclusive.

TABLE 8. Web Citation by Types of Domains.

Journal	.com/.co	.org	.edu	Domain unknown	Total
Information Processing & Management	30	15	15	81	141
Information Society	8	11	16	22	57
MIS Quarterly	12	12	35	26	85
Journal of the American Society for Information Science	25	39	98	172	334
Libri	3	9	30	45	87
Canadian Journal of Information and Library Science	3	3	7	20	33
Journal of Scholarly Publishing	8	6	39	25	78
Restaurator	0	6	8	25	39
Total	**89**	**101**	**248**	**416**	**854**

Discussion and Conclusions

For most journals (57%) in the study, there was a significant correlation between bibliographic citations and Web citations to their articles. The lack of correlation for some journals may be attributable to such factors as very low numbers of bibliographic citations or publication from outside the U.S./Canada/U.K. This leads to the conclusion that normally bibliographic and Web citation counts are correlated; however, it is not clear that bibliographic and Web citations are measuring “the same thing.” The journal impact factor (JIF) correlated with the average number of Web citations to the journal and the number of Web citations was typically higher than the number of bibliographic citations for the same article. Among the 46 journals in the study, 14 had a median of zero ISI citation counts (median of zero means more than 50% of the articles in the journal have no ISI citations, and thus one cannot distinguish these articles by citation counts). In contrast, no journal had a median of zero for Web citation counts. This can be viewed as an advantage of Web citation analysis, in that it can make finer distinctions among articles. The number of Web citations to 1997 articles was typically higher than to 1992 articles, suggesting that more recent articles are more likely to be cited on the Web.

When the Web citations were classified according to the source of citation, the two largest groups (about 30% for each) were citations in papers posted on the Web and in listings of Web bibliographic services such as ResearchIndex. If we consider citations from other papers and class reading lists to be indicators of the intellectual impact of the cited article, then at least 42% (possibly as high as 49%) of Web citations represented this kind of influence. Sources of Web citations to 1992 articles were similar to those of 1997 articles, although citations to 1992 articles tended to have fewer class reading list citations and more citations from Web-posted papers. A major difference between journals with higher Web citations and those with lower Web citations was that the former tended to have Web listings of their articles (often in the form of tables of contents) while the latter had no Web promotion of their articles. However, this was not the main cause of the lower Web citations to the latter because Web citations by the journal itself accounted for less than 3% of the total Web citations (21 out of 854, as shown in Table 4).

We attempted to classify Web citations by country and domain type, but the large number of “unknowns” rendered these analyses inconclusive. Among the citations that have known country designations, the vast majority were from North America and Europe (about 47% each) while only 6% were from the rest of the world. The most common type of citing domain was .edu, followed by .org and .com.

Returning to the question posed at the beginning of the paper, is Web citation ready to replace bibliographic citation? From this study, it is fair to conclude that Web citations parallel or confirm analyses of bibliographic citations reported by ISI. However, Web citation practices are far from uniform, and it is difficult to distinguish what Garfield (2002, online) called “research citations” from “mere mentions of names.” Thus Web citation analysis is not, or not yet, a replacement for the study of bibliographic citations, especially in assessing academic impact in promotion and tenure deliberations.

Web citations do have advantages over bibliographic citations and contain complementary information. The “faster turnaround” with Web citations may be particularly helpful with the time lag in bibliographic citations, which has been a recognized concern for at least 20 years (Garfield, 1983a). Moreover, Web citation searches provide more access points (such as words from the article title) to improve discrimination and enhance recall. On the other hand, the potential impermanence of some Web citations may introduce problems for bibliometric or Webometric research and for use of citation measures in evaluation—what confidence can promotion committees, or even the readers of this paper have that the Web citation counts reported were accurate? The development of online archives may alleviate some of these problems. Web citation analysis is certainly promising and has considerable potential due to the continued development of the Web. Further research in this area is probably inevitable, and is also needed to contribute to our knowledge in this new domain.

Acknowledgments

We wish to thank Mary Beth Cox and Elena Saunders for their work in data collection. Blaise Cronin, Charles H. Davis, and an anonymous referee provided helpful advice on drafts of this paper.

References

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

Anderson, A. (1991, May 3). No citation analyses please, we're British. *Science, 252* (5006), 639. Retrieved August 24, 2002 from http://links.jstor.org/sici?sici=0036_8075%2819910503%293%3A252%3A5006%3C639%3ANCAPWB%3E2.0.CO%3B2_K

Bar-Ilan, J. (in press). The use of Web search engines in information science research. *Annual Review of Information Science and Technology, 38*.

Chu, H., He, S., & Thelwall, M. (2002). Library and information science schools in Canada and the USA: A Webometric perspective. *Journal of Education for Library and Information Science, 43*(2): 110-125.

Cole, J. R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In B. Cronin & H. B. Atkins (Eds.). *The web of Knowledge: A Festschrift in honor of Eugene Garfield* (pp. 281-300). Medford, NJ: Information Today, Inc.

Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on Web-based citation analysis. *Journal of Information Science*, *27*(1), 1-7.

Cronin, B. (1999). The Warholian moment and other proto-indicators of scholarly salience. *Journal of the American Society for Information Science, 50*(10), 953-955.

Cronin, B. & Overfelt, K. (1994). Citation-based auditing of academic performance. *Journal of the American Society for Information Science, 45*(2), 61-72.

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

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

Egghe, L. (2000). New informetric aspects of the Internet: Some reflections–many problems. *Journal of Information Science*, *26*(5), 329-335.

Garfield, E., 1965. Can citation indexing be automated? In: M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.). *Statistical Association Methods for Mechanized Documentation, Symposium*

Proceedings, Washington 1964 (pp. 189-192).NBS Miscellaneous Publication 269. Washington D.C.: National Bureau of Standards.

Garfield, E. (1983a). How to use citation analysis for faculty evaluations, and when is it relevant? Part 1. *Essays of an information scientist* (vol. 6, pp. 354-362). Philadelphia: Institute for Scientific Information. Retrieved August 26, 2002 from http://www.garfield.library.upenn.edu/essays/v6p354y1983.pdf

Garfield, E. (1983b). How to use citation analysis for faculty evaluations, and when is it relevant? Part 2. *Essays of an information scientist* (vol. 6, pp. 363-372). Philadelphia: Institute for Scientific Information. Retrieved August 26, 2002 from http://www.garfield.library.upenn.edu/essays/v6p363y1983.pdf

Garfield, E. (2002, June 19). Web citation. [Posting to ASIS Special Interest Group on Metrics SIGMETRICS@LISTSERV.UTK.EDU]. Retrieved August 24, 2002 from http://listserv.utk.edu/cgi-bin/wa?A2=ind0206&L=sigmetrics&D=1&O=D&F=&S=&P=2512

Global Reach (2002, September 30). Global Internet statistics (by language). Retrieved May 2, 2003, from http://www.glreach.com/globstats

Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing & Management*, *37*(5), 661-675.

Harnad, S. & Carr, L. (2000). Integrating, navigating and analyzing open eprint archives through open citation linking (The OpCit Project). *Current Science*, *79,* 629-638. Retrieved August 24, 2002 from http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.citation.htm

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

Kim, H. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science*, *51*(10), 887-899.

Larson, R. R. (1996).Bibliometrics and the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, 71-83. Retrieved August 24, 2002 from http://sherlock.berkeley.edu/asis96/asis96.html

Lawrence, S. (2001). Online or invisible? *Nature*, *411*(6837), 521.

Li, X., Thelwall, M., Musgrove, P., & Wilkinson, D. (2002, September). The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001. Paper presented at the Seventh International S&T Indicators Conference, Karlsruhe, Germany.

MacRoberts, M. H. & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342-349.

Notess, G. R. (2002). Search Engine Statistics: Relative Size Showdown. Retrieved September 2, 2002 from http://www.searchengineshowdown.com/stats/size.shtml

OCLC Web Characterization Project (2002). Top 50 List. Retrieved September 4, 2002 from http://wcp.oclc.org/

Owen, T. & Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science, 26*(6), 421-428.

Rousseau, R. (1997). Sitations: An exploratory study. *Cybermetrics,1,* 1-9. Retrieved August 24, 2002 from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1998/99). Daily time series of common single word searches in Alta Vista and Northern Light. *Cybermetrics*, 2/3(1), Retrieved September 9, 2002 from www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638.

Sloan, B. (2001). Personal citation index: Exploring the impact of selected papers. Retrieved

August 24, 2002 from http://www.lis.uiuc.edu/~b_sloan/pci2.html

Snyder H., & Rosenbaum, H. (1999). Can search engines be used as tools for Web-link analysis? A critical view. *Journal of Documentation*, *55*(4), 375-384.

Sullivan, D. (2003, April 23). The major search engines and directories. Retrieved May 2, 2003, from http://www.searchenginewatch.com/links/article.php/2156221

Taubes, G. (1993, May 14). Measure for measure in science. *Science, 260*(5110), 884-886. Retrieved August 24, 2002 from http://links.jstor.org/sici?sici=0036_8075%2819930514%293%3A260%3A5110%3C884%3AMFMIS%3E2.0.CO%3B2_D

Thelwall, M. (2001a). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, *52*(13), 1157-1168.

Thelwall, M. (2001b). The responsiveness of search engine indexes. *Cybermetrics*, *5*(1). Retrieved May 2, 2003, from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2002). Research dissemination and invocation on the Web. *Online Information Review*, *26*(6), 413-420.

van Raan, A. F. J. (2001). Bibliometrics and Internet: Some observations and expectations.

Scientometrics, 50(1), 59-63.

Vaughan, L., & Thelwall, M. (2002). Web link counts correlate with ISI impact factors: Evidence from two disciplines. *Proceedings of the 65th Annual Meeting of the American Society for Information Science and Technology*, 436-443.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology,* 54(1), 29-38.

Vaughan, L., & Wu, G. (in press). Link counts to commercial Web sites as a source of company information. *Proceedings of the 9th International Conference of Scientometrics and Informetrics*.

Vaughan, L. (in press). New measurements for search engine evaluation proposed and tested. *Information Processing & Management*.

Virgo, J. A. (1977). A statistical procedure for evaluating the importance of scientific papers. *Library Quarterly, 47*(4), 415-430.

Wade, N. (1975, May 2). Citation analysis: A new tool for science administrators. *Science, 188* (4187), 429-432. Retrieved August 24, 2002 from http://links.jstor.org/sici?sici=0036_8075%2819750502%293%3A188%3A4187%3C429%3ACAANTF%3E2.0.CO%3B2_S

Wouters, P. (2002). Problems in the use of Web indicators for research assessment [abstract]. EASST 2002 Conference, January 18, 2002. Retrieved September 3, 2002 from http://www.york.ac.uk/org/satsu/easst2002/abstracts/Wouters.pdf

Zhao, D., & Logan, E. (2002). Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. *Scientometrics, 54*(3), 449-472.