Challenges of Scholarly Publications on the Web to the Evaluation of Science — A Comparison of Author Visibility on the Web and in Print Journals

Dangzhi Zhao*

School of Library and Information Studies, University of Alberta, Edmonton, AB, T6H 1J5, Canada

Abstract

This article reveals different patterns of scholarly communication in the XML research field on the Web and in print journals in terms of author visibility, and challenges the common practice of exclusively using the ISI's databases to obtain citation counts as scientific performance indicators. Results from this study demonstrate both the importance and the feasibility of the use of multiple citation data sources in citation analysis studies of scholarly communication, and provide evidence for a developing "two tier" scholarly communication system.

Keywords: Scholarly communication; Citation analysis; Science evaluation; Web publishing; Author visibility

1. Introduction

As the accelerated development of information technology, especially the rapid growth of the Web, is changing the circumstances and consequently the structures and processes of scholarly communication, there is renewed interest in the study of scholarly communication to see how it is being transformed, what the similarities or differences between the new formats of communication and the traditional ones might be, and how the new formats facilitate or inhibit the scholarly communication process (Borgman & Furner, 2002; Cronin, 2001; Zhao, 2003).

Citation analysis and other informetric techniques have been applied successfully to the study of this new phenomenon in scholarly communication. As Zhao & Logan (2002) point out, some studies apply, often with modifications, informetric principles and techniques to study the characteristics and link structures of the Web. Examples include studies on search engines that make use of hyperlink structure (Clever, 1999), and so-called "Webometrics" studies (Almind & Ingwersen, 1997; Cronin et al., 1998, Egghe, 2000; Larson, 1996; Rousseau, 1997; Thelwall & Harries, 2004; Turnbull, 2000; Wilkinson et al., 2003). Other studies consider "electronic ingredients" in journal articles – either in reference lists or in abstracts – to gauge the impact of electronic publications on traditional print journal-based scholarly communication (Harter, 1992; Harter & Kim, 1996; ISI, 2004a; Lu, 1999; McCain, 2000; Youngen, 1997). Still others examine scholarly communication patterns demonstrated in research papers published on the Web, and study the differences from, and similarities to, what we have come to expect from print journal-based communication (Goodrum et al., 2001; Lawrence, 2001; The Open Citation Project, 2001; Zhao & Logan, 2002; Zhao, 2003 & 2004).

As part of a larger research project that aims to systematically compare scholarly communication patterns between the Web and the print world, the present study focuses on the comparison of author visibility between the Web and print journals as revealed from citation analysis, and discusses the challenges of scholarly communication being increasingly conducted over the Internet to traditional scholarly communication system in general and to the common practice of science evaluation based on the databases of the Institute for Scientific Information (ISI) in particular. The present study along with other parts of the project (Zhao, 2004; Zhao & Logan, 2002; Zhao & Strotmann, 2004) may contribute to the understanding of the transition of scholarly communication from print to electronic media, to advancing citation analysis theory and methodology, and to information organization and retrieval on the Web.

2. Research questions

Within a citation analysis framework, author visibility can either be measured in terms of how frequently authors have been publishing or in terms of how often their published works have been used (cited) by other scholars. Based on this consideration, the research questions to be explored in the present study are as follows.

- Are there any significant correlations between author rankings by number of publications identified from the Web and those identified from print journals in the field of XML research?

* Email: dzhao@ualberta.ca, Phone: 1-780-4922814, Fax: 1-780-4922430

- What is the degree of correlation between author rankings by number of citations identified from the Web and those identified from print journals in the XML research field?
- What has contributed to the differences in author visibility between the Web and the print world?

A study we reported on earlier (Zhao & Logan, 2002) compared author rankings between the Web and print journals in the XML research field based on a visual inspection of a small set of highly visible authors. We found a considerable difference in publication patterns between these two views, but at the same time similar rankings of the top ten authors as ranked by their number of citations. We also noted the importance of examining the characteristics of author groups with different publication and citation patterns. The present study builds on this earlier study and examines the degree of correlation through the use of statistical approaches and more controlled data, and explores possible contributing factors to differences in author visibility by examining specific characteristics of authors.

3. Methodology

3.1. Data collection

The ISI's *Science Citation Index* (*SCI*) and The NEC Research Institute's *CiteSeer*[†] were used in the present study to collect information on research papers published in print journals and on the Web, respectively. To date, the ISI databases including *SCI* have been used as the single data source for most of the citation analysis studies reported in the literature. *SCI* was originally designed for print journals, and the majority of journals covered by *SCI* nowadays are still print-based (in print format or having a print version), although it now also selectively indexes e-journals (ISI 2004b). *CiteSeer* is a *SCI*-like tool freely available on the Web. It automatically indexes research papers of any type (journal articles, technical reports, conference papers, etc.) that is in the broadly defined computer science field and publicly available on the Web. Our previous studies have provided evidence that citation analysis studies using *CiteSeer* as a data source are as valid as those using *SCI* (Zhao & Logan, 2002; Zhao, 2003). More information about *CiteSeer* can be found in Bar-Ilan (2001), Goodrum et al (2001), Lawrence et al (1999), and Zhao & Strotmann (2004).

Although XML technology has applications in a wide range of areas, the core of the XML research field belongs to computer science. Since *CiteSeer* covered only broadly defined computer science research while *SCI* covered all sciences, three sets of source data were collected in order to control for data scope in our comparisons. They were all documents (along with their references) indexed under the term "XML" or "eXtensible Markup Language" from (1) *CiteSeer*, (2) the entire *SCI* database, and (3) journals indexed and classified in *SCI* as representing computer science research.

Thus, the search terms "XML" and "extensible Markup Language" were used to identify papers (citing papers) on XML. The actual searches were conducted on December 18, 2001. Papers that met the searching criteria were retrieved from the databases (*SCI* or *CiteSeer*) and downloaded into a local machine. Since the abundance of duplicates had previously been identified as one of the major differences between traditional citation databases and those indexing the Web automatically, paper entries retrieved from *CiteSeer* were examined first by a Java program and then manually to remove possible duplicates. Programs were then developed in Java to convert the data formats of the retrieved paper entries to a data structure that was convenient for subsequent data analyses such as counting citations and co-citations. The data structures used and the algorithms underlying these programs can be found in Zhao (2003).

To better control for data scoping issues, the present study limited the search for citing papers in *CiteSeer* to "header" fields rather than searching in the full text of the documents as we had done in an earlier study (Zhao & Logan, 2002). The reason for this change in our data collection method was that *SCI* only goes as far as abstracts when indexing citing papers, and "header" fields in the *CiteSeer* database appeared to be similar in scope. We hoped that this way of collecting data would result in more comparable data from the two data sources.

3.2. Data analysis

An author ranking by the number of publications using fractional counts and another ranking by the number of citations using straight counts were produced based on each of the three data sets — the data set from *CiteSeer*, that from the entire *SCI* database and that from a subset of *SCI* addressing computer science research, resulting in three

[†] It was also known as *ResearchIndex*, and is now a joint effort of NEC and the School of Information Science and Technology at Pennsylvania State University. URL: http://citeseer.ist.psu.edu/.

author rankings by each of the two ranking criteria. Within each of the two groups of rankings, each author ranking was compared with every other ranking in that group, and for each such comparison, Pearson's r was calculated for the common authors in order to examine the degree of correlation between the two rankings being compared.

Authors' characteristics such as age, nationality, research topic, publishing history, nature of affiliation, relationship with the World Wide Web Consortium (W3C), and collaboration preferences, were then examined to identify possible contributing factors to differences in authors' visibility. For this purpose, authors were classified as belonging to one of three groups: (1) authors who are highly visible both in *SCI* and in *CiteSeer*, (2) authors who are highly visible in *SCI* but not in *CiteSeer*, and (3) authors who are highly visible in *CiteSeer* but not in *SCI*.

Among the various methods of counting citations and publications, fractional counts are the most preferred and recommended by many studies (Egghe & Rousseau, 1990; Lindsey, 1980; van Hooydonk, 1997). The concept of fractional counts is therefore chosen here for counting authors' publications. Specifically, the number of publications of each of the N authors of a publication increases by 1/N when this publication is counted, where N is normally the actual number of authors of each publication, although the present study took a simplified approach in that it only took into account the first five authors rather than all authors. It was hoped that this approach would approximate strict fractional counts sufficiently as publications with more than five authors were not expected to occur too frequently based on the statistics we had (Table 1), and even if its approximation were insufficient it would still help us to see beyond the straight counts, where only the first author's citation count increases by 1 when a paper is cited.

On the other hand, only straight counts were used here for counting authors' citations because they are the only citation counting method supported by *SCI*. While author rankings by fractional citation counts and by complete citation counts were obtained as well from *CiteSeer*, results from comparisons between different citation counting methods will be reported in a separate paper.

Table 1: Distribution of citing papers by number of authors

# papers / # authors	ResearchIndex		SCI	
	#	%	#	%
0	4	1	0	0
1	83	27	75	20
2	77	25	99	26
3	78	25	86	23
4	36	12	54	14
5 or more	34	11	60	16

4. Findings

A search on "XML" or "eXtensible Markup Language" resulted in 312 papers using *CiteSeer* (after removing duplicates) and in 374 papers with reference lists using *SCI*, 268 of which from computer science journals. Among these papers, only 26 were common to both *CiteSeer* and *SCI*. The papers from *CiteSeer* made 4,578 citations, and those from *SCI* made 6,782 citations. Among the cited papers from *CiteSeer*, 22% (987) were proceedings, 22% (991) were Web publications, 2.3% (105) were technical reports, and 2.5% (115) were from books. Among the cited papers from *SCI*, 20.6% (1399) were proceedings. We were not able to calculate the percentage of other types of documents in *SCI* due to the limited amount of information *SCI* provides about cited papers. It can be seen that papers in both data sources in the present study cite roughly the same percentage of proceedings.

The percentage of citing papers shared by the two data sources was very low (less than 10%), which means that in the XML research field, papers published in journals were not largely made available on the Web and papers published on the Web were not well represented in journals indexed by *SCI*. Since papers publicly available on the Web have been found to tend to have more impact on research (Lawrence, 2001), there appears to be a need to promote the public availability of research papers on the Web in order to improve the efficiency and effectiveness of scholarly communication in the XML research field.

4.1. Author visibility indicated by number of publications

An author ranking by fractional counts of publications was produced based on each of the three data sets — the data set from *CiteSeer*, that from the entire *SCI* database and that from a subset of *SCI* addressing computer science research. Parts of these rankings — authors whose fractional publication counts are higher than one — are presented in Table 2. More complete lists can be found in Zhao (2003).

Table 2: Authors ranked by number of publications
(fractional counts greater than 1)

ResearchIndex		SCI		Computer science journals in SCI	
Name	**#p**	**Name**	**#p**	**Name**	**#p**
Wenfei Fan	4.07	H. S. Rzepa	4.07	A. Hunter	2
D. Fensel	2.9	P. Murray-rust	3.57	M. Rezayat	2
Dan Suciu	2.82	D. Suciu	3.2	P. T. Wood	2
Serge Abiteboul	2.68	G. V. Gkoutos	2.15	S. J. Derose	2
M. Murata	2.67	R. H. Dolin	2.05	W. Weitz	2
J. Simeon	2.45	A. Hunter	2	J. Dudeck	1.95
Angela Bonifati	2.33	A. Kristensen	2	H. S. Rzepa	1.92
Harold Boley	2.33	J. Hunter	2	S. Paraboschi	1.62
Dongwon Lee	2.25	M. Rezayat	2	P. Murray-rust	1.58
Mark Huckvale	2	P. T. Wood	2	M. Fernandez	1.53
S. Ceri	1.9	S. J. Derose	2	H. Kim	1.5
Amarnath Gupta	1.75	W. Weitz	2	K. Canfield	1.5
Victor Vianu	1.73	J. Dudeck	1.95	N. Sundaresan	1.5
Daniela Florescu	1.7	M. F. Fernandez	1.87	E. Bertino	1.33
M. F. Fernandez	1.65	J. Simeon	1.75	F. A. Fontana	1.33
Wolfgang Emmerich	1.58	S. Ceri	1.7	G. Weikum	1.33
L. Libkin	1.5	E. Bertino	1.67	E. Damiani	1.28
Leonidas Fegaras	1.5	S. Paraboschi	1.62	L. Kerschberg	1.25
Torsten Schlieder	1.5	M. Wright	1.57	L. Rutledge	1.25
W. van der Aalst	1.5	A. Sahuguet	1.5	S. Ceri	1.2
Elena Ferrari	1.45	H. Kim	1.5	J. Simeon	1.17
John Miller	1.33	J. R. Smith	1.5	M. Shields	1.15
A. Finkelstein	1.25	K. Canfield	1.5	R. H. Dolin	1.05
B. Ludascher	1.25	N. Sundaresan	1.5		
F. Tian	1.25	C. M. Chiu	1.33		
I. Schena	1.25	F. A. Fontana	1.33		
M. Mani	1.25	G. Weikum	1.33		
Letizia Tanca	1.2	E. Damiani	1.28		
Marin Dimitrov	1.2	A. Zisman	1.25		
S. Saeyor	1.2	L. Kerschberg	1.25		
D. Kossmann	1.17	L. Rutledge	1.25		
E. Damiani	1.15	A. Y. Halevy	1.25		
S. Paraboschi	1.12	M. Shields	1.15		
Piero Fraternali	1.07				
Philip Wadler	1.03				

For practical reasons, we only aimed to select the top-ranked 100 authors in each ranking for closer examination. Although the goal was to select 100 authors, that number was not used as a strict cut-off but as a guideline so that authors with the same number of publications were treated identically: either all or none being selected. As a result, the number of publications used in selecting the top-ranked authors was different from one ranking to another. Specifically, the criterion of "fractional publication counts greater than 0.5" was used for the ranking from the computer science journals in SCI, resulting in 101 authors, and "fractional publication counts greater than 0.8" and "fractional publication counts of at least 0.9" were used respectively for rankings from the entire SCI database and from *CiteSeer*, resulting in 104 and 100 authors, respectively. These criteria were those that each resulted in a number of authors that was the closest to the goal of 100 top-ranked authors.

4.1.1. Common authors and correlations between author rankings

Among these top authors, 13 authors were common to all three lists, 15 were common to the lists from *CiteSeer* and from the entire *SCI*, 15 were common to the lists from *CiteSeer* and the computer science journals in *SCI*, and 77 were common to the lists from the entire *SCI* and from the computer science journals in *SCI*. Only about 8% of the authors who actively published were highly visible both on the Web and in print journals. This confirms at a larger scale our observation in an earlier study (Zhao & Logan, 2002) that two very different groups of scholars were actively publishing on the Web or in journals.

The multidisciplinary nature of the *SCI* database may have led one to expect that results from the *CiteSeer* database, which only covered computer science research, should be closer to those from computer science journals in *SCI* than to those from the entire *SCI* database. Data from the present study reveals that this is not the case.

The Pearson's r for author rankings of the 113 authors shared by all three full datasets is 0.478 between *CiteSeer* and the entire *SCI* database, 0.258 between *CiteSeer* and the computer science journals in *SCI*, and 0.824 between *SCI* and its computer science journals. Clearly, the author ranking resulting from *CiteSeer* is more similar to that from the entire *SCI* database than to that from *SCI* restricted to computer science journals. We can take that as an indication that *CiteSeer* is really a database for computer science literature in a quite broadly defined sense, and that the differences we observed in an earlier study (Zhao & Logan, 2002) between results from *CiteSeer* and those from the entire *SCI* database cannot be explained by the multidisciplinary nature of the *SCI* database as we surmised in that study. We will derive a different explanation for this phenomenon below based on the more controlled data underlying the present study.

4.1.2. Impact of authors who publish in different media

Our earlier study (Zhao & Logan, 2002) observed that there were more authors in the Web group than in the journal group who have great impact on XML research in terms of the frequency with which they have been cited in the literature. The current data, as discussed below, show that this is true only when very highly visible authors in terms of number of publications are considered. If authors who are not as highly visible are included, more authors in the journal group would be among the highly cited authors.

A list of highly cited authors was obtained by taking authors whose number of citations divided by the total number of citing papers in the corresponding dataset is greater than 0.018 in *CiteSeer* or in *SCI*. We call these "high-impact authors" for the convenience of discussion. This list of high-impact authors contains 42% of the authors in *CiteSeer* and 36% of the authors in *SCI* whose fractional publication counts are greater than one. However, the same list of high-impact authors contains 19% of the authors in *CiteSeer* and 22% of those in *SCI* who were among the top 100 or so authors ranked by fractional publication counts.

This indicates that among those scholars who were publishing on the Web, only very highly visible ones are likely to be recognized by the community, and that regular scholars' publications in print journals tend to be more widely accepted than those on the Web. This shows that it is still of some concern among XML scholars whether work published in venues other than print journals will be recognized by the community.

If the top authors from both databases (ranked by number of publications) are grouped into three categories: (1) authors who actively publish both on the Web and in journals as indexed by *CiteSeer* and by SCI, (2) authors who actively publish in journals but not on the Web, and (3) authors who actively publish on the Web but not in journals, then the same list of high-impact authors contains 47% of the authors in group 1, 19% of those in group 2, and 13% of those in group 3. Clearly, there are considerably more influential authors in the group who publish actively in both media than in those groups who only publish in one of the media. This indicates that the public availability of

scholars' work on the Web can well contribute to scholars' becoming more influential, though it may not be a decisive factor.

4.1.3. Contributing factors

The interesting question now is what may have contributed to this pattern of publication.

In order to find an answer to this question, data about authors' characteristics, such as age, research topics, publishing history, affiliation, and collaboration preferences, were collected and examined. These data are not presented here in a table due to space limitations but they can be found in Zhao (2003) and will be discussed below.

One might expect to see relatively more young scholars on the Web than in journals as it would appear easier for younger scholars to adapt to new technologies than for scholars who have had a long publishing history and who therefore may not be attracted to publishing on the Web as easily. One might also expect to find more scholars on the Web than in journals who have been involved in large group collaboration, as that requires open and effective communication and the Web should be a perfect medium for this. It might also be expected that relatively more scholars from countries outside of North America be seen on the Web as there is a well-documented bias in *SCI* toward North American journals.

However, none of these expected patterns emerged from the data we collected in the XML research field. Neither in fact did we see any clear patterns about how authors' age, publishing history, nationality, affiliation or collaboration preferences contributed to their medium preference. For example, it appears to be true with both the Web group and the journal group that there are almost as many prolific experienced scholars as there are young scholars with short publishing history, and that most of the highly visible scholars were professors or students when they published the articles in this study.

The only really clear factor that we were able to discover is that an author's research area was closely related to his or her publishing behavior. Scholars who studied the application of computer science in general, and of XML in particular, tended to publish mainly in journals. For example, ranked at the very top in *SCI* as shown in Table 2, P. Murray-Rust is a scientist in Computational Biology (Bioinformatics, Molecular informatics) and H. S. Rzepa one in Computational Chemistry. They both edited the Chemical Markup Language (CML) — a formal XML language in Chemistry. Gkoutos who is ranked 4th in *SCI*, was Rzepa's student and was also involved in CML related research. Similarly, Dolin, who is ranked right after Gkoutos in *SCI*, is a researcher in the area of medical informatics who has done research on XML for medical information exchange and was involved in the development of related standards. None of these top ranked scholars in *SCI* in terms of their number of publications appeared at all in *CiteSeer*.

This is not difficult to understand. Scholars in an application area of a technology (e.g. computational biology) may have adapted to the publishing tradition within that field (here, biology or chemistry) which may be different from that in the core field of the technology they apply (here, computer science). Although XML researchers in the computer science field are heavily publishing on the Web, scholars in the application areas of XML may not do so because they act more like, say, biologists than like computer scientists in terms of their publishing behavior.

4.2. Author visibility indicated by number of citations

An author ranking by straight citation counts was produced based on each of the three data sets — the data set from *CiteSeer*, the one from the entire *SCI* database and that from a subset of *SCI* addressing computer science research. Part of these rankings — authors whose citation counts, divided by number of citing papers in corresponding datasets, were 0.035 or higher — is presented in Table 3. More complete lists can be found in Zhao (2003).

Again, as discussed earlier, for reasons of convenience, we only aimed to select the top ranked 100 authors in each ranking for closer examination. Although the goal was to select 100 authors, the number "100" was not used strictly as a cut-off but as a guideline so that authors with the same number of citations are treated identically: either all or none being selected. As a result, the number of citations used in selecting the top ranked authors was different from one ranking to another. Specifically, the criterion of "6 or more citations" was used for the ranking from the computer science journals in *SCI* and that from *CiteSeer*, resulting in 100 and 90 authors respectively, and "7 or more citations" was used for the ranking from the entire *SCI* databases resulting in a list of 103 authors. These criteria were those that each resulted in a number of authors that was the closest to the goal of 100 top-ranked authors.

Table 3: Authors ranked by number of citations
(straight counts, divided by total # citing papers, 0.035 or greater)

ResearchIndex		SCI		Computer science journals in SCI	
Name	**#c**	**Name**	**#c**	**Name**	**#c**
S. Abiteboul	0.351	S. Abiteboul	0.222	S. Abiteboul	0.25
P. Buneman	0.242	T. Bray	0.206	T. Bray	0.209
A. Deutsch	0.208	A. Deutsch	0.152	P. Buneman	0.179
T. Bray	0.199	P. Buneman	0.152	A. Deutsch	0.164
J. Clark	0.186	P. Murrayrust	0.131	J. Clark	0.127
R. Goldman	0.143	J. Clark	0.123	M. Fernandez	0.119
M. F. Fernandez	0.134	M. Fernandez	0.12	J. Robie	0.097
D. Florescu	0.115	J. Robie	0.088	Y. Papakonstantinou	0.097
Stefano Ceri	0.106	H. S. Rzepa	0.086	P. Murrayrust	0.093
J.Shanmugasundaram	0.093	Y. Papakonstantinou	0.08	R. Goldman	0.086
J. Robie	0.09	R. Goldman	0.08	S. Ceri	0.082
J. McHugh	0.087	D. Florescu	0.067	S. Cluet	0.075
Y. Papakonstantinou	0.081	R. H. Dolin	0.064	S. J. Derose	0.075
H. Thompson	0.078	S. Cluet	0.064	J. Bosak	0.071
Sophie Cluet	0.078	T. J. Berners-Lee	0.064	R. H. Dolin	0.071
S. S. Chawathe	0.071	J. Bosak	0.061	C. Goldfarb	0.067
Makoto Murata	0.068	S. Ceri	0.061	T. J. Bernerslee	0.063
D. D. Chamberlin	0.065	C. Goldfarb	0.059	G. Wiederhold	0.06
Wenfei Fan	0.065	S. J. Derose	0.059	H. S. Rzepa	0.06
R. G. G. Cattell	0.053	D. D. Chamberlin	0.053	D. Florescu	0.056
S. DeRose	0.053	C. Friedman	0.051	P. Wadler	0.056
C. Beeri	0.05	J. Shanmugasundara	0.051	T. Milo	0.056
Tova Milo	0.05	G. V. Gkoutos	0.045	E. Maler	0.052
W. van der Aalst	0.05	T. Milo	0.045	H. Hosoya	0.052
C. Brew	0.047	A. Y. Levy	0.043	A. Bruggemannklein	0.049
H. Hosoya	0.047	G. Wiederhold	0.043	A. Hunter	0.049
O. Lassila	0.047	H. Hosoya	0.043	S. S. Chawathe	0.045
P. Wadler	0.047	J. Mchugh	0.04	C. Friedman	0.041
V. Christophides	0.047	L. Liu	0.04	D. Calvanese	0.041
E. Maler	0.043	P. Wadler	0.04	F. Neven	0.041
Angela Bonifati	0.04	S. S. Chawathe	0.04	G. Hripcsak	0.041
Jennifer Widom	0.04	D. Gardner	0.037	P. Atzeni	0.041
T. Berners-Lee	0.04	E. Maler	0.037	V. Christophides	0.037
D. Brickley	0.037	A. Bruggemannklein	0.035		
Michael Hanus	0.037	A. Hunter	0.035		
		R. GG. Cattell	0.035		
		V. Christophides	0.035		

4.2.1. Common authors and correlations between their rankings

Among the top authors thus found, 49 authors were common to all three lists, 55 were common to the lists from *CiteSeer* and from the entire *SCI*, 49 were common to the lists from *CiteSeer* and from the computer science journals in *SCI*, and 84 were common to the lists from the entire *SCI* and from its computer science journals. Through visual inspection of a small set of highly cited authors, our earlier study (Zhao & Logan, 2002) observed a high correlation for the top 10 authors between the entire *SCI* database and *CiteSeer*. The present study examined the correlation statistically on a much larger scale: Pearson's r's were calculated both for common authors between the three lists of top ranked roughly 100 authors and for all authors that were common to all three full datasets. The Pearson's r for the 49 top ranked common authors is 0.92 between *CiteSeer* and *SCI*, 0.91 between *CiteSeer* and the computer science journals in *SCI* and 0.98 between *SCI* and its computer science journals. The Pearson's r's for all the 576 authors that are common to the three datasets turned out to be very similar to those for the 49 top ranked authors: 0.92 between *CiteSeer* and *SCI*, 0.91 between *CiteSeer* and the computer science journals in *SCI* and 0.99 between *SCI* and its computer science journals.

This shows that the Pearson's r between the author ranking from the entire *SCI* database and that from the portion of *SCI* addressing computer science research is very high. The number of top ranked authors shared by the two rankings is fairly high as well. This indicates that current XML research is still mostly limited to computer science or that studies on the application of XML technology in different fields have been publishing in journals that are considered by *SCI* as belonging to computer science research, or it may mean that *SCI*, like *CiteSeer*, defines computer science journals rather broadly.

The Pearson's r between the author ranking resulting from *CiteSeer* and that from *SCI* computer science data was unexpectedly lower than that between *CiteSeer* and the *SCI* entire database although both were significant and their difference was very small. This suggests that publishing medium may have played a more important role than discipline in shaping the citing authors' perceptions of cited authors' visibility in the XML research field.

Since results from the entire *SCI* database and those from computer science journals were very highly correlated and the results obtained from the *CiteSeer* database were more closely correlated to those obtained from the entire *SCI*, which has also been shown in the comparison between author rankings by number of publications discussed earlier, the following discussion will no longer include the dataset from the computer science journals in *SCI*.

The high correlations between the author rankings indicate that, when the same citation counting methods are used, a group of authors will be ranked very similarly no matter which of the two data sources is used, *CiteSeer* or *SCI*. This confirms our interpretation of the results we found in our earlier study that citation analysis using *CiteSeer* as a data source is as valid for evaluating scholars as is citation analysis based on *SCI* data, the data source still most widely used in the literature and widely validated in evaluation studies for scholars and scholarly contributions. This also suggests that publications on the Web should no longer be ignored either as part of the literature for research or as a data source for the study of scholarly communication because they are similar to those in print journals in terms of the way they refer to earlier publications. If this can be confirmed even more strongly in the future, it is good news for informetric scholars who either might not have easy access to *SCI* data, especially those in some developing countries, or who might investigate research areas or researcher populations that are under-represented in *SCI*, because they may now be able to conduct citation analysis studies using data and tools freely available on the Web and still be confident of getting valid results.

However, although the correlations between author rankings are high for the common authors, common authors only account for about half of the top 100 or so highly cited authors from each dataset as seen from the numbers above. This challenges the common practice in science evaluation in which the ISI database is used exclusively to obtain citation data on performance, and clearly indicates that the best way to evaluate scholars using a citation analysis approach is to combine multiple data sources, such as *SCI* and *CiteSeer*, so that the data sources can complement each other and the evaluation results become less biased.

4.2.2. Different citation patterns on the Web and in journals

Clearly there are three groups of authors: (1) authors who are highly cited by both documents in *SCI* and those in *CiteSeer*, (2) authors who are highly cited by documents in *SCI* but not by those in *CiteSeer*, and (3) authors who are highly cited by documents in *CiteSeer* but not by those in *SCI*.

Our earlier study (Zhao & Logan, 2002) pointed out the importance of examining the characteristics of author groups with different citation patterns. In the present study, some characteristics of the three groups of authors were

studies in an attempt to identify factors that contribute to the differences in author visibility between *SCI* and *CiteSeer*. Due to space constraints these data are not presented here as a table but discussed below (see Appendix H of Zhao, 2003, for complete data).

An examination of authors' characteristics suggests some interesting aspects of scholarly communication in the XML research field.

- The majority of the authors who are highly cited in both of the data sources either belong to one or more of several interrelated research groups or have been involved in World Wide Web Consortium (W3C) working groups for XML related standards or specifications.

Bray, Clark, Fernandez, Robie, Thompson, Chamberlin, Murata, DeRose, Lassila, Maler, Wadler, Brickley, Apparao, Berners-Lee, Bosak, and Decker are all members of W3C working groups for XML related standards or specifications. Abiteboul, Florescu, Cluet, Milo, and Christophides belong to the French Project Verso research group; Buneman, Deutsch, Hosoya, and Sahuguet belong to the database group at the University of Pennsylvania; Goldman, McHugh, Papakonstantinou, Chawathe, Widom, and Ullman belong to the database group at Stanford University (mostly the Lore project); and Ceri and Bonifati belong to a group of Italian researchers. All these groups are interrelated not only intellectually as indicated by co-citation but also socially as indicated by co-authorship.

- Foundational and historical materials and general opinion papers are highly cited in journals but not on the Web.

This can be seen quite clearly from Bosak, Goldfarb and Berners-Lee being highly cited in *SCI* but not in *CiteSeer*.

Charles F. Goldfarb is the father of "markup languages" and the main author of the Standard Generalized Markup Language (SGML), on which the Web's HTML and XML are based. Although XML as a subset of SGML is much simpler and specifically designed for the Web, and as a result has gained wide recognition and application in the Web context, SGML is still heavily being used in the publishing industry. Goldfarb's two handbooks about SGML and XML respectively are highly cited in *SCI* but not in *CiteSeer*. Handbooks represent more mature and secondary materials and therefore are very useful in the industry. Scholars at the research front however may only refer to them as historical background as these scholars may have found the original material such as the ISO standard on SGML more convenient to use, and hence, to cite.

Tim Berners-Lee is renowned as the father of the World Wide Web and retains today an influential position as the director of the World Wide Web Consortium, while Bosak organized and led the XML working group in the development of the seminal XML specification. Both Berners-Lee and Bosak have thus laid the foundation of XML and other Web related technologies and have written influential opinion papers, such as *XML, Java, and the Future of the Web* (Bosak, 1997), *XML and the second-generation web* (Bosak & Bray, 1999), *Weaving the Web* (Berners-Lee, et al., 1999) and *The Semantic Web* (Berners-Lee, et al., 2001).

These scholars being highly cited in *SCI* but less so in *CiteSeer* suggests that papers on the Web are perhaps more at the research front than those in journals which are still referring to a considerable extent to foundational and historical materials and to opinion papers and may therefore contain more reviews and research at earlier stages.

- Authors in application areas of XML, such as Chemical Markup Language (CML) and XML for medical information exchange, are not as well represented in *CiteSeer* as in *SCI*.

As we can see from the earlier discussion about author visibility as indicated by number of publications, scholars in application areas of XML have not published often on the Web but more in journals. Examples include Murray-Rust and Rzepa who led the CML specification effort and Friedman and Dolin who have been involved in research on XML for medical informatics. The citation patterns of scholars in these areas are very similar to their publication patterns: highly cited in *SCI* but not in *CiteSeer*. This reveals two things clearly.

First, scholars in these areas have primarily been cited from within their own areas, suggesting that these areas are relatively independent of the rest of XML research and may be quite narrow.

Second, application areas of XML are clearly not well represented on the Web. That means that citation analysis using data and tools on the Web as a data source is currently limited to certain research fields where Web publishing is well accepted by the communities, and may be biased by leaving out either partly or completely certain specialties

in which scholars have different publishing behaviors. This limitation would exist until scholarly publishing on the Web is as widely accepted and practiced as publishing in journals.

- Authors in *the Semantic Web* area and in the *Programming / processing of XML data* specialty are not as well represented in *SCI* as in *CiteSeer*.

Many of the authors in *the Semantic Web* area were ranked much higher by number of citations in *CiteSeer* than in *SCI*. Examples include Lassila (24.5 in RI vs. 35.5 in SCI), Brickley (31 vs. 53), and Fensel (34 vs. 46). Others were highly cited in *CiteSeer* but rarely or not at all in *SCI*. For example, Hanus and Horrocks received 12 and 7 citations in *CiteSeer* but only 0 and 2 in *SCI* respectively. This suggests that research in *the Semantic Web* area is better represented on the Web. Since this area of research was still emerging at the time our data was collected, and was aiming to develop "the next generation" of the Web, we have further evidence that research reported on the Web is perhaps more at the cutting-edge than that reported in journals.

Among the authors who were categorized by an SPSS factor analysis routine into the factor that has been identified as representing the specialty *XML and programming*, most, including Megginson, Klarlund, Bourret, Aho, Schmidt, and Carl-Christian Kanne, were only highly cited on the Web. Those who were highly cited both in *CiteSeer* and in *SCI* were ranked much higher in *CiteSeer* than in *SCI*. Examples include Murata (17 in RI vs. 35.5 in *SCI*) and Lee (34 vs. 46). This indicates that research in this area is also better represented on the Web than in journals.

Therefore, the discussion above about the possible bias caused by data and tools on the Web applies to *SCI* as well. In other words, citation analysis using *SCI* data as the only data source may also be biased by leaving out in part or completely certain specialties in which scholars publish heavily in venues other than the journals indexed by *SCI*.

- Research areas are the major contributing factor to differences in author visibility between different media. Age, nationality, collaboration preferences did not seem to have had much effect.

Research on XML database design and implementation is well represented in both media, but research on XML applications is better represented in journals, and research on *the Semantic Web* and *Programming / processing of XML data* is more visible on the Web.

In both media, highly cited authors appear to be mostly from North America and Europe (especially France, Italy and the United Kingdom). French and Italian researchers have been very active in research on XML database design and implementation, while scholars from the UK, such as Clark, Thompson, Brickley, Murray-Rust and Rzepa, have been actively involved in the development of many of the XML-related standards or specifications.

Although there do appear to be more authors among the "print-only" scholars than among the "Web-only" scholars who are relatively older and more experienced or less active in large group collaboration, the difference does not seem to be significant. For example, as discussed earlier, both Suciu and Fan have been involved in large group collaboration but one (Fan) was highly cited only on the Web and the other (Suciu) only in journals. Another example is that young scholars can be highly cited both on the Web (e.g. Fan and Nestorov) and in journals (e.g. Dolin and Gkoutos), probably depending on the topics of their research.

5. Discussion

The findings presented above can shed some light on issues of both citation analysis and scholarly communication in transition.

5.1. Citation analysis

Citation analysis has had a long history in the study of scholarly communication, and until recently the ISI databases have served as virtually the only data source for citation analysis studies. The incompleteness, bias, and limitations of this data source have been well studied in the literature; nevertheless, they remained the basis for most such studies, partly because these databases have been the only ones available for this purpose. Despite these problems, studies based upon ISI databases have provided valuable insight into scholarly communication patterns in many fields.

With the Web becoming a more and more powerful communication medium, full text research papers (including reference lists) are increasingly available on the Web. Search engines and even citation indexes are emerging to help researchers make full use of these resources. It is therefore natural for scholarly communication researchers to be

tempted to use these papers and indexes as a data source for citation analysis studies as they may avoid many of the problems associated with the ISI databases, such as their "first author only" approach to indexing cited papers with multiple authors. However, we have only seen a small number of studies that have made extensive use of Web data sources. The concerns citation analysts may have about their use may include: (1) Lacking a formal refereeing process, Web-publishing is not as well controlled as journal publishing and therefore might be viewed as being flawed for citation analysis. (2) Citation analysis of data from the ISI databases is considered complete enough to get a valid picture of scholarly communication patterns in a research field as papers indexed in these databases are considered to be "the most important" portion of the literature in the field. (3) Data and tools on the Web do not cover as many disciplines and are often not as easy to use as the ISI databases.

Findings from the present citation analysis study of XML research may help address some of these concerns.

First, the author ranking by number of citations that resulted from *CiteSeer* data is highly correlated with that obtained from *SCI*. In other words, regarding the impact of a group of scholars on research in the XML field, the collective view of citers on the Web is very similar to that of citers in journals. Evaluation of scholars based on this view should thus be considered as equally valid, provided the discipline being studied is well published on the Web.

Second, the two groups of XML scholars who actively publish on the Web or in journals, respectively, share very few publications, and are concerned with different issues. While all study XML related standards and specifications or XML database design and implementation, research on XML applications is a focus only in print journals, and research on *the Semantic Web* and on *programming for, and processing of, XML data* is better represented on the Web. That means that, in order to gain a complete picture of the scholarly communication pattern in a research field and to obtain a less biased indicator of scholars' performance in the field, multiple data sources should ideally be used rather than only the ISI database or only *CiteSeer*.

Third, although there are many advantages to using data and tools on the Web for citation analysis, it is true that they do not currently cover as many disciplines and are not as easy to use as the ISI database in some sense (Zhao & Logan, 2002; Zhao, 2003). However, these are precisely some of the aspects in scholarly communication systems that need to be improved upon and to which citation analysis can contribute. For example, we can investigate how to design and implement a problem solving environment (PSE) for scholarly communication research that would put together in a user-friendly environment all the computational facilities needed for studying problems in scholarly communication, including easy access to *SCI* and *CiteSeer* and support for seamless integration of new data sources, tools for constructing citation indexes from existing full text contents, programs for various citation and co-citation counting methods, statistical analysis tools, and visualization tools (Zhao & Strotmann, 2004).

5.2. Scholarly communication in transition

The differences in XML research focus in the two media as discussed above along with other findings from the present study may also shed some light on issues of scholarly communication in transition. As mentioned above, XML applications were found to be a focus only in print journals, while research into the Semantic Web and the programming for, and processing of, XML data was seen to be more visible on the Web. From the point of view of XML research, unlike the Semantic Web and the programming for and processing of XML data, XML applications are about relatively mature rather than cutting-edge technologies. This may be evidence of research on the Web being more at a research front than that in journals. This was also suggested by foundational and historical material being more highly cited in journals than on the Web, such as handbooks and opinion papers by Goldfarb, Bosak and Berners-Lee — some of the inventors of the Web and XML related technologies.

These results appear to provide evidence for a "two-tier system" in scholarly communication that is believed by some scholars to be a probable future model of the scholarly communication system (Poultney, 1996; van Raan, 2001). In this model, the first tier is predicted to be a "free space" which represents the scholarly enterprise in "real time" and is most likely to feature free and timely Web-based publications, while the second tier is thought to be the world of more formal publications that is most likely to continue to be dominated by journals (van Raan, 2001, p. 61). As suggested by the present study and by other studies (Chan, 2004; Crow, 2002; Zhao, 2004), the first tier would primarily serve as an information distribution medium that improves the effectiveness and efficiency of the informal communication, on which scholars have relied heavily to obtain the information they need for their research, while the second tier would primarily serve as an archive and evaluation rather than information distribution device. The faster and wider distribution of information on the Web makes the Web a perfect medium for the initial publication of new research results in the first tier, while the journal has served well as an archive and evaluation device for a long time, which makes it natural to continue its role in the second tier.

As we concluded in our study of the intellectual structure of XML research (Zhao, 2004), if this system evolves, journals that currently do not accept papers published on the Web may have to change their policies, and all journals may eventually implement new procedures to reduce or eliminate the time scholars spend reformatting their research papers for journal acceptance after they have been published on the Web. This could significantly improve the efficiency of scholarly communication.

6. Conclusion

Scholarly communication is increasingly being conducted over the Internet, certainly in the core research areas of computer science but increasingly also in other sciences. This has brought some challenges as well as opportunities to the traditional scholarly communication systems and to the associated evaluation approaches. The present study has explored these through a citation analysis of author visibility in one particular research field — the XML research field. It has revealed different scholarly communication patterns in this field on the Web and in print journals in terms of author visibility, has challenged the common practice of using the ISI's databases exclusively to obtain citation counts as scientific performance indicators, and has found evidence for an emerging "two tier" scholarly communication system.

As demonstrated in this study, with scholarly communication increasingly being carried out over the Internet, it has become both important and feasible to use multiple citation data sources in citation analysis studies of scholarly communication, including scholarly publications on the Web, in digital libraries and in institutional repositories in addition to the journal articles indexed by the ISI databases. The inclusion of these new formats of scholarly publications in the evaluation of science not only can contribute to obtaining a more balanced and more complete evaluation, but also may promote a more efficient two-tier scholarly communication system. Scholars may be more willing to publish their works in these new formats if they can rest assured that their work is counted in the evaluation of science, just as scholars now prefer to publish in journals indexed by the ISI databases. We hope that the present study motivates further research in this area to contribute to the transition of the scholarly communication system from one that has evolved with print journals as the center to one that works best for the networked digital environment.

Acknowledgments

The author wishes to thank Dr. Andreas Strotmann of the Center for Applied Computer Science at the University of Cologne, for his many helpful insights.

This work was supported in part by a Fellowship of the School of Computational Science and Information Technology, Florida State University.

References

Abrams, M., Allison, D., Kafura, D., Ribbens, C., Rosson, M.B., Shaffer, C., Watson L. (n.d.). PSE Research at Virginia Tech: An Overview. Retrieved January 2003, from http://research.cs.vt.edu/pse/intro.html

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

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes — a review and analysis. *Scientometrics, 50*, 7-32

Borgman, C.L., & Furner, J. (2002). Scholarly communication and bibliometrics. In *Annual Review of Information Science and Technology, 36* (pp. 3-72). Medford, NJ: Information Today

Chan, L. (2004). Supporting and enhancing scholarship in the digital age: the role of open-access institutional repositories. *Canadian Journal of Communication, 29*, 277-300

Clever Project. (1999). Hypersearching the Web. Retrieved 2000, from http://www.sciam.com/1999/0699issue/0699raghavan.html

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

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

Crow, R. (2002). The case for institutional repositories: a SPARC position paper. Retrieved Feb. 8, 2005, from The Scholarly Publishing & Academic Resources Coalition website http://www.arl.org/sparc/IR/ir.html

Dahal, TM. (2000). Cybermetics: the use and implications for Scientometrics and Bibliometrics: a study for developing science & technology information system in Nepal. Retrieved 2000, from http://www.panasia.org.sg/nepalnet/ronast/cyber.html

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

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics*. New York: Elsevier Science Pub, 1990

Gallopoulos, E., Houstis, E., & Rice, J.R. (1994). Problem-solving environments for computational science. *IEEE Computational Science & Engineering, 1*, 11-23

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 and Management, 37*, 661-675

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*, 602-615

Harter, S. P., & Kim, H. J. (1996). Electronic Journals and Scholarly Communication: A citation and reference study. In *Proceedings of the Midyear Meeting of the American Society for Information Science* (pp. 299-315)

Institute for Scientific Information (2004a). The Impact of Open Access Journals: A Citation Study from Thomson ISI. Retrieved May 8, 2004, from http://www.isinet.com/media/presentrep/acropdf/impact-oa-journals.pdf

Institute for Scientific Information (2004b). The ISI Database: the journal selection process. Retrieved May 30, 2004, from http://www.isinet.com/essays/selectionofmaterialforcoverage/199701.html/

Larson, R. R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the 59th the American Society for Information Science Annual Meeting* (pp. 71-78). Medford, NJ: Information Today

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

Lawrence, S., Bollacker, K., & Giles, C. L. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6), 67-71

Lindsey, D. (1980). Production and citation measures in the sociology of science: the problem of multiple authorship. *Social Studies of Science, 10*, 145-162

Lu, S. (1999). *The transition to the virtual world in formal scholarly communication: a comparative study of the natural sciences and the social sciences*. Dissertation, University of California, Los Angeles, 1999

McCain, K. W. (2000). Sharing digitized research-related information on the World Wide Web. *Journal of the American Society for Information Science, 51*, 1321-1327

The Open Citation Project. (2001). Mining the social life of an eprint archive. Retrieved October 20, 2001, from http://opcit.eprints.org/tdb198/opcit/

Poultney, R. W. (1996). Front-ends are the way to go. *Europhysics News, 27*, 24-25

Rice, J.R., & Boisvert, R.F. (1996). From scientific software libraries to problem-solving environments. *IEEE Computational Science & Engineering, 1996*(Fall), 44--53.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics. 1*(1). Retrieved October 10, 2001, from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Thelwall, M. & Harries, G. (2004). Do the Web sites of higher reated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology, 55*, 149-159

Turnbull, D. (2000). Bibliometrics and the World-Wide Web. Retrieved 2000, from http://donturn.fis.utoronto.ca/research/bibweb.html

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: consequences for the impact of authors. *Journal of the American Society for Information Science, 48*, 944-945

Van Raan, A. F. J. (2001). Bibliometrics and Internet: some observations and expectations. *Scientometrics, 50*, 59-63

Youngen, G. (1997). Citation patterns of the physics preprint literature with special emphasis on the preprints available electronically. Retrieved 2000, from http://www.physics.uiuc.edu/library/preprint.html

Zhao, D. (2003). *A comparative citation analysis study of Web-based and print journal-based scholarly communication in the XML research field*. Doctoral dissertation, School of Information Studies, Florida State University

Zhao, D. (2004). Web-based and print journal-based scholarly communication in the XML research field: a look at the intellectual structure. In *Proceedings of the American Society for Information Science and Technology 2004 Annual Meeting: Managing and Enhancing Information: Cultures and Conflicts*, (pp. 72-83). Medford, NJ: Information Today

Zhao, D., & Logan, E. (2002). Citation analysis of scientific publications on the Web: a case study in XML research area. *Scientometrics, 54*, 449-472

Zhao, D. and Strotmann, A. (2004). Towards a Problem Solving Environment for Scholarly Communication Research. In *Proceedings of Canadian Association for Information Science (CAIS) 2004 Annual Conference:*

Access to Information: Technologies, Skills, and Socio-Political Context (available online at http://www.cais-acsi.ca/proceedings/2004/zhao_2004.pdf)

Table 1
Distribution of citing papers by number of authors

# papers / # authors	ResearchIndex		SCI	
	#	%	#	%
0	4	1	0	0
1	83	27	75	20
2	77	25	99	26
3	78	25	86	23
4	36	12	54	14
5 or more	34	11	60	16

Table 2
Authors ranked by number of publications

ResearchIndex		SCI		Computer science journals in SCI	
Name	**#p**	**Name**	**#p**	**Name**	**#p**
Wenfei Fan	4.07	H. S. Rzepa	4.07	A. Hunter	2
D. Fensel	2.9	P. Murray-rust	3.57	M. Rezayat	2
Dan Suciu	2.82	D. Suciu	3.2	P. T. Wood	2
Serge Abiteboul	2.68	G. V. Gkoutos	2.15	S. J. Derose	2
M. Murata	2.67	R. H. Dolin	2.05	W. Weitz	2
J. Simeon	2.45	A. Hunter	2	J. Dudeck	1.95
Angela Bonifati	2.33	A. Kristensen	2	H. S. Rzepa	1.92
Harold Boley	2.33	J. Hunter	2	S. Paraboschi	1.62
Dongwon Lee	2.25	M. Rezayat	2	P. Murray-rust	1.58
Mark Huckvale	2	P. T. Wood	2	M. Fernandez	1.53
S. Ceri	1.9	S. J. Derose	2	H. Kim	1.5
Amarnath Gupta	1.75	W. Weitz	2	K. Canfield	1.5
Victor Vianu	1.73	J. Dudeck	1.95	N. Sundaresan	1.5
Daniela Florescu	1.7	M. F. Fernandez	1.87	E. Bertino	1.33
M. F. Fernandez	1.65	J. Simeon	1.75	F. A. Fontana	1.33
Wolfgang Emmerich	1.58	S. Ceri	1.7	G. Weikum	1.33
L. Libkin	1.5	E. Bertino	1.67	E. Damiani	1.28
Leonidas Fegaras	1.5	S. Paraboschi	1.62	L. Kerschberg	1.25
Torsten Schlieder	1.5	M. Wright	1.57	L. Rutledge	1.25
W. van der Aalst	1.5	A. Sahuguet	1.5	S. Ceri	1.2
Elena Ferrari	1.45	H. Kim	1.5	J. Simeon	1.17
John Miller	1.33	J. R. Smith	1.5	M. Shields	1.15
A. Finkelstein	1.25	K. Canfield	1.5	R. H. Dolin	1.05
B. Ludascher	1.25	N. Sundaresan	1.5		
F. Tian	1.25	C. M. Chiu	1.33		
I. Schena	1.25	F. A. Fontana	1.33		
M. Mani	1.25	G. Weikum	1.33		
Letizia Tanca	1.2	E. Damiani	1.28		
Marin Dimitrov	1.2	A. Zisman	1.25		
S. Saeyor	1.2	L. Kerschberg	1.25		
D. Kossmann	1.17	L. Rutledge	1.25		
E. Damiani	1.15	A. Y. Halevy	1.25		
S. Paraboschi	1.12	M. Shields	1.15		
Piero Fraternali	1.07				
Philip Wadler	1.03				

Note: authors with fractional counts greater than 1 are presented.

Table 3
Authors ranked by number of citations

ResearchIndex		SCI		Computer science journals in SCI	
Name	**#c**	**Name**	**#c**	**Name**	**#c**
S. Abiteboul	0.351	S. Abiteboul	0.222	S. Abiteboul	0.25
P. Buneman	0.242	T. Bray	0.206	T. Bray	0.209
A. Deutsch	0.208	A. Deutsch	0.152	P. Buneman	0.179
T. Bray	0.199	P. Buneman	0.152	A. Deutsch	0.164
J. Clark	0.186	P. Murrayrust	0.131	J. Clark	0.127
R. Goldman	0.143	J. Clark	0.123	M. Fernandez	0.119
M. F. Fernandez	0.134	M. Fernandez	0.12	J. Robie	0.097
D. Florescu	0.115	J. Robie	0.088	Y. Papakonstantinou	0.097
Stefano Ceri	0.106	H. S. Rzepa	0.086	P. Murrayrust	0.093
J.Shanmugasundaram	0.093	Y. Papakonstantinou	0.08	R. Goldman	0.086
J. Robie	0.09	R. Goldman	0.08	S. Ceri	0.082
J. McHugh	0.087	D. Florescu	0.067	S. Cluet	0.075
Y. Papakonstantinou	0.081	R. H. Dolin	0.064	S. J. Derose	0.075
H. Thompson	0.078	S. Cluet	0.064	J. Bosak	0.071
Sophie Cluet	0.078	T. J. Berners-Lee	0.064	R. H. Dolin	0.071
S. S. Chawathe	0.071	J. Bosak	0.061	C. Goldfarb	0.067
Makoto Murata	0.068	S. Ceri	0.061	T. J. Bernerslee	0.063
D. D. Chamberlin	0.065	C. Goldfarb	0.059	G. Wiederhold	0.06
Wenfei Fan	0.065	S. J. Derose	0.059	H. S. Rzepa	0.06
R. G. G. Cattell	0.053	D. D. Chamberlin	0.053	D. Florescu	0.056
S. DeRose	0.053	C. Friedman	0.051	P. Wadler	0.056
C. Beeri	0.05	J. Shanmugasundara	0.051	T. Milo	0.056
Tova Milo	0.05	G. V. Gkoutos	0.045	E. Maler	0.052
W. van der Aalst	0.05	T. Milo	0.045	H. Hosoya	0.052
C. Brew	0.047	A. Y. Levy	0.043	A. Bruggemannklein	0.049
H. Hosoya	0.047	G. Wiederhold	0.043	A. Hunter	0.049
O. Lassila	0.047	H. Hosoya	0.043	S. S. Chawathe	0.045
P. Wadler	0.047	J. Mchugh	0.04	C. Friedman	0.041
V. Christophides	0.047	L. Liu	0.04	D. Calvanese	0.041
E. Maler	0.043	P. Wadler	0.04	F. Neven	0.041
Angela Bonifati	0.04	S. S. Chawathe	0.04	G. Hripcsak	0.041
Jennifer Widom	0.04	D. Gardner	0.037	P. Atzeni	0.041
T. Berners-Lee	0.04	E. Maler	0.037	V. Christophides	0.037
D. Brickley	0.037	A. Bruggemannklein	0.035		
Michael Hanus	0.037	A. Hunter	0.035		
		R. GG. Cattell	0.035		
		V. Christophides	0.035		

Note: numbers are straight counts, divided by total number of citing papers; 0.035 or greater are presented

Dangzhi Zhao is an Assistant Professor at the School of Library and Information Studies of the University of Alberta, Canada. Her research interests are in the areas of information systems, scholarly communication, and information technology. Recent research concentrates on the comparison of scholarly communication between the Web and the print world.