This is a preprint of an article published in *Journal of the American Society for Information Science and Technology, 58* (9): 1285-1302, 2007

# Can citation analysis of web publications better detect research fronts?

**Dangzhi Zhao**[1]

School of Library and Information Studies, University of Alberta, Edmonton, Alberta, Canada T6G 2J4. Phone: (780) 492-2814, Fax: (780) 492-2430, Email: dzhao@ualberta.ca

**Andreas Strotmann**

School of Business, University of Alberta, Edmonton, Alberta, Canada T6G 2J4.
Phone: (780) 492-26924, Fax: (780) 492-3325, Email: andreas.strotmann@ualberta.ca

## Abstract

We present evidence that, in some research fields, research published in journals and reported on the Web may collectively represent different evolutionary stages of the field with journals lagging a few years behind the Web on average; and that a "two-tier" scholarly communication system may therefore be evolving. We conclude that, in such fields, (a) for detecting current research fronts, author co-citation analyses (ACA) using papers published on the Web as a data source can outperform traditional ACAs using articles published in journals as data; and that (b), as a result, it is important to use multiple data sources in citation analysis studies of scholarly communication for a complete picture of communication patterns.

Our evidence stems from comparing the respective intellectual structures of the XML research field, a subfield of computer science, as revealed from three sets of ACA covering two time periods: (1) from the field's beginnings in 1996 to 2001; and (2) from 2001 to 2006. For the first time period, we analyze research papers both from journals as indexed by the *Science Citation Index (SCI)* and from the Web as indexed by *CiteSeer*. We follow this up by an ACA of *SCI* data for the second time period. We find that most of the trends in the evolution of this field from the first to the second time period that we find when comparing ACA results from *SCI* between the two time periods, were already apparent in the ACA results from *CiteSeer* during the first time period.

[1] Corresponding author

## 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 has been renewed interest in the study of scholarly communication in recent years (Borgman, 2000) to see the types of communication that are evolving in the new media and to compare them to what we have come to expect from journal based communication.

Citation analysis and other bibliometric techniques have been applied successfully to the study of this new phenomenon in scholarly communication (Cronin, 2001). Some studies have applied citation analysis and other bibliometric principles and techniques (often with modifications) to the analysis of characteristics and link structures of the Web. Among these are studies on search engines making use of hyperlink structure (Clever, 1999) and so-called "Webometric" studies (Cronin et al., 1998; Egghe, 2000; Larson, 1996; Rousseau, 1997; Thelwall, Vaughan, & Bjorneborn, 2005). Other studies have looked at "electronic ingredients" in journal articles – either in reference lists or in abstracts – to study the impact of electronic publications on traditional print journal-based scholarly communication (Harter, 1992; Harter & Kim, 1996; Youngen, 1997; McCain, 2000; ISI, 2004a). Still others have taken advantage of research papers that are increasingly available on the Web to conduct citation analysis studies using such papers as a data source in order to explore the opportunities brought to scholarly communication research by such publications and by tools for searching for citations from these papers (The Open Citation Project, 2001; Goodrum et al., 2001; Zhao, 2003; Zhao, 2004; Zhao, 2005; Zhao, 2006a; Zhao, 2006b; Zhao & Logan, 2002). Examples of such tools include *Google Scholar* (http://scholar.google.com/), *CiteSeer* (http://citeseer.ist.psu.edu/), and *CiteBase* (http://www.citebase.org/).

These tools are different from the citation indexes, such as *Science Citation Index (SCI)*, that are produced by the former Institute for Scientific Information (ISI), now Thomson Scientific (http://www.isinet.com/), and which have been used as the main source of data for most of the citation analysis studies published to date. For example, these new tools may cover a wider range of document types such as degree theses, technical reports, conference papers, or preprints in addition to journal articles, which represent different stages in the scholarly communication

process; they may provide more information about documents cited in the publications they index, such as all authors, full titles, and full names of sources, as compared with the limited information provided by the ISI indexes (i.e., first authors only, and abbreviations of source journal titles); and their source paper selection and indexing process may be highly automatic and more inclusive as compared with the manual and intentionally highly selective process employed by the ISI indexes (ISI, 2004b).

These data and tools have opened up opportunities for a variety of new inquiries: we can study how scholarly communication in some fields is being transformed by the new media, what the similarities or differences between the new formats and the traditional ones may be, how the new formats facilitate or inhibit the scholarly communication process, and how citation analysis theory and methods may be extended with the support of new citation data sources. Just as the advent of the ISI indexes greatly advanced the theory and broadened the applications of citation analysis, data and tools for citation analysis studies increasingly available on the Web may lead to another such step forward (Zhao & Logan, 2002).

The present study attempts to explore this possibility through a comparison of scholarly communication patterns in the eXtensible Markup Language (XML) research field between research papers publicly available on the Web on the one hand and those in scientific journals on the other. We seek (a) to identify similarities and differences in the intellectual structure between Web-based and journal-based scholarly communication in the XML research field; and (b) to explore possible contributing factors. This study may contribute to the understanding of scholarly communication in transition and to the advancement of citation analysis theory and methodology.

## Methodology

### *Research field analyzed*

The eXtensible Markup Language (XML) research field, a subfield of computer and information science, was chosen for this comparative study. This field was “digitally born” and has been growing with the Web – indeed, the XML technology was developed with the hope that it would provide the basis for the next generation of the Web.

As a result, the Web is naturally one of the main communication channels used by the XML research community, which makes it ideal for our particular study for the following reasons. (a) The number of research papers published on the Web in this field is likely to be large enough for applying a citation analysis approach even during the very early stages of its evolution. (b) Emerging models for scholarly communication, if there are any, are more easily identified in this particular field. At the same time, however, this also means that the extent to which results from a study of this field can be generalized or extrapolated to other fields may be limited because Web publications in these other fields may not represent their fields as completely as we would expect for the XML field due to inherent field differences in scholarly communication (Kling & McKim, 2000).

However, this fortunately does not undermine the value of the present study which lies not so much in the identification of characteristics in scholarly communication that apply universally, but more in its implications as an early indicator for understanding the transition of the scholarly communication system, even if more in some fields than in others (Kling & McKim, 2000), from a traditional journal-centered model to a model that utilizes the networked digital environment of the Web as a powerful medium for disseminating research results (Goodrum, et al., 2001). The potential of citation analysis based on publications on the Web demonstrated in the present study on the XML research field may also be applicable in other fields such as mathematics, physics, or computer science that are similar to the XML research field in terms of the degree to which they adopt Web publishing.

Thus, the XML research field was chosen here purposely to maximize the visibility of differences between the new and traditional publishing media and thus to amplify any effects of these differences, while ensuring for practical reasons that the authors of the present paper have the domain knowledge necessary for providing a reasonable interpretation of the citation analysis results in this field. In the future, we plan to expand this study to include both research fields that are similar to the XML research field and others that are different in terms of the degree to which they are related to the Web and have adopted Web publishing, to better judge the extent to which results agree across a range of research fields. In this paper, we observe intriguing differences between journal and Web-based publishing, and present a likely interpretation of the phenomena

we found. We thus point the way to important further research aimed at understanding more deeply the issues we raise here, perhaps raising more new questions than we answer.

### *Data sources*

Two separate citation indexes, namely the *Science Citation Index* (*SCI*) and *CiteSeer*, were queried in the present study to collect information on research papers published in journals and publicly on the Web, respectively, including their reference lists. These source papers are referred to as "citing papers" in the present study, and those included in their reference lists as "cited papers". The structure of the relationship between citing and cited papers as perceived by the authors of citing papers was then analyzed to reveal the intellectual structure of the XML research field as published in these two different venues and as indexed by these two different citation databases.

*SCI* is one of the ISI databases described earlier which primarily index journal articles while *CiteSeer* is one of the citation indexes for research papers found on the Web discussed above. To date, *SCI* and other ISI databases have been used as the data source for most of the citation analysis studies reported in the literature.

Originally developed by the NEC Corporation Research Institute, and now a joint effort of NEC and Pennsylvania State University's School of Information Sciences and Technology, *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.) which are in the broadly defined computer science field and are publicly available on the Web. It does so by identifying repositories of publicly available research publications on the Web which are dedicated mainly to the research field of computer science, and by automatically analyzing the full-text documents they provide in order to extract relevant metadata, including references. This way, *CiteSeer* retroactively indexes all the publications made available on the websites or repositories that it processes. As a result, the *CiteSeer* database is able to provide significant coverage even of those parts of the literature that predate its introduction, given the relatively long history of online pre-publications in computer science that it can rely on. More information about *CiteSeer* can be found in Bar-Ilan (2001), Goodrum et al. (2001), and Lawrence, Giles, & Bollacker (1999), as well as in Zhao & Strotmann (2004).

Since we find below that a comparable search of both *SCI* and *CiteSeer* returns a similar number of citing papers in the XML research field for the same time period, coverage of this field is therefore in all likelihood quite comparable between these two databases, both in terms of the proportion of the literature indexed and in terms of the publication time periods covered. We can therefore be quite confident that the main difference between these two citation databases with respect to the XML research field is indeed that between the publication venues they each cover, namely the Web and the journal, respectively.

As for whether research papers found on the Web and associated citation indexes such as *CiteSeer* can be used as valid data sources for citation analysis purposes despite their comparative lack of consistent quality control when compared with the journal articles indexed by the well-established ISI databases, the following reasons lead us to believe that they can.

Although it is generally accepted that the formal peer review process involved in the journal publication process often contributes significantly to the improved quality of research papers, the main thrust of a research paper tends to remain unchanged during this process, which is probably one reason why research papers found on the Web are perceived as a valuable source of information for research in some research fields (Youngen, 1997). As a result, the main part of the reference list of a preprint or a conference paper that appears on the Web with little delay is likely to remain largely the same even in a considerably reworked version of that paper finally published in a journal. In other words, the peer review process is not very likely to greatly affect the citing authors' view of a research field as seen from their citation patterns.

And indeed, Zhao (2005) found a very high correlation between author rankings by number of citations in the XML research field based on data from *SCI* and *CiteSeer* respectively, indicating that citation analysis studies using *CiteSeer* as a data source can be as valid as those using *SCI*, provided of course research fields studied are covered as well as the XML field by *CiteSeer*.

### *Data collection*

Three sets of source data were collected: all documents (along with the references they cite) as indexed under the terms "XML" or "eXtensible Markup Language" by (a) *CiteSeer* for the time period 2001 and earlier; (b) *SCI* for the same time period (i.e., years 2001 and earlier); and

(c) *SCI* for the years 2001 to 2006. For convenience and clarity, we will below refer to these two time periods as "the first time period" (i.e., years 2001 and earlier) and "the second time period" (i.e., years 2001 to 2006) respectively. These two time periods overlap somewhat (i.e., in the year 2001).

The actual searches for the first time period were conducted on December 18, 2001 for both *SCI* and *CiteSeer*. We did not specify a citation window at that time, indicating that publications from all years up to the date of the searches were used. However, the citation window corresponded in practice to roughly five years because the first phase of the World Wide Web Consortium (W3C)'s XML activity, which coined the terms we chose for searching for papers on XML (i.e., "XML" or "eXtensible Markup Language"), began in June 1996 (W3C, 2001). The search in the *SCI* database for the second time period (years 2001 to 2006) was conducted on February 24, 2006. All citing papers indexed by *SCI* under the terms "XML" or "eXtensible Markup Language" that had been published since 2001 were retrieved, for an actual citation window comparable in length to the first time period, i.e., roughly 5 years.

Despite their similar natures as citation indexes for scientific research publications, there are considerable differences between *SCI* and *CiteSeer* in the way they are designed and implemented. In order to obtain more comparable datasets from these two data sources, we took into account three considerations.

(a) Since *CiteSeer* only indexes research papers in the computer science field, we used citing papers from journals indexed by *SCI* in the "Computer Science" subject category rather than all citing papers from the entire *SCI* database. Since computer science is defined in *CiteSeer* in a very broad sense, and since the XML research field has broad applications in Computer Science, we used all of the seven sub-categories of Computer Science category in *SCI*: Artificial Intelligence, Cybernetics, Hardware & Architecture, Information Systems, Interdisciplinary Applications, Software Engineering, and Theory & Methods. Data collected this way should cover the core of XML research which belongs to the computer science field even though XML technology has a wide range of applications even outside Computer Science.

(b) The search for citing papers in *SCI* was limited to "topic" (TS) while the search in *CiteSeer* was limited to "header". We assumed that "header" fields in the *CiteSeer* database are sufficiently similar in scope to "topic" fields in *SCI* although the "header" fields in *CiteSeer*

include author names, of course, while the "topic" fields in *SCI* do not. Given our search terms, there was little danger of any author's name providing spurious matches in *CiteSeer*, since corporate authors were filtered out from both datasets for our study in order to focus on individual authors' contributions to the field. Because *SCI* only goes as far as abstracts in indexing citing papers, we did not use *CiteSeer*'s full text search facility for the present study despite its potential to support a wider view of the impacts of the XML research field.

(c) In our analysis of citation data provided by *CiteSeer* we restricted ourselves to counting only the first author of each cited paper despite the support provided by *CiteSeer* for counting all authors of a cited paper to potentially improve the view of the research field (Zhao, 2006a, 2006b). We did so because *SCI* does not directly provide the names of any but the first authors of cited papers, and for a large percentage of the cited papers it indexes, names of non-first authors can not be obtained even through matching cited papers with the *SCI* source papers. This is shown by Persson (2001) who found that in the Library and Information Science field, for example, only about one in ten cited references in the *Social Science Citation Index* (*SSCI*) were to papers indexed by ISI as source papers (and thus available for all-author citation counting). Since therefore only a minority of cited papers has sufficiently complete information available in the ISI databases, and since we are not aware of any method that would have allowed us to reliably automate the conversion of every brief *SCI* cited paper record into a full record outside of the ISI database, any attempt to remedy this problem would have involved considerable manual processing and would thus have been prohibitively expensive. However, as classical ACA has traditionally been using first-author-only citation counts anyway, this was an acceptable limitation for the study reported here as well, although future in-depth studies should attempt to avoid this problem in such a highly collaborative field.

Citing papers which met the searching criteria described above were retrieved along with their reference lists from the respective databases (i.e., *SCI* and *CiteSeer*) and downloaded into a local computer. Since the existence of duplicate citing papers was found to be one of the major differences between *SCI* and *CiteSeer* (Zhao & Logan, 2002), citing paper entries retrieved from *CiteSeer* were examined first by a Java program and then manually to remove duplicate citing papers and their associated reference lists from the dataset. Programs were then developed in Java to convert the data for the retrieved papers to a format that was convenient for subsequent

data analysis such as counting citations and co-citations. Details about the algorithms and data structure can be found in Zhao (2003).

This way we collected 268 citing papers from *SCI* and 312 from *CiteSeer* for the first time period, and 2,475 from *SCI* for the second time period. Since these two databases (i.e., *SCI* and *CiteSeer*) were comparable in terms of the number of non-duplicate citing papers retrieved for the first time period, and since our previous studies had found high correlations between author rankings by number of citations resulting from these two databases, we feel confident that *CiteSeer* represented the XML research area as published on the Web about as well as *SCI* represented the same research field as published in journals during that period. Given that most scientometric studies to date have relied on the ISI databases as their sole source of data on journal publications, we therefore feel justified in phrasing our comparisons in the following in terms of "Web publications" *vs*. "journal publications" when we are in reality comparing the contents of two citation databases.

Nevertheless, we do recognize that while *CiteSeer* may well represent a fairly random sample of computer science papers available on the Web, the ISI purposely indexes a non-random selection of "elite" journals. This tends to bias the ISI databases towards established research journals at the expense of those that are introduced to cover emerging fields, because such journals need time to pass through the ISI journal selection process. In our discussion below, we attempt to take into account the systematic time-delay in *SCI* that this bias causes for the recognition of an emerging research field like XML.

### *Data analysis*

We describe below the data analysis steps and techniques we employed within the ACA framework. Readers are referred to other studies (Kreuzman, 2001; McCain, 1986; McCain, 1990a; McCain, 1990b; White, 2003; White & Griffith, 1981; White & Griffith, 1982; White & McCain, 1998) for a detailed review of classic ACA techniques and their underlying rationale.

For each dataset we retrieved, a core set of authors was selected to represent the XML research field based on "citedness" – the number of citations they received. Citedness above some threshold is a good criterion for selecting authors to include in ACA although the resulting authors may not be "wholly definitive" of the research field being studied (White & McCain,

1998, p. 332). Three sets of highly-cited authors were selected in this manner from the three data sets respectively – the data set from *CiteSeer*, the one from *SCI* for the first time period, and the one from *SCI* for the second time period. There are no strict rules regarding thresholds for citation-based author selection in classical ACA studies (McCain, 1990b). Although the more authors studied the better a research field may be represented, the present study chose the 100 most highly cited individual authors from each of the three data sets to be included in the final multivariate analyses. This number is higher than in most previously published ACA studies and should suffice for our main purpose, which is to compare the specialty structures as revealed through the analysis of representative authors' oeuvres found on the Web and in journals.

In classic ACA, two authors are considered as being co-cited when at least one document from each author's oeuvre occurs in the same reference list, an author's oeuvre being defined as all the works with the author as the first author (McCain, 1990a). Based on this definition, a Java program was developed to determine co-citation frequencies of the highly-cited authors selected and to record them in co-citation matrices. As in McCain (1986), authors who were co-cited with very few other authors were deleted from these matrices. Specifically, an author was deleted if the corresponding row/column contained less than 5% non-zero value cells – a lower cutoff than the 33% cutoff that was used in McCain (1986). The lower threshold was used in order to allow a larger number of authors to be included in the study.

The resulting matrices were then converted to Pearson's *r* correlation matrices that were in turn used as input to the multivariate analysis procedure employed: Factor Analysis (FA). We used Factor Analysis here as a way to map the intellectual structure of a research field because it has been shown to provide clear and revealing results as to the nature of the discipline when applied in ACA (White & McCain, 1998). We chose Pearson's *r* as the similarity measure for this study as many ACA studies have done, and decided that the recent controversy around the use of Pearson's *r* correlations in ACA (Ahlgren, Jarneving, & Rousseau 2003; White, 2003) is beyond the scope of the present study whose focus is essentially the *comparison* of results from ACA between two data sources rather than a detailed analysis of the structure of a research field that those results themselves allow.

Factors were extracted by Principal Component Analysis (PCA) with an oblique rotation (SPSS Direct OBLIMIN). As in White & Griffith (1982), an oblique rotation was employed, for

two reasons. (a) It is often more appropriate than an orthogonal rotation when it can be expected theoretically that the resulting factors (in this case, specialties) would in reality be correlated (Hair et al., 1998). Especially in an emerging research field like XML, fairly high correlations between specialties are in fact expected. (b) The degree of correlation between the resulting factors is indicated by the component correlation matrix produced by an oblique rotation, which is clearly a very useful indicator in the analysis of the intellectual structure of a research field (McCain, 1990b; White & Griffith, 1982).

The number of factors extracted was determined based on Kaiser's rule of eigenvalue greater than 1 because the resulting model fit was adequate in all cases as indicated by total variance explained, communalities, and correlation residuals (Hair et al., 1998). We will present the numbers below when we discuss the results.

Although some previous ACA studies have used factor structure matrices (White & Griffith, 1982), the factor analysis results in the present study are presented and discussed below based on the factor pattern matrices provided by the Factor Analysis routine in SPSS. Hair et al. (1998) explain the differences between these two types of factor matrices as follows:

> When an oblique rotation has been used, two factor matrices of loadings are provided. The first is the factor pattern matrix, which has loadings that represent the unique contribution of each variable to the factor. The second is the factor structure matrix, which has simple correlations between variables and factors, but these loadings contain both the unique variance between variables and factors and the correlation among factors. As the correlation among factors becomes greater, it becomes more difficult to distinguish which variables load uniquely on each factor in the factor structure matrix. Most researchers report the results of the factor pattern matrix. (p. 113)

As factors in ACA are interpreted as scientific specialties (White & McCain, 1998), the correlations among factors in ACA can be expected to be fairly high in general. And indeed, we found quite high correlations between factors in the XML field in the present study (e.g., 0.65 between factors 1 and 3, and 0.47 between factors 2 and 5 in Table 2). We therefore concluded that the factor pattern matrix that represents the unique contribution of individual authors (variables) to specialties (factors) is more appropriate for ACA purposes in the present study than the factor structure matrix.

With the aid of these factor analysis results, we analyzed the grouping of scholars within each set of authors, and the specialties that these groups represent, and compared results from the three datasets.

## Findings

ACA allows a detailed analysis of the intellectual structure of a research field as well as its development over time as shown in White & McCain (1998) and in other studies (e.g., Culnan, 1987), and indeed the ACA results obtained in this study are very informative as shown in the tables presented in this article. However, we will leave a detailed analysis of the intellectual structure of the XML research field and its evolution over time to a separate article as the focus here is on identifying similarities and differences in the intellectual structure between Web-based and journal-based scholarly communication. We therefore emphasize here the overall picture of the XML field.

We will nevertheless examine first the perspective that the *SCI* database affords on the development of the XML research field from the first time period (i.e., the years until 2001) to the second time period (i.e., the years 2001 to 2006), by comparing its intellectual structures as revealed from *SCI* for these two time periods. We will then examine the intellectual structure found from *CiteSeer* during the first time period to see how it is related to that from *SCI* for each of these two time periods.

The intellectual structure of the XML research field will be discussed based on factor analysis results presented in Tables 1 (*SCI*, first time period), Table 2 (*SCI*, second time period), and Table 3 (*CiteSeer*, first time period).

The factors are assigned numbers and names as shown in the caption of each of the three tables. Their numbers appear as column headings in the table, while their names were assigned upon examining the frequently cited articles written by authors in the corresponding factors. Following White & McCain (1998)'s example, authors are listed in the factor on which they load most highly, and are ranked within each factor by the strength of their loadings in it. Loadings higher than 0.3 on additional factors, if any, are also presented, indicating the contributions of these authors to more than one specialty.

With large factors interpreted as specialties, the results of the factor analyses presented in these tables reveal the specialty structure of the XML research field and the associated authors' memberships in one or more specialties as judged by citing authors in the three datasets (White & McCain, 1998).

### *Evolving intellectual structure of XML research as revealed from SCI*

Using search terms "XML" or "eXtensible Markup Language", a "topic" search in *SCI* from journals indexed in its "computer science" category resulted in 268 citing papers with reference lists for the first time period, and 2,475 for the second time period, a nine-fold increase in just five years and a sign of the initial exponential growth typical of a successfully emerging discipline (De Solla Price, 1961). On average, each paper in these two sets cited about 18 or 19 references, respectively.

Table 1 presents the results of a factor analysis of the 100 most highly cited XML researchers found in *SCI* during the first time period, and Table 2 results of an identical analysis for the second time period. Kaiser's rule of eigenvalue greater than one resulted in an eight-factor model for the first time period which explains 94.6% of the total variance, and a nine-factor model for the second time period which explains 96.3% of the total variance. In both cases, the differences between observed and implied correlations were smaller than 0.05 for the most part (almost 100%). Clearly, the model fit was good in both cases.

As mentioned earlier, the factors in these two tables were named based on an examination of the titles and abstracts of the frequently cited articles written by authors in the corresponding factors, based on our expertise in the XML research field. It can be seen that some identified factors are highly coherent groups, some are less coherent, and others pick up some interesting isolates.

Major specialties that we identified from *SCI* for the first time period as shown in Table 1 include (1) *Management of semi-structured or XML databases*, (2) *Web standards, specifications and guidelines*, (3) *Intelligent Web service management and integration*, (4) *Foundations*, e.g., formal languages, (5) *XML for medical decision support*, and (6) *XML for medical data exchange*. Small and less coherent groups include (a) *The Semantic Web*, (b) *Intelligent software agents on the Web*, and (c) *Hypermedia*.

**Table 1: Factor Analysis of 100 authors in the XML research field** (*SCI computer science <=2001*)

**1**. XML databases; **2**. Web standards, specifications & guidelines; **3**. Intelligent Web service management & integration; **4**. Foundations (formal languages); **5**. XML for medical decision support; **6**. Intelligent software agents on the Web; **7**. The Semantic Web; **8**. XML for medical data exchange / Hypermedia

| Authors | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A. Salminen | 0.85 | | | | | | | |
| A. Bonifati | 0.85 | | | | | | | |
| W. B. Frakes | 0.80 | | | | | | | |
| S. Ceri | 0.75 | | | | | | | |
| D. Maier | 0.74 | | | | | | | |
| D. Chamberlin | 0.73 | | | | | | | |
| R. Goldman | 0.73 | | | | | | | |
| J. Robie | 0.73 | | | | | | | |
| N. Ide | 0.72 | 0.34 | | | | | | -0.30 |
| D. Florescu | 0.72 | | | | | | | |
| J.Shanmugasundaram | 0.70 | | | | | | | |
| B. Ludascher | 0.69 | | | | | | | |
| A. Deutsch | 0.69 | | | | | | | |
| D. Beech | 0.68 | | | | | | | |
| R. Kimball | 0.68 | | | | | | | |
| J. Mchugh | 0.67 | | | | | | | |
| K. Bohm | 0.66 | | | | | | | |
| M. Kifer | 0.62 | | | -0.31 | | | | |
| P. PS. Chen | 0.62 | | | 0.38 | -0.32 | | -0.36 | |
| M. Fernandez | 0.61 | | | | | | | |
| S. Abiteboul | 0.57 | | | | | | | |
| C. Beeri | 0.57 | | | -0.44 | | | | |
| S. Chawathe | 0.56 | | | | | | -0.47 | |
| P. Buneman | 0.55 | | | -0.30 | | | | |
| D. Suciu | 0.55 | | | | | | | |
| D. Calvanese | 0.54 | | | -0.41 | | | | |
| J. Widom | 0.53 | | -0.39 | | | | | |
| H. Thompson | 0.52 | 0.33 | | | | | 0.33 | |
| V. Christophides | 0.52 | | | -0.33 | | | | |
| F. Bancilhon | 0.52 | | | -0.36 | | | | |
| S. J. Derose | 0.51 | 0.34 | | 0.32 | | | | -0.47 |
| S. Cluet | 0.50 | | | -0.37 | | | | |
| R. GG. Cattell | 0.49 | | | | | | | |
| T. Milo | 0.49 | | | -0.49 | | | | |
| D. Lee | 0.49 | | -0.42 | | -0.30 | | | |
| Y.Papakonstantinou | 0.49 | | -0.31 | | | | | |
| M. Murata | 0.48 | | -0.35 | -0.33 | | | | |
| G. Wiederhold | 0.44 | | | | | | | |
| D. Dubois | 0.41 | | -0.36 | | | | | |
| A. Gupta | 0.34 | | | | | | -0.30 | |
| D. Knuth | | 0.89 | | -0.42 | | | | |
| P. Ciancarini | | 0.65 | | | | | | |
| C. F. Goldfarb | 0.50 | 0.63 | | | | | | |
| V. Apparao | | 0.60 | | | | | 0.31 | |
| E. Maler | | 0.59 | | | | | | |
| T. Bray | | 0.59 | | | | | | |
| B. Bos | | 0.51 | | | | | 0.31 | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| C. M. Sperberg-McQueen | | 0.50 | | | | | | |
| J. Bosak | | 0.47 | | | | | 0.35 | |
| J. Clark | | 0.47 | | | | | 0.36 | |
| A. Layman | -0.32 | 0.47 | | | | | 0.40 | |
| P. Murray-Rust | | 0.41 | | | | | | |
| L. Wood | | 0.40 | | | | | | -0.34 |
| R. Khare | | 0.35 | | | | | 0.32 | |
| E. R. Harold | | 0.34 | | | | | 0.32 | |
| C. Knoblock | | | -1.02 | | | | | |
| D. Raggett | | | -0.96 | | | | | |
| P. Atzeni | | | -0.95 | | | | | |
| A. Sahuguet | | | -0.95 | | | | | |
| D. Konopnicki | | | -0.91 | | | | | |
| N. Kushmerick | | | -0.91 | | | | | |
| A. O. Mendelzon | | | -0.90 | | | | | |
| G. Arocena | | | -0.75 | | | | | |
| B. Adelberg | | | -0.66 | | | | | |
| H. Garcia-Molina | | | -0.65 | | | | | |
| A. Dogac | | | -0.60 | | | | | |
| P. A. Bernstein | 0.38 | | -0.56 | | | | | |
| A. Levy | 0.31 | | -0.53 | | | | | |
| L. Liu | | | -0.51 | | | | -0.31 | |
| A. Hunter | | | -0.46 | -0.35 | | | | |
| D. Fensel | | -0.32 | -0.43 | | | -0.34 | | |
| D. Harel | | -0.31 | -0.40 | -0.37 | | -0.31 | | |
| D. Remy | | | | -0.98 | | | | |
| A. Bruggemann-Klein | | 0.35 | | -0.85 | | | | |
| L. Cardelli | | | | -0.80 | | | | |
| A. Ohori | | | | -0.78 | | | | |
| A. Aiken | 0.34 | | | -0.68 | | | | |
| H. Hosoya | 0.41 | | | -0.62 | | | | |
| F. Neven | 0.44 | | | -0.60 | | | | |
| P. Wadler | 0.48 | | | -0.51 | | | | |
| D. F. Lobach | | | | | 0.84 | | | |
| R. N. Shiffman | | | | | 0.83 | | | |
| L. Ohnomachado | | | | | 0.83 | | | |
| G. Hripcsak | | | | | 0.59 | | | 0.37 |
| D. Connolly | -0.36 | | | | -0.40 | 0.32 | | |
| M. Wooldridge | | | | | | -1.00 | | |
| D. A. Benson | | | | | | | -0.80 | |
| S. B. Davidson | | | | | | | -0.75 | |
| O. Lassila | | | | 0.30 | | | 0.55 | |
| S. Decker | 0.52 | -0.3 | | | -0.32 | -0.48 | 0.54 | |
| D. Brickley | 0.34 | 0.36 | | 0.35 | | | 0.36 | -0.30 |
| S. B. Johnson | | | | | | | | 0.76 |
| L. Alschuler | | | | | | | | 0.74 |
| C. Friedman | | | | | | | | 0.72 |
| J. J. Cimino | | | 0.34 | 0.32 | -0.30 | | | 0.65 |
| R. H. Dolin | | | | | 0.36 | | | 0.60 |
| S. Chakrabarti | | | | | | 0.44 | | -0.63 |
| K. Gronbaek | | | | | | 0.40 | | -0.58 |
| F. Halasz | | | | | | 0.31 | | -0.40 |
| T. Berners-Lee | | 0.31 | | | | | | -0.39 |

Note: Data in this table was also used in Zhao (2004) for different purposes.

**Table 2: Factor Analysis of 100 authors in the XML research field** (*SCI computer science 2001-2006*)

**1**. Semi- structured or XML databases; **2**. XML database Theory; **3**. Information retrieval from XML documents and databases; **4**. XML processing theory; **5**. XML data engineering for data-driven websites; **6**. Core XML standards; **7**. Access control; **8**. The Semantic Web; **9**. Intelligent integration

| Authors | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| M. Yoshikawa | -0.86 | | | | | | | | |
| A. R. Schmidt | -0.81 | | | | | | | | |
| I. Tatarinov | -0.79 | | | | | | | | |
| T. Grust | -0.78 | | | | | | | | |
| P. Bohannon | -0.77 | | | | | | | | |
| C. Zhang | -0.76 | | | | | | | | |
| Q. Li | -0.75 | | | | | | | | |
| H. Jiang | -0.74 | | | | | | | | |
| Y. M. Chen | -0.74 | 0.43 | | | | | | | |
| S. Al Khalifa | -0.73 | | | | | | | | |
| N. Bruno | -0.71 | | | | | | | | |
| H. V. Jagadish | -0.71 | | | | | | | | |
| S. Y. Chien | -0.71 | | | | | | | | |
| B. Cooper | -0.68 | | -0.35 | | | | | | |
| J. Shanmugasundaram | -0.67 | | | | | | | | |
| D. Florescu | -0.67 | | | | | | | | |
| R. Kaushik | -0.62 | | -0.35 | | | | | | |
| I. Manolescu | -0.60 | | | | 0.32 | | | | |
| M. Carey | -0.60 | | | | | | | | |
| R. Goldman | -0.57 | | | | | | | | |
| J. McHugh | -0.54 | | | | | | | | |
| M. Fernandez | -0.41 | | | | | | | | |
| V. Christophides | -0.41 | | | | 0.32 | | | | |
| A. Deutsch | -0.40 | | | | | | | | |
| A. Sahuguet | -0.40 | | | | | | | | |
| S. Abiteboul | -0.35 | | | | | | | | |
| D. Chamberlin | -0.33 | | | | | | | | |
| M. W. Vincent | | 1.08 | | | | | | | |
| M. Arenas | | 0.97 | | | | | | | |
| W. Fan | | 0.80 | | | | | | | |
| J. Widom | | 0.70 | | | | | | | |
| S. Chawathe | | 0.47 | | | | | | | 0.33 |
| P. Buneman | | 0.39 | | -0.31 | | | | | |
| T. Schlieder | | | -1.05 | | | | | | |
| N. Fuhr | | | -0.94 | | | | | | |
| R. Baeza-Yates | | | -0.85 | | | | | | |
| G. Salton | | | -0.79 | | | | | | |
| S. Ameryahia | -0.36 | | -0.66 | | | | | | |
| A. Bonifati | | | -0.45 | | 0.36 | | | | |
| J. Robie | | | -0.37 | | 0.31 | -0.32 | | | |
| J. Engelfriet | | | | -0.91 | | | | | |
| A. Bruggemann-Klein | | | | -0.90 | | | | | |
| A. V. Aho | | | | -0.80 | | | | | |
| F. Neven | | | | -0.80 | | | | | |
| M. Murata | | | | -0.76 | | -0.32 | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| H. Hosoya | | | | -0.71 | | | | | |
| J. E. Hopcroft | | | | -0.67 | | | | | |
| G. Gottlob | | | -0.33 | -0.66 | | | | | |
| P. Wadler | | | | -0.61 | | | | | |
| F. Bry | | 0.43 | | -0.58 | | | | | |
| D. Suciu | | | | -0.54 | | | | | |
| Y. Papakonstantinou | | | | -0.52 | | | | | |
| D. Calvanese | | | | -0.51 | | | | | |
| C. Beeri | | 0.41 | | -0.49 | | | | | |
| T. Milo | -0.38 | | | -0.46 | | | | | |
| S. Cluet | | | | -0.43 | | | | | |
| P. Fankhauser | | | | -0.33 | | | | | |
| K. Zhang | 0.34 | 0.38 | | 0.51 | 0.33 | | | -0.32 | 0.39 |
| Y. Wang | | 0.32 | | 0.49 | 0.35 | | | -0.43 | 0.31 |
| M. Altinel | | | | | 0.53 | | | -0.38 | |
| A. Gupta | | | | | 0.51 | | | | 0.31 |
| L. Liu | 0.31 | 0.41 | | | 0.48 | | | | |
| C. Chan | | | | | 0.48 | | | -0.39 | |
| S. Ceri | | | -0.34 | | 0.41 | | | | |
| B. Ludascher | | | | -0.31 | 0.40 | | | | |
| R. Cattell | | | | | 0.30 | | | | |
| J. R. Ullman | | | | | 0.30 | | | | |
| P. V. Biron | | | | | | -0.87 | | | |
| E. Gamma | | | | | | -0.79 | | 0.33 | |
| D. C. Fallside | | | | | | -0.78 | | | |
| H. Thompson | | | | | | -0.72 | | | |
| M. Kay | | | | | | -0.69 | | | |
| S. Derose | | | -0.37 | | | -0.67 | | | |
| R. Bourret | -0.47 | | | | | -0.67 | | | |
| J. Clark | | | | | | -0.58 | | | |
| T. Bray | | | | | | -0.55 | | | |
| D. Lee | -0.31 | 0.31 | | | | -0.43 | | | |
| S. Boag | -0.40 | | -0.30 | | | -0.42 | | | |
| E. Damiani | | | | | | | -1.04 | | |
| E. Bertino | | | | | | | -1.02 | | |
| R. H. Dolin | | | | | 0.37 | 0.35 | -0.72 | | |
| R. Agrawal | -0.33 | 0.40 | -0.40 | | | | -0.46 | | |
| S. Decker | | | | | | | | 0.99 | |
| I. Horrocks | | | | | | | | 0.94 | |
| T. R. Gruber | | | | | | | | 0.91 | |
| D. Fensel | | | | | | | | 0.89 | |
| D. Brickley | | | | | | | | 0.78 | |
| H. Zhuge | | | | | | | | 0.76 | |
| T. Berners-Lee | | | | | | -0.34 | | 0.74 | |
| O. Lassila | | | | | | -0.35 | | 0.74 | |
| I. Foster | | | | | | | | 0.67 | |
| J. Lee | | | | 0.40 | | -0.52 | | 0.58 | |
| W. MP. Van der Aalst | | | | | | | | 0.54 | |
| G. Cabri | | | 0.40 | | | | | 0.50 | 0.31 |
| E. Rahm | | | | | | | | | -0.97 |
| A. Doan | | | | | | | | | -0.95 |
| G. Wiederhold | | | | | 0.37 | | | 0.35 | -0.51 |
| H. Garcia-Molina | | | | -0.31 | 0.35 | | | | -0.40 |
| A. Y. Levy | | | | -0.31 | 0.35 | | | | -0.38 |
| P. A. Bernstein | | | | | | -0.33 | | | -0.37 |

In Table 2, major specialties identified from *SCI* for the second time period include (1) *XML & relational databases,* (2) *XML database theory,* e.g., normal forms or constraints, (3) *Information retrieval from XML documents and databases*, (4) *XML processing theory,* e.g., formal languages, tree walkers, or regular expressions, (5) *Core XML standards,* e.g., XML Schema, XSLT, web services, and design patterns (6) *The Semantic Web,* (7) *Intelligent mediation between XML databases,* i.e., database mediation with an emphasis on intelligent translation and integration. A small but distinct group, *Access control,* is also identified.

We found it difficult to label the small group in Table 2 that has low loadings from all authors in the group (factor 5). Two major themes in the research of this group of authors appear to be XML data engineering and data-driven websites. We thus give it a tentative label as shown in the table.

**Major changes from first to second time period**

As one would expect, after a five year jump in time from the first to the second time period in the emerging XML research field, some new specialties have appeared while some old ones disappeared; some specialties have shifted focus while others split or merged.

A major development in the XML research field during this time period sees the *Semantic Web* specialty growing from a very weak group (as indicated by the low loadings of just three members) to a large and distinct specialty. Nine current members of this specialty load significantly higher than any of its three members during the first period.

Specialties that have newly emerged during this time include (a) *XML database theory*, (b) *Information retrieval for XML documents*, and (c) *Access control*. These three areas are distinguished by the emergence of authors who load highly on these respective specialties, such as Vincent and Arenas for the *XML database theory* specialty, Schlieder, Fuhr, and Baeza-Yates for the *Information retrieval for XML documents* specialty, and Damiani and Bertino for the *Access control* group. Works in these areas by several authors previously lumped in the *XML data management* specialty have also been recognized, and these authors have therefore been regrouped into these emerging specialties. As an example, long-time highly cited authors Widom, Chawathe, and Buneman have been placed into the *XML database theory* specialty, and Bonifati and Robie into the *Information retrieval for XML documents and databases* specialty.

A few previously recognized specialties have essentially disappeared, including the *Intelligent Web service management* group, the *Intelligent software agents on the Web* group, and both of the specialties on applications of XML in the medical field. The very few remaining highly-cited authors from these four specialties are now perceived as belonging to some of the newly emerging specialties, which in most cases appears to be a more appropriate acknowledgement of their more recent research. For example, Fensel is now found in the *Semantic Web* specialty, and Dolin from the *XML for medical data exchange* group has relocated to the *Access control* group, a fundamental issue in medical data management and exchange and therefore a logical new focus for him within the overall XML research area.

Authors who study data integration have been recognized in results from both time periods, but were joined by authors with slightly different research foci in each period – those studying intelligent Web services during the first time period, and those interested in intelligent mediation between XML databases during the second. This shows a tightening of the focus away from the practical applications (web services) to the underlying theoretical challenges of data integration and its efficient automation (intelligent mediation).

While its size has remained essentially unchanged, the research within the *XML standards* group has reorganized itself in the view of the citing authors in this field from covering standards and guidelines in a very general sense to a clear focus on core standards for processing XML data, such as XML Schema and eXtensible Stylesheet Language – Transformations (XSLT).

More than half of the authors previously in the *XML data management* specialty have moved on or lost their ranking among the 100 most highly cited authors in their field. While some joined the newly emerging specialties discussed above, including the specialty *XML database theory*, many of them (e.g., Beer, Suciu and Murata) have split off into the specialty labeled *XML processing theory*. There they have been joined by some authors with very high loadings on this specialty who only made the cut to the top 100 authors during the second time period, such as Engelfriet and Aho. This large and distinct specialty has apparently formed from the specialty labeled *Foundations* during the first time period, which at that time already included Bruggemann-Klein, Hosoya, Neven, and Wadler, and which has now recruited new members from other sub-fields.

The remaining authors in this previously dominant *XML data management* specialty have been joined by other scholars with a research focus on relational XML databases and their efficient implementation, to form a group with an overall focus on the applied aspects of research on XML data management.

Research on *XML data management* has thus apparently split into two groups, one focusing on theoretical and the other on applied aspects of that field. The theoretical aspect of XML research has further divided into sub-specializations on data retrieval and data processing, respectively.

It appears that the XML research field has been evolving as follows during its first decade: In its early formative stages, the XML research field featured wide-ranging exploration of interesting problems, frequently derived from early experimentation with potential practical applications. As the field matured over time, experience gained from early experimentation appears to have identified common underlying problems that required deeper research, and, as a direct consequence, the field appears to have focused more on its theoretical aspects.

**Major trends over a decade of XML research**

We have thus identified several major trends in the evolution of the XML research field from the first to the second time period as found in journals indexed as Computer Science research by *SCI*, each trend exemplified by a number of authors included in the analysis:

(a) Research on *the Semantic Web* has strengthened considerably.

(b) Research on applications of XML outside of computer science (e.g., in the medical field) has essentially disappeared.

(c) Research on access control has been clearly recognized.

(d) Research on *information retrieval for XML documents* has become a focus.

(e) Research on theoretical aspects of XML data management and processing has become pronounced.

(f) Research on data integration has shifted to show an emphasis on intelligent mediation and translation between XML databases.

Interestingly, as we will see below, most of these trends were already apparent in the results of an ACA based on computer science publications on the Web as indexed by *CiteSeer* during the first time period.

### *Relationship between Intellectual structures of XML research revealed from CiteSeer and SCI*

For the first time period (i.e., the roughly five years until 2001), a "header" search in *CiteSeer* using the search terms "XML" or "eXtensible Markup Language" resulted in 312 papers after removing duplicates, producing a dataset comparable in size with that from *SCI* for the same time period (268). The average length of reference lists in these papers is, however, significantly smaller than that in *SCI* for the same time period: about 15 *vs.* about 18 references per paper. Apparently, papers publicly available on the Web have made 20% fewer citations on average than publications in journals in the same field and during the same time period. This difference may be the result of several factors: (a) journal articles are generally more carefully constructed arguments requiring more references; (b) conference papers, one of the major types of research papers found on the Web, often have tighter length restrictions; (c) "courtesy citations," citations made based on considerations other than intellectual relevance, arguably occur more frequently in journals.

Considering that the absolute number of citing papers retrieved simultaneously from the two databases covering the two publishing media was very similar, the overlap of citing papers between the two publishing media is surprisingly small: only 26 (or about 8%) of the citing papers from *CiteSeer* for this period overlap with the citing papers from *SCI* for the same time period, while 37 (or about 12%) of them overlap with those from *SCI* for the second time period. This indicates that in the XML research field, papers published in journals were not largely available publicly on the Web, and papers published openly on the Web were not well represented in *SCI*. Nevertheless, the fact that more than half of the 100 most highly cited authors are common to these datasets indicates a common perspective on the research field as viewed by these different sets of citing papers.

We can unfortunately only speculate about why the overlap here is so small. One obvious factor is that *CiteSeer* goes beyond journal articles and conference papers, i.e., those document types that are covered by the ISI databases in the Computer Science research field, when indexing research papers. In addition, since there is no equivalent of the ISI journal selection process in *CiteSeer*, the range of journals or conference proceedings it covers is naturally broader than that of *SCI*. The differing degrees to which different journals or conference proceedings demand or enforce exclusive copyright to their publications could amplify this difference between the two databases. Authors may be more willing to "risk" putting up publications on the web if they subsequently get formally published in journals with fewer copyright restrictions. In this case, one might suspect that the more "prestigious" venues indexed by the ISI databases may well be stricter in this respect than "lesser" publication venues that have not "earned" the ISI stamp of approval, but this supposition remains to be tested in future research.

It is perhaps interesting to note that the overlap between *CiteSeer* and *SCI* for the first time period is significantly smaller (by 11 papers out of 37) than that between *CiteSeer* for the first period and *SCI* for the second period. In other words, the dataset from *CiteSeer* for the first time period appears to be more similar to that from *SCI* for the second time period than to that from *SCI* for the first time period (although it is still very different from both). This already provides us with a first small piece of evidence for our main finding in this paper, namely, that, collectively, research reported on the Web on the one hand and that published in journals on the other hand appear to represent different stages of the development of the XML research field, with publications on the Web a few years more current.

Table 3 shows the results of a factor analysis of the 100 most highly cited XML researchers found in *CiteSeer* during the first time period. Kaiser's rule of eigenvalue greater than one resulted in an eleven-factor model which accounted for 96% of the total variance. The differences between observed and implied correlations were for the most part (almost 100%) smaller than 0.05. These indicate a good model fit.

**Table 3: Factor Analysis of 100 authors in the XML research field (*CiteSeer <=2001*)**

**1.** XML or semi-structured databases; **2.** Foundations of XML data management; **3.** The Semantic Web;

**4.** Programming for / processing of XML data; **5.** Natural language processing; **6.** Version management;

**7.** Functional and logic programming.; **8.** Database and information retrieval foundations;

**9.** Knowledge management; **10.** Access control; **11.** Data integration

| Authors | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L. Fegaras | 0.84 | | | | | | | | | | |
| S. Adler | 0.79 | | | | | | | | | | |
| J. F. Naughton | 0.74 | | | | | | | | | | |
| P. Atzeni | 0.61 | | | | | | | 0.34 | | | |
| D. Maier | 0.56 | | | | | | | | | | |
| D.D.Chamberlin | 0.56 | | | | | | | | | | |
| R. Cattell | 0.55 | | | | | | | | | | |
| M. J. Carey | 0.53 | | | | | | | | 0.32 | | |
| D. Beech | 0.48 | | | | | | | | | | |
| F. Bancilhon | 0.45 | 0.43 | | | | | | | | | 0.31 |
| J. Shanmugasundaram | 0.43 | | | | | | | | | | |
| V. Christophides | 0.42 | | | | | | | | | | |
| J. Widom | 0.41 | | | | | | | | | | |
| J. Miller | 0.40 | 0.36 | | | | | | 0.37 | -0.35 | | |
| S. Cluet | 0.39 | 0.38 | | | | | | | | | |
| A. Y. Levy | 0.38 | | | | | | | | | | |
| D. Florescu | 0.36 | | | | | | | | | | |
| J. McHugh | 0.36 | | | | | | | | | | |
| S. Abiteboul | 0.35 | 0.31 | | | | | | | | | |
| A. Deutsch | 0.34 | | | | | | | 0.33 | | | |
| M. Fernandez | 0.34 | | | | | | | | | | |
| R. Goldman | 0.33 | | | | | | | | | | |
| C. Baru | | 1.03 | | | | | | | | | |
| S. Cosmadakis | | 0.99 | | | | | | | | | |
| F. Neven | | 0.97 | | | | | | | | | |
| W. Fan | | 0.92 | | | | | | | | | |
| R. Ramakrishnan | | 0.90 | | | | | | | | | |
| D. Calvanese | | 0.89 | | | | | | | | | |
| V. Apparao | | 0.88 | | | | | | | | | |
| J. Ullman | | 0.84 | | | | | | | | | |
| P. Wadler | | 0.84 | | | | | | | | | |
| H. Thompson | | 0.74 | | | | | | | | | |
| T. Bray | | 0.72 | | | | | | | | | |
| C. Beeri | 0.30 | 0.67 | | | | | | | | | |
| J. E. Hopcroft | | 0.67 | | -0.43 | | | | | | | |
| J. Clark | | 0.64 | | | | | | | | | |
| H. Hosoya | | 0.61 | | | | | | | | | |
| P. Fankhauser | 0.32 | 0.57 | | | | | | 0.30 | | | |
| P. Buneman | | 0.56 | | | | | | | | | |
| A. Davidson | | 0.53 | 0.31 | -0.39 | | | | | | | |
| A. Sahuguet | | 0.49 | | -0.44 | | | | | 0.32 | | |
| L. Cardelli | | 0.42 | | | -0.37 | | | | | | |
| Y. Papakonstantinou | | 0.40 | | | | | | | | | |
| T. Milo | | 0.38 | | | | | | | | | |
| D. Fensel | | | 0.92 | | | | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D. Brickley | | | 0.92 | | | | | | | | |
| O. Lassila | | | 0.91 | | | | | | | | |
| I. Horrocks | | | 0.85 | | | | | | | | |
| T. Berners-Lee | | | 0.84 | | | | | | | | |
| P. Biron | | | 0.75 | -0.31 | | | | | | | |
| S. Decker | | | 0.75 | | | | 0.37 | | | | |
| D. Megginson | | | | -0.84 | | | | | | | |
| D. Lee | | | | -0.74 | | | | | | | |
| A. Aho | | | | -0.70 | | | | | | | |
| N. Klarlund | | 0.34 | | -0.67 | | | | | | | |
| M. Murata | | 0.42 | | -0.64 | | | | | | | |
| R. Bourret | | | | -0.62 | | | | | | | |
| E. Maler | | | 0.38 | -0.61 | | | | | | | |
| D. Fallside | | | | -0.51 | | | | | | | |
| C. C. Kanne | 0.34 | | | -0.44 | | | | | | | |
| L. Wood | | | | -0.43 | 0.43 | | | | | | |
| A. Schmidt | 0.34 | | | -0.41 | | | | | 0.39 | | |
| J. Bosak | 0.35 | | | -0.37 | | | | | | | |
| H. Jagadish | | | | | 0.82 | | | | 0.45 | | |
| M. Kay | | | 0.34 | | 0.74 | | | | | | |
| N. Walsh | | | | | 0.67 | | | | | | |
| J. K. Ousterhout | | | | | 0.66 | | | | | | |
| C. Barras | | | | | 0.63 | | | | | | |
| D. McKelvie | | | | | 0.36 | | | | | | |
| A. Albano | 0.30 | | | | -0.33 | | | | | | |
| G. Ghelli | 0.30 | | | | -0.33 | | | | | | |
| L. Liu | | | | | | 0.95 | | | | | |
| A. Marian | | | | | | 0.94 | | | | | |
| S. Y. Chien | | | | | | 0.88 | | -0.35 | | | |
| J. Chen | 0.34 | 0.30 | | | | 0.60 | | | | | |
| S. S. Chawathe | | | | | | 0.36 | | | | | |
| E. Harold | | | | | | | 0.91 | | | | |
| H. Boley | | | | | | | 0.72 | | | | |
| M. Hanus | | | | | | | 0.69 | | | | |
| C. Goldfarb | | | | | | | | 0.84 | | | |
| H. Meuss | | | | | | | | 0.76 | | | |
| M. P. Marcus | | | | | | | | 0.75 | | | |
| G. Navarro | | | | | | | | 0.74 | | | |
| R. Baeza-Yates | | | | | | | | 0.59 | | | |
| E. Baralis | | | | | | | | 0.57 | | | |
| J. Paredaens | | | | | | | | 0.53 | | | |
| A. Bonifati | | | | | | | | 0.51 | | | |
| S. Ceri | | | | | | | | 0.45 | | | |
| A. Aiken | | | | | | | | 0.45 | | | |
| H. Liefke | | | | | | | | 0.44 | | | |
| S. DeRose | | | | -0.30 | | | | 0.44 | | | |
| J. Robie | | 0.33 | | | | | | 0.38 | | | |
| C. Freitag | | | | | | | | | 0.81 | | |
| P. McBrien | 0.32 | | | | | | | | 0.51 | | |
| E. Bertino | | | | | | | | | | 1.03 | |
| E. Damiani | | | | | | | | | | 0.90 | |
| A. Gupta | | | | | | | | | | | 1.06 |
| M. Kifer | 0.31 | | | | | | | | | | 0.57 |
| B. Ludascher | | | | | | | | | | | 0.52 |
| S. Nestorov | | | | | | | | | | | 0.34 |

Note: Data in this table was also used in Zhao (2004), Zhao (2005b), and Zhao (2006a) for different purposes.

As in other tables, the factor names in Table 3 were given based on an examination of the titles and abstracts of the frequently cited articles written by authors in the corresponding factors, based on our expertise in the XML research field. As shown in this table, major specialties identified from *CiteSeer* for this first time period include: (1) Management of semi-structured or XML databases; (2) Foundations of XML data management & processing; (3) The Semantic Web; (4) Programming for and processing of XML data; (5) Natural language processing; (6) Version management; and, (7) Database (DB) and Information Retrieval (IR) foundations. Smaller groups include (a) Functional and Logic Programming; (b) Knowledge management; (c) Access control; and (d) Data integration.

Earlier in this paper, we identified several major trends in the evolution of the XML research field from the first to the second time period by comparing the ACA results from *SCI* between the two time periods. Interestingly, most of these trends were already apparent in the ACA results from *CiteSeer* for the first time period when compared with *SCI* for the same first time period. We discuss this in the order of the major trends identified in the section above.

The very weak *Semantic Web* group identified from *SCI* for the first time period was already apparent as a very distinct specialty in results from *CiteSeer* for the same first time period as indicated by its size (more than double the size of the corresponding group in *SCI* during this period) and by the much higher author loadings in this group. In the results from *SCI* for the first time period, Fensel was placed into the group *Intelligent Web service management and integration* while Berners-Lee appeared in the group *Hypermedia.* Here in the results from *CiteSeer*, however, both Berners-Lee and Fensel are grouped into the *Semantic Web* specialty just as they were in results from *SCI* for the second time period, indicating a greater emphasis on their more recent research foci. In other words, we appear to see a larger time-lag in *SCI* than in *CiteSeer* in acknowledging the shift of these authors away from earlier research foci to more recent ones.

The two groups representing XML applications in the medical sciences, which appeared in *SCI* during the first time period, did not show up at all in *CiteSeer* at this time. These applications had also disappeared from the *SCI* results in the second time period.

Similarly, research on *Access control* was clearly recognized both in results from *CiteSeer* during the first time period and in results from *SCI* during the second time period, but did not appear at all in the results from *SCI* during the first time period. In both cases when this area of research was recognized, Bertino and Damiani loaded very highly in this factor, indicating that they are the two most representative authors of this area of research.

Research on *Information retrieval for XML documents* did not appear in *SCI* during the first time period, but was already recognized in *CiteSeer* during the same first time period as shown in the *Database and information retrieval foundations* specialty. This specialty here contains three of the seven authors (Baeza-Yates, Bonifati, and Robie) of the *Information retrieval for XML documents* specialty identified from *SCI* during the second time period. In other words, research on information retrieval was already recognized on the Web during the first time period, and became visible in journals as a clear research focus in the XML research field only during the second time period.

Although research on XML data management was identified from both *SCI* and *CiteSeer* during the first time period as the most active research area in the XML research field, as indicated by the high percentage (about 40%) of the 100 most highly cited authors being placed in this general research area, authors working in this area formed a single group in the results from *SCI*, but split into two in those from *CiteSeer*. These two areas, *XML databases* and *Foundations of XML data management and processing*, corresponded to the practical and theoretical aspects, respectively, of XML data management.

This was also observed in our earlier studies (Zhao & Logan, 2002; Zhao, 2004) on the comparison of intellectual structures revealed from the Web vs. journals for the same time period. There we offered one possible explanation of this observation: the intellectual difference between the two groups seen from the Web was blurred or weakened in journals by factors such as “diplomatic citing” (Edge, 1979, p. 120) or “courtesy citations” which would arguably happen more frequently in journals, with the result that the two database groups could be pulled together there if the distinctions between them were not very clear to start with, as indicated by fairly low loadings of authors and a number of co-loading authors in the two separate groups identified from *CiteSeer*.

Now that we can take into account the subsequent development of XML research, we see another possible explanation for this phenomenon: the distinction between theoretical and applied aspects of research on *XML data management* was already becoming visible on the Web whereas research in journals was still showing a more exploratory investigation of problems of practical interest with less of a clear focus on underlying theoretical issues.

This alternative explanation appears to be supported by the following three observations.

(a) All authors studying *XML data management* were grouped into a single factor in *SCI* during the first time period. Low loadings of all authors in this factor identify it as a general research topic rather than a tightly focused group.

(b) Many authors in this single general group identified from *SCI* during the first time period later moved to the *XML processing theory* specialty identified from *SCI* during the second time period. This move contributed about one half of the authors of this theoretically focused specialty.

(c) The *Foundations of XML data management and processing* group in *CiteSeer* was a mix of authors studying XML database theory (e.g., Fan and Buneman), XML data processing theory (e.g., Neven and Calvanese), or foundational standards on XML data processing (Thompson and Bray). A distinction between theoretical and practical aspects of research on XML data management was thus visible in the results from *CiteSeer* for this first time period, although it was not yet as clearly defined as in *SCI* five years later.

In other words, a trend that has become very clear in *SCI* for the second time period, namely, research on theoretical aspects of XML data management becoming more pronounced, was already apparent in *CiteSeer* during the first time period, even if not as clearly.

To summarize, most of the trends in the evolution of the XML research field from the first to the second time period that we were able to identify when comparing ACA results from *SCI* between the two time periods, were already apparent to some degree in the ACA results from *CiteSeer* for the first time period. These trends were visible in quite some detail, as the trends were not just identified via a few factor labels or subfield descriptions, but more importantly through the many individual authors that constituted these trends becoming reclassified in the factor analyses as their research foci changed over time.

Thus, we conclude that *CiteSeer* quite likely provided us with a more current picture of research areas in the XML research field than *SCI* did using data from the same time period. This finding constitutes evidence that research reported on the Web during the first time period represented a more current stage of XML research at the time of data collection than did studies published in the "best" computer science journals during the same time period, since, as we have argued at length earlier in the paper, these two databases are quite comparable in their coverage of this research field in these respective publication venues.

## Discussion

Findings from the present study reinforce some of the findings from our study on the visibility of XML scholars (Zhao, 2005), and shed more light on issues both of citation analysis and of scholarly communication in transition.

### *Citation analysis*

Findings from this study suggest that citation analysis studies of research papers published on the Web can be a valuable complement to citation analysis studies of journal articles indexed by *SCI* by revealing a more current picture of a research field such as XML which is well-represented on the Web and is covered by a citation database publicly available on the Web. Citation analysis of research papers published on the Web appears to reveal a somewhat more current picture of the XML research field, whereas citation analysis of journal articles seems to provide a more detailed historical view of the field, although other factors will almost certainly play a complicating role in the interpretation of our data. As a result, research papers on the Web are quite likely a better data source for the detection of current research fronts in research areas where Web publication is becoming accepted, whereas journal articles may be better suited for a retrospective study of the evolution of a field. However, given the small overlap of citing papers between the two databases that we observed, a combination of data sources is likely to produce significantly better results than either of the two on its own for an analysis that is less focused on timeliness.

Thus we find that different data sources have different strengths, so that different research questions may well require the use of different data sources or the combination of multiple data sources. Citation analysis using either one of the two data sources alone would not reveal the complete and current communication structure of the XML research field. In other words, in order to gain a full view of scholarly communication patterns in, say, the XML research field, multiple data sources should be used rather than, say, only *SCI* or only *CiteSeer*.

The importance of using multiple data sources as opposed to relying solely on the ISI databases has also been suggested by the very different results produced when using different citation counting methods, as well as by the constantly increasing importance of scholarly publications on the Web, if more in some fields than in others.

Studies (e.g., Garfield, 1979; Lindsey, 1980; Persson, 2001) have shown that different citation counting methods can result in divergent author rankings and different pictures of the specialty structure of a research field. These differences can be particularly pronounced in areas where multiple authors are the norm rather than the exception (Zhao, 2003; Zhao, 2006b). Data sources that support more than one citation counting method such as *CiteSeer* should therefore be used to allow authors to be ranked and mapped based on more than one citation or co-citation counting method, and thus to permit results to cross-validate and complement each other. This way, a more accurate evaluation of scholars can be achieved, and a clearer and more complete intellectual structure obtained. This is true, again, more of some fields than others, but it is likely the degree to which multiple authors are the norm in a field that determines this rather than the degree to which a field is published on the Web.

Moreover, the rapid development of information technology has been revolutionizing the way that information is produced and exchanged. As a result, the scholarly communication system has been changing to a new model which "emphasizes conference papers, preprint archives, and the online availability of articles" – more in some fields than others (Goodrum et al., 2001, p. 662). In physics or computer science, for example, the Web is often a researcher's first choice for literature searching (Youngen, 1997). This means that the study of scholarly communication patterns demonstrated in this part of the literature is increasingly important, and

that it becomes a more serious problem to use the "journal only" ISI databases as the only citation analysis data source when studying this kind of fields.

### *Scholarly communication in transition*

As discussed earlier, results of the present citation analysis study indicate that, collectively, research reported on the Web and in journals appears to represent different stages of the development of the XML research field, with journals on average lagging a few years behind the Web. This mirrors current common practice in scholarly communication in some research fields such as mathematics, physics, and computer science. In these fields, scholars make available on the Web papers they have just finished writing, and immediately send a link to the papers to those people in their field who may be interested in working with the new results reported there while the papers make their way to formal publication in either conferences or journals. In other words, research in these fields is now largely initially reported on the Web to obtain priority and fast recognition, and subsequently distributed more gradually through other more formal channels such as journals to gain formal acceptance. As a result, in these fields, as Youngen (1997, p. 1) points out, "the Web is often the first choice for finding information on current research, for breaking scientific discoveries, and for keeping up with colleagues (and competitors) at other institutions," even if journals such as *Science* and *Nature* do provide breaking news services to some scientific communities.

Our current study does not provide enough data to allow us to calculate a concrete number of months for the time lag we observe between research reported on the Web and in journals. In part this is certainly due to the fundamental problem that this lag would almost certainly be different in different subfields, as publication venues and patterns do differ between them even in such tightly defined areas as the one we have studied. If we were to venture a guess, we would estimate an average time lag of two to three years in this particular field during this particular phase of its evolution. This time lag is likely the effect of more than one cause. The peer review process involved in the journal publication process would be expected to contribute from several months up to a year on average to the overall time lag, depending on different subfields with different publication patterns. We suspect that an additional contribution to the observed time lag

of similar magnitude may have resulted from the extra lead time required for the ISI's advisory boards to recognize as sufficiently "important" any new publication venues that might specialize on a newly emerging field like the one we studied here. We leave it to future research to attempt a more stringent quantitative analysis of the multiple causes for the effect we evidence here.

We have thus provided evidence for the rise of a "two-tier system" of scholarly communication in some research fields, a system that some scholars believe is a model of the emerging scholarly communication system (Poultney, 1996; van Raan, 2001; Zhao & Logan, 2003). In this model, the first tier is a "free space" which represents the scholarly enterprise in "real time" and is most likely to feature free Web-based publications, while the second tier is "the world of more formal publications" that is thought most likely to continue to be dominated by journals (van Raan, 2001, p. 61). Although some authors caution that this two-tier model may be adopted much more easily in some research fields than in others, and may never be adopted in some (Kling & McKim, 2000), here we provide evidence that it is indeed being adopted in at least one such field, albeit one that may be particularly prone to doing so, and it is reasonable to expect that a similar effect may be observable in other fields if and when they do start adopting Web publication of research findings.

As indicated by the present study, at least in the XML research field, the first tier appears to primarily serve as a dissemination medium that improves the efficiency and effectiveness of the informal scholarly communication on which scholars rely heavily to obtain the information they need for their current research, especially in a fast-moving field of research. The second tier may primarily serve as an archive and evaluation rather than information distribution device, although a few journals such as *Nature* and *Science* have also provided a "breaking news" service to some scholarly communities with all the corresponding risks with regard to quality control that are inherent in such a service (e.g., the cold fusion controversy). The potentially significantly faster and wider distribution of information on the Web can make the Web a perfect medium for the initial publication of new research results in the first tier, and the journal has long served well as an archive and evaluation device, which makes it a natural candidate for continuing in this role in the second tier.

As we concluded in a study on the visibility of XML scholars (Zhao, 2005), if this system continues to evolve in more research fields, journals in these fields that currently do not accept papers already published on the Web may eventually need to change their policies.

## Conclusions

New data sources and tools for scholarly communication research are increasingly becoming available on the Web to such an extent that they are now opening up opportunities for conducting a wider variety of studies on scholarly communication than before. This may contribute to the discovery of new research methods and lead to new theories in this area (Borgman & Furner, 2002; Zhao, 2003). The present study took this opportunity by conducting an author co-citation analysis of the intellectual structure of the XML research field using data from both journals as indexed by *SCI* and the Web as indexed by *CiteSeer*.

Findings from this study indicate that research in the XML research field has boomed during the second time period we have studied (i.e., years 2001 to 2006), and has clearly been in a stage of exponential growth indicative of a successfully emerging research field. Compared to the first time period we studied (i.e., from the birth of the field in approximately 1996 to 2001), new specialties have appeared such as *Information retrieval from XML documents and databases,* and *Access Control*, while old ones such as XML applications in the medical sciences disappeared. Theoretical aspects of XML research have attracted much more attention, as indicated by the emergence of two theoretical groups. Some existing specialties, such as *the Semantic Web,* have strengthened considerably, while others have found clearer emphases, such as intelligent mediation and translations in research on data integration. The XML research field has clearly matured.

Perhaps surprisingly, this pattern of development of the XML research field was already largely apparent from data provided by *CiteSeer* during the first time period (i.e., approximately 1996 to 2001), although sometimes less clearly. Examples include a stronger *Semantic Web* specialty, the existence of research foci such as *information retrieval* and *access control*, and the lack of specialties in XML applications in the medical sciences. Each of these trends is linked to

a number of individual authors in the analyses who exemplify the trends we observe and who thus contribute to the confidence level for our results.

This indicates that, in our study, research papers available during the same time period in the two different publishing media, namely the Web and the journal, represented two different stages in the evolution of the XML research field, with publications on the Web collectively offering a view more current by a few years. Consequently, we find that a citation analysis of research papers published on the public Web can be more successful in detecting current research fronts in a research field like XML than a traditional citation analysis based on traditional citation indexes of journal articles.

We have thus demonstrated both the importance and the feasibility of the use of multiple data sources in citation analysis studies of scholarly communication, and have found further evidence for a developing "two-tier" scholarly communication system in some research fields.

Future research is necessary to confirm the results from the present study, both by collecting data from *CiteSeer* for the second time period (i.e., 2001-2006), and by examining the extent to which results from the present study may be generalized to other research fields, as there are large differences in scholarly communication patterns between fields and between the different phases of the evolution of a field. It would also be interesting to find out if other bibliometric tools such as co-word analysis of abstracts or full texts of research publications would produce different results regarding the time lag between web and journal publications given that co-word analysis might be able to take into account a paper's forward-looking future research discussions rather than its inherently backward-looking references.

## Acknowledgements

This work was supported in part by a Fellowship of the School of Computational Science and Information Technology of the Florida State University, as well as by the Social Sciences and Humanities Research Council of Canada. We would also like to thank our anonymous reviewers for their insightful comments on an earlier version of this paper.

## References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*, 550-560.

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

Borgman, C.L. (2000). Digital libraries and the continuum of scholarly communication. Journal of Documentation, 56, 412-430.

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

Clever Project. (1999). *Hypersearching the Web*. Retrieved March 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.

Culnan, M.J. (1986). The intellectual development of management information systems, 1972-1982: A co-citation analysis. *Management Science, 32*(2), 156-172.

Culnan, M.J. (1987). Mapping the intellectual structure of MIS, 1980-1985: A cocitation analysis. *MIS Quarterly, 11*(3), 341-353.

De Solla Price, D.J. (1961). *Science Since Balylon*. New Haven: Yale University Press

Edge, D. (1979). Quantitative measures of communication in science: a critical review. *History of Science, 7*, 102-134.

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

Garfield, E. (1979). Citation indexing — its theory and application in science, technology, and humanities. New York: John Wiley & Sons.

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.

Hair, J.F. Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate data analysis* (5th edition). Upper Saddle River, NJ: Prentice Hall.

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. *Proceedings of the Midyear Meeting of American Society for Information Science, 1996,* 299-315.

ISI (2004a). *The Impact of Open Access Journals: A Citation Study from Thomson ISI*. Retrieved May 19, 2006, from http://www.isinet.com/media/presentrep/acropdf/impact-oa-journals.pdf

ISI (2004b). *The Thomson Scientific journal selection process*. Retrieved May 19, 2006, from http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/.

Kreuzman, H. (2001). A co-citation analysis of representative authors in philosophy: examining the relationship between epistemologists and philosophers of science. *Scientometrics, 51,* 525-539.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *Proceedings of the 59th ASIS Annual Meeting* (pp71-78). Medford, NJ: Information Today/ASIS.

Lawrence, S., Giles, C. L., & Bollacker, K. (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.

McCain, K. W. (1986). Cocited author mapping as a valid representation of intellectual structure. *Journal of the American Society for Information Science, 37*(3), 111-122.

McCain, K. W. (1990a). Mapping authors in intellectual space: population genetics in the 1980s. In C. L. Borgman (ed.), *Scholarly Communication and Bibliometrics* (pp.194-216). Newbury Park, CA: Sage.

McCain, K. W. (1990b). Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science, 41,* 433-443.

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/.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics, 50*(2), 339-344.

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

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., Vaughan, L., & Bjorneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology, 39,* 81-135.

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

W3C (2001). *Extensible Markup Language (XML) Activity Statement*. Retrieved December 2, 2001, from http://www.w3.org/XML/Activity.

White, H.D. (2003). Author cocitation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 54,* 1250-1259.

White, H.D., & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32,* 163-171.

White, H.D., & Griffith, B.C. (1982). Authors as markers of intellectual space: Co-citation in studies of science, technology and society. *Journal of Documentation, 38*(4), 255-272.

White, H. D. & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*, 327-355.

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*. Dissertation, 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. *Proceedings of the American Society for Information Science and Technology 2004 Annual Meeting* (pp 72-83), November 13-18, 2004, Providence, Rhode Island, USA.

Zhao, D. (2005). Challenges of scholarly publications on the web to the evaluation of science – a comparison of author visibility on the web and in print journals. *Information Processing & Management, 41*(6): 1403-1418.

Zhao, D. (2006a). Towards all-author co-citation analysis. *Information Processing & Management, 42*, 1578-1591.

Zhao, D. (2006b). Dispelling the myths behind straight citation counts. To appear in *Proceedings of the American Society for Information Science and Technology 2006 Annual Meeting*.

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., & Strotmann, A. (2004). Towards a Problem Solving Environment for Scholarly Communication Research. *Proceedings of the Canadian Association for Information Science 2004 Annual Conference*, June 3-5, 2004, Winnipeg, Canada.