Web-Based and Print Journal-Based Scholarly Communication in the XML Research Field: a Look at the Intellectual Structure

Dangzhi Zhao
School of Library and Information Studies, University of Alberta, Edmonton, Alberta, Canada T6G 2J4.
Email: dzhao@ualberta.ca

As part of a research project that aims to identify the similarities and differences between Web-based and print journal-based scholarly communication, this paper compares the intellectual structure of the XML research field revealed from author co-citation analysis of research papers published on the Web as indexed by *ResearchIndex* and that derived from print journals as indexed by *SCI*. Considerable differences are observed and some media specific features are identified. Results from this study demonstrate the importance and the feasibility of the use of multiple data sources in citation analysis studies of scholarly communication, and evidence for a developing "two-tier" scholarly communication system.

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 the types of communication that are taking place and the similarities to what we have come to expect from print based communication. Citation analysis and other bibliometric techniques have been successfully applied to the study of this new phenomenon in scholarly communication. As Zhao & Logan (2002) point out, such applications roughly fall into three categories of study. One is to apply, often with modifications, citation analysis and other bibliometric principles and techniques to study the characteristics and link structures of the Web. Examples include studies on search engines making use of hyperlink structure (Clever, 1999), and so-called "Webometric" studies (Almind & Ingwersen, 1997; Cronin et al., 1998, Egghe, 2000; Larson, 1996a; Rousseau, 1997; Thelwall & Harries, 2004; Turnbull, 2000; Wilkinson et al., 2003). The second category of studies looks at "electronic ingredients" in journal articles — either in reference lists or in abstracts — to see the impact of electronic publications on traditional print journal-based scholarly communication (Harter, 1992; Harter & Kim, 1996; ISI, 2004a; McCain, 2000; Youngen, 1997).

A third important category of study — citation analysis using research papers published on the Web as a data source — has recently begun (The Open Citation Project, 2001; Goodrum et al., 2001; Zhao & Logan, 2002, 2003; Zhao, 2003). Full text research papers along with corresponding tools for searching for citations from these papers are becoming increasingly available on the Web; examples include *ResearchIndex*[1] and *CiteBase*[2]. These citation indexes are different from those for print journals such as the ISI databases[3] in that, among others, papers they index use the Web as their communication medium, which affords higher speeds of communicating and wider distribution of information than the journal; they cover a wider range of document types such as degree theses, technical reports, conference papers, and preprints in addition to journal articles which may represent different stages in the scholarly communication process; they contain more information about cited documents such as all authors, full titles and full names of sources as compared with the limited information (first authors only, and abbreviations of source journal titles) provided by the ISI databases; and their source paper selection and indexing process is highly automatic and inclusive as compared with the manual and highly selective process employed by the ISI databases. These data and tools have opened up the possibility of a larger variety of inquiries, such as how scholarly communication is being transformed, what are the similarities or differences between the new formats and the traditional ones, and how the new formats facilitate or inhibit the scholarly communication process. Just as the advent of the ISI databases 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 advance (Zhao, 2003; Zhao & Logan, 2002).

We have conducted a research project that attempts to explore this possibility and to systematically compare scholarly communication patterns between the Web and the print world in the eXtensible Markup Language (XML) research field (Zhao, 2003). Some of the results from this project have been reported, including a pilot study that demonstrated the feasibility of such studies and raised many further issues to explore (Zhao & Logan, 2002), and a study that identified some of the similarities and

[1] http://www.researchindex.com
[2] http://citebase.eprints.org/cgi-bin/search
[3] http://www.isinet.com/

differences in author visibility between Web-based and print journal-based scholarly communication (Zhao & Logan, 2003).

The present paper discusses further results from this research project. The objectives are (1) to identify the similarities and differences in the intellectual structure between Web-based and print journal-based scholarly communication in the XML research field; and (2) to explore possible contributing factors. This study may contribute to the understanding of scholarly communication in transition and to the advance of citation analysis theory and methodology.

Research questions

The research questions to be explored in the present study are as follows.

- What are the differences between the intellectual structure of the XML research field revealed from the Web and that derived from print journals?
- What are the similarities in the intellectual structure of the XML research field revealed from the Web and print journals?
- What has contributed to the differences in the intellectual structure between the Web and the print world?

Although the pilot study mentioned above also compared the intellectual structure between the Web and print journals, the comparison in the present study is carried out controlling for data scope and citation counting method, and attempts to explore possible contributing factors to differences in the intellectual structure.

Methodology

Research field to be analyzed

The XML research field was chosen for this comparative study. This field was "digitally born" and has been growing with the Web. As a result, the Web is naturally one of the major communication channels used by the XML research community. This, on the one hand, ensures that the number of research papers published on the Web in this field is likely to be large enough for applying citation analysis approach and that new models for scholarly communication, if any, should be more easily identified. On the other hand, however, this also means that the extent to which results from the study of this field can be generalized to other fields may be limited because Web publications in these other fields may not fully represent the entire fields. The value of the present study lies not so much in the identification of characteristics in scholarly communication that apply universally, but more in its implications for understanding the transition of scholarly communication systems from print-centered to Web-based format, as using the Web to formally communicate research results represents an important trend in scholarly communication.

Thus, the XML research field was chosen purposely to maximize the visibility of the differences between the new and traditional publishing media. In the future, we may expand this study to include more research fields, both those that are similar to the XML research field and those that are different in terms of the degree to which they are related to the Web and have adopted Web publishing. As this would involve multiple research fields and data sources, such a study, while significant, would not be an easy task, and therefore would greatly benefit from a Problem Solving Environment (PSE) for scholarly communication research (Zhao & Strotmann, 2004).

Data collection

Science Citation Index (*SCI*) and *ResearchIndex* were used in the present study to collect information on research papers published in print journals and on the Web, respectively. *SCI* is one of the ISI's databases and *ResearchIndex* one of the citation indexes for Web publications discussed in the *Introduction* section. To date, *SCI* along with other ISI databases has been used as the 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; Pringle, 2004). Developed by the NEC Corporation Research Institute, *ResearchIndex* 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 are in the broadly defined computer science field and are publicly available on the Web. Our previous studies have provided evidence that citation analysis studies using *ResearchIndex* as a data source are as valid as those using *SCI* (Zhao & Logan, 2002, 2003; Zhao, 2003). More information about *ResearchIndex* can be found in these studies and also in Bar-Ilan (2001), Goodrum et al (2001), and Lawrence et al (1999).

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

Thus, the 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 *ResearchIndex*) and downloaded into a local machine. Since the existence of duplicates was found to be one of the major differences between traditional databases and the Web, paper entries

retrieved from *ResearchIndex* 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 analysis such as counting citations and co-citations.

Note that in the present study, the search for citing papers in *ResearchIndex* was limited to "Header" fields rather than searching in the full text of the documents as we did in the pilot study. The reason for this change in our data collection method was that *SCI* only goes as far as abstracts in indexing citing papers, and "Header" fields in the *ResearchIndex* database were assumed to be similar in scope. We hoped that this way of collecting data would result in more comparable data from the two data sources.

Also note that no citation windows were specified in this study, indicating that publications from all years were used. This design was based on the fact that XML research was a fairly young field of study and had a history of only about six years since the first phase of W3C's XML activities started in June 1996 (W3C, 2001). A six-year period is among the citation windows commonly used in citation analysis studies.

Data analysis

The commonly used steps and techniques of author co-citation analysis (McCain, 1990b; White & McCain, 1998; White, 2003; Zhao, 2003) were followed in this study. Core sets of authors were selected based on "citedness" — the number of citations they received. Citedness above some threshold is a good criterion for selecting authors in author co-citation analysis although the resulting authors may not be "wholly definitive" of the research field being studied (White & McCain, 1998, p. 332). Three sets of highly visible authors were thus selected from the three data sets — the data set from *ResearchIndex*, the one from the entire *SCI* database, and the one from a subset of *SCI* addressing computer science research. There are no strict rules regarding thresholds for citation-based author selection in author co-citation analysis studies (McCain, 1990b). Assuming that the more authors the better a research field is represented, the present study used low thresholds to allow 100 authors to be included in the final multivariate analysis, the maximum number of variables possible when using ALSCAL, the multidimensional scaling routine in SPSS (version 10.0).

A Java program was developed to count author co-citation frequencies and to record them in matrixes (Zhao, 2003). These co-citation matrixes were then cleaned by deleting authors who were co-cited with very few other authors based on the assumption that authors who have little connection with the rest of the field are not good representatives of the field. Specifically, an author was deleted if the corresponding row/column contained more than 95% zero value cells. The resulting matrixes were then converted to Pearson r correlation matrixes that were in turn used as input to the two multivariate analysis procedures employed: Factor Analysis (FA) and Multi Dimensional Scaling (MDS).

Factors were extracted by Principal Component Analysis (PCA) with an oblique rotation (SPSS Direct OBLIMIN) because of the theoretical expectation that the resulting factors (specialties) would in reality be correlated. The number of factors extracted was determined based on Kaiser's rule of eigenvalue greater than 1 because the resulting model fit was adequate as represented by total variance explained, communalities, and correlation residuals (Hair et al., 1998).

The multidimensional scaling procedure used in this study was SPSS ALSCAL as many studies have done (White & McCain, 1998; Kreuzman, 2001), and the two-dimensional maps (MDS maps) were generated using LaTeX from the coordinates resulting from the ALSCAL procedure.

With the aid of both factor analysis and multidimensional scaling techniques, the grouping of the scholars within each set of authors was analyzed, and results from the three datasets were compared.

Note that because straight counts are the only citation counts supported by *SCI*, straight counts were used here for counting authors' citations and co-citations to ensure comparable data analysis method between the two publishing media although complete counts and fractional counts are also supported by *ResearchIndex* and results using these two counting methods were also obtained. Results from the comparisons between different citation and co-citation counting methods will be reported in a separate paper.

Results and discussion

A search on "XML" or "eXtensible Markup Language" resulted in, after removing duplicates, 312 papers using *ResearchIndex* and 374 papers with reference lists using *SCI*, 268 of which were from computer science journals. The papers from *ResearchIndex* made 4,578 citations, and those from *SCI* made 6,782 citations. Among the citing papers, there are 26 common to both *ResearchIndex* and *SCI*. The percentage of citing papers shared by the two data sources is very low (7% of papers in *SCI* and 8% in *ResearchIndex*). This means that in the XML research field, papers published in journals are not largely made available on the Web and papers published on the Web are not well represented in *SCI*.

As Factor Analysis when applied in author co-citation analysis has been shown to provide clear and revealing results as to the nature of the discipline (White and McCain, 1998), the intellectual structure of XML research will be discussed mainly based on factor analysis results presented in Table 1 and Table 2, complemented by the MDS maps, only one of which is presented here due to limited space (Figure 1).

The factor names shown in the column headings of the tables were given based on the examination of the cited articles written by authors in the corresponding factors. Following White and McCain's example, authors are ranked in the factor on which they load most highly and their loadings on other factors that are above 0.4, if any, are also presented, indicating their contributions to more than one specialty (White & McCain, 1998). If an author does not load 0.4 or higher on any of the factors, the author's highest loading, whatever it is, is presented. If large factors are interpreted as specialties, the results of the factor analysis presented in the tables reveal the specialty structure of the XML research field and the associated authors' memberships in one or more specialties as seen by citing authors in the three datasets (White & McCain, 1998).

Table 1 presents the results of a factor analysis of highly cited XML researchers selected from *ResearchIndex*, and Table 2 presents those from the computer science journals in *SCI*. Kaiser's rule of eigenvalue greater than one resulted in an eleven-factor model from *ResearchIndex* which accounts for 96% of the total variance, and an eight-factor model from *SCI* which explains 94.6% of the total variance. In both cases, the differences between observed and implied correlations were for the most part (almost 100%) smaller than 0.05. Results from the entire *SCI* database are not presented here due to the limited space and also because they are very similar to those from the portion of *SCI* indexing computer science journals.

Table 1: Factor Analysis of 100 authors in the XML research field (*ResearchIndex*)

Authors	Semi-structured or XML DBs	Foundations of XML data mgt. & proc.	The Semantic Web	Prog. /proc. XML data	NLP	Version mgt.	Functional and Logic Prog.	DB & IR founda-tions	KM	Access ctl.	Data integration
L. Fegaras	0.84										
S. Adler	0.79										
J. F. Naughton	0.74										
P. Atzeni	0.61										
D. Maier	0.56										
D. D. Chamberlin	0.56										
R. Cattell	0.55										
M. J. Carey	0.53										
D. Beech	0.48										
F. Bancilhon	0.45	0.43									
J. Shanmugasundaram	0.43										
V. Christophides	0.42										
J. Widom	0.41										
J. Miller	0.40										
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		0.67									
J. E. Hopcroft		0.67		-0.43							
J. Clark		0.64									
H. Hosoya		0.61									
P. Fankhauser		0.57									
P. Buneman		0.56									

A. Davidson		0.53									
A. Sahuguet		0.49		-0.44							
L. Cardelli		0.42									
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								
S. Decker			0.75								
D. Megginson				-0.84							
D. Lee				-0.74							
A. Aho				-0.70							
N. Klarlund				-0.67							
M. Murata		0.42		-0.64							
R. Bourret				-0.62							
E. Maler				-0.61							
D. Fallside				-0.51							
C.-C. Kanne				-0.44							
L. Wood				-0.43	0.43						
A. Schmidt				-0.41							
J. Bosak	0.35			-0.37							
H. Jagadish					0.82				0.45		
M. Kay					0.74						
N. Walsh					0.67						
J. K. Ousterhout					0.66						
C. Barras					0.63						
D. McKelvie					0.36						
A. Albano					-0.33						
G. Ghelli					-0.33						
L. Liu						0.95					
A. Marian						0.94					
S.-Y. Chien						0.88					
J. Chen						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								0.44			
J. Robie		0.33						0.38			
C. Freitag									0.81		
P. McBrien									0.51		
E. Bertino										1.03	
E. Damiani										0.90	
A. Gupta											1.06
M. Kifer											0.57
B. Ludascher											0.52
S. Nestorov											0.34

Differences

We will first compare the results from *ResearchIndex* data and those from computer science journals in *SCI* as this comparison filters out data coverage concerns in addition to the concerns regarding data retrieval and citation counting method as discussed in the *Methodology* section. Then we are going to look at the results from the entire *SCI* database to see if any additional differences may appear.

As seen from the two tables, eleven factors emerged from *ResearchIndex* and eight from computer science journals in *SCI.* Some of them are highly coherent groups, some are less coherent, and others pick up some interesting isolates.

ResearchIndex — (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,* (7) *Functional and Logic Programming,* (8) *DB and IR foundations,* (9) *Knowledge Management,* (10) *Access control, and* (11) *Data integration.*

SCI — (1) *Management of semi-structured or XML databases,* (2) *Web standards, specifications and guidelines,* (3) *Intelligent Web service management and 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*.

Table 2: Factor Analysis of 100 authors in the XML research field (*SCI computer Science*)

Authors	XML DBs	Web standards	Intelligent web service mgt. & integration	Foundations (formal languages)	XML for medical decision support	Intelligent Software Agents on the Web	The semantic Web	XML for medical data exchange / Hypermedia
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							
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							
P. PS. Chen	0.62							
M. Fernandez	0.61							
S. Abiteboul	0.57							
C. Beeri	0.57			-0.44				
S. Chawathe	0.56						-0.47	
P. Buneman	0.55							
D. Suciu	0.55							
D. Calvanese	0.54			-0.41				
J. Widom	0.53							
H. Thompson	0.52							
V. Christophides	0.52							
F. Bancilhon	0.52							
S. J. Derose	0.51							-0.47
S. Cluet	0.50							
R. GG. Cattell	0.49							
T. Milo	0.49			-0.49				
D. Lee	0.49		-0.42					
Y. Papakonstantinou	0.49							
M. Murata	0.48							
G. Wiederhold	0.44							
D. Dubois	0.41							

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						
E. Maler		0.59						
T. Bray		0.59						
B. Bos		0.51						
C. M. Sperberg-McQueen		0.50						
J. Bosak		0.47						
J. Clark		0.47						
A. Layman		0.47					0.40	
P. Murray-Rust		0.41						
L. Wood		0.40						
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			-0.56					
A. Levy			-0.53					
L. Liu			-0.51					
A. Hunter			-0.46					
D. Fensel			-0.43					
D. Harel			-0.40					
D. Remy				-0.98				
A. Bruggemann-Klein				-0.85				
L. Cardelli				-0.80				
A. Ohori				-0.78				
A. Aiken				-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			
D. Connolly					-0.40			
M. Wooldridge						-1.00		
D. A. Benson							-0.80	
S. B. Davidson							-0.75	
O. Lassila							0.55	
S. Decker	0.52					-0.50	0.54	
D. Brickley	0.34	0.36		0.35			0.36	
S. B. Johnson								0.76
L. Alschuler								0.74
C. Friedman								0.72
J. J. Cimino								0.65
R. H. Dolin								0.60
S. Chakrabarti						0.44		-0.63
K. Gronbaek								-0.58
F. Halasz						0.31		-0.40
T. Berners-Lee		0.31						-0.39

Clearly, XML research represented in *ResearchIndex* and that reported by *SCI* have been concerned with very different issues. This is indicated by the different factors that emerged and by the size of the factors and how clearly they are separated from each other.

Web standards, specifications and guidelines does not appear separately as a specialty on the Web. Instead, authors in this area are scattered into several groups. For example, Bray (XML, namespaces in XML), Clark (XPath, XSLT), and Apparao (DOM) are in the second specialty (*Foundations of XML data management & processing*), Maler and Wood are in the *Programming for and processing of XML data* specialty, and Brickley, Lassila (RDF) and Berners-Lee are in *The Semantic Web* group. In the print journals, however, this is still a distinct group although some of the standards people such as DeRose and Lassila have been placed in other specialties. It seems that on the Web, groups of interacting standards or specifications have been perceived as the foundation of different specialties once they were formulated while in journals this distinction is not so clear yet.

The Semantic Web group is very distinct on the Web but quite weak in journals. Unlike on the Web, authors in journals appear to still refer regularly to general computer science foundations (formal languages). Moreover, applications of XML in medical science form two distinct groups in journals but do not occur at all on the Web. This is true the other way around with the specialty *Programming for and processing of XML data.*

In addition to indicating the different research focuses, all of this seems to also suggest that studies reported on the Web are perhaps more at the research front than those in print journals, considering that *The Semantic Web* is an emerging research focus and that, unlike *the Semantic Web* and *the programming for and processing of XML data*, *XML applications* are about relatively mature rather than cutting-edge technologies from the point of view of XML research. This makes sense because the Web can afford higher speeds of communication and publishing on the Web does not have to go through the time-consuming formal publication process, which can take years.

Of course, there might be other factors that have caused the different visibility of *XML applications* in the two media. Scholars in an application area of a technology (e.g. computational biology) may have adapted to the publishing tradition within that field (e.g. biology) which may be different from that in the field of the technology (e.g. 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 because they act more like, say, biologists than like computer scientists in terms of their publishing behavior. However, this is just one of the observations that have suggested that studies published on the Web are more at a research front than those published in print journals (Zhao, 2003; Zhao & Logan, 2003).

Although the oblique rotation method used in the present study allows the examination of the interrelationships between the specialties, results are not reported because the emphasis here is on the comparison between the two media and there are not many specialties that are common to these two media. Actually there are only two as seen in the tables and the relationship between them will be discussed later.

When the entire *SCI* database is used in searching for citing papers, as mentioned earlier, the intellectual structure of the XML research field revealed from author co-citation analysis largely remains the same as that from the portion of *SCI* indexing computer science journals. The only specialty that is identified from the entire *SCI* database that does not occur separately as a specialty from the computer science data is *XML and Chemistry*. Specialties that have been identified from computer science data but not from the entire *SCI* database include *The Semantic Web*, *Hypermedia and hypertext*, and *XML for medical decision support.* It appears that research on *XML and Chemistry* is not largely published in journals that are considered by the *SCI* database as belonging to computer science while studies on *XML for medical decision support* are. The fact that *The Semantic Web* and *Hypermedia and hypertext* are not distinct groups in the entire *SCI* database suggests that research in these areas is relatively more intensive in computer science journals. The work of these groups is treated in the entire *SCI* database more as general guidelines, as indicated by authors in these groups such as Lassila and Berners-Lee being placed in the *Web standards, specifications and guidelines* group in the results from the entire database, which corresponds to a greater emphasis on their earlier work.

Thus, the differences identified from the comparison between the computer science journals in *SCI* and *ResearchIndex* as discussed earlier become slightly larger when the entire *SCI* database is considered. For example, the distinct specialty in the results from *ResearchIndex* data, *The Semantic Web*, was also identified from the computer science journals in *SCI* although not as distinctly, but does not appear separately in the entire *SCI* database as a specialty, and the specialty, *Web standards, specifications and guidelines*, is very distinct in the results from the entire *SCI* database, but is not as clear a grouping any more in those from the computer science journals in *SCI*, and completely disappears on the Web. Another example is that the specialty, *application of XML in chemistry*, is not identified in either *ResearchIndex* or computer science journals in *SCI*, but emerges quite distinctly in the entire *SCI* database.

It appears that the time-lag in acknowledging the shift of authors away from their earlier research focuses to their new ones is larger in the *SCI* computer science journals than on the Web, and is even larger in the entire *SCI*

database as it normally takes longer for the current XML research to reach people from outside computer science. It also appears that the analysis of the entire *SCI* database may capture more interdisciplinary aspects of XML research (e.g. *XML and chemistry*) than that of computer science data, since there exist publications on XML in fields outside computer science.

Different concerns and emphases on the Web and in print journals can also be seen from the regrouping of the highly visible authors who are common to both media, as indicated on the map in Figure 1. This map depicts the moves of these authors, indicated by arrows, from their positions resulting from *ResearchIndex* data to those from the entire *SCI* database. It clearly shows the changes of position of author-points mentioned earlier, that is authors, including Bosak and Goldfarb, moving from *The Semantic Web* or other groups to the scattering across the top to form a distinct *Web standards, specifications and guidelines* group.

Other significant moves include that of Decker and Fensel, both starting from *The Semantic Web* specialty but heading to different groups. Actually, Fensel along with Liu has moved to join some other scholars that emerged in the analysis of *SCI* data in the specialty *Intelligent web service management and integration*, and Decker has been related to Wooldridge for research on *Intelligent Software Agents on the Web*. It seems that although the new *The Semantic Web* specialty has not yet been as well represented in journals as on the Web, the trend of well-established specialties, such as *Software agents* and *Data management and integration on the Web*, heading to the Semantic Web direction has been reflected in print journals. This might be because the inertia of a specialty makes well-established specialties tend to continue publishing in traditional media such as the journal that may have served them quite well, resulting in more publications in journals than through new channels such as the Web. This inertia effect does not exist in emerging specialties such as *The Semantic Web*.

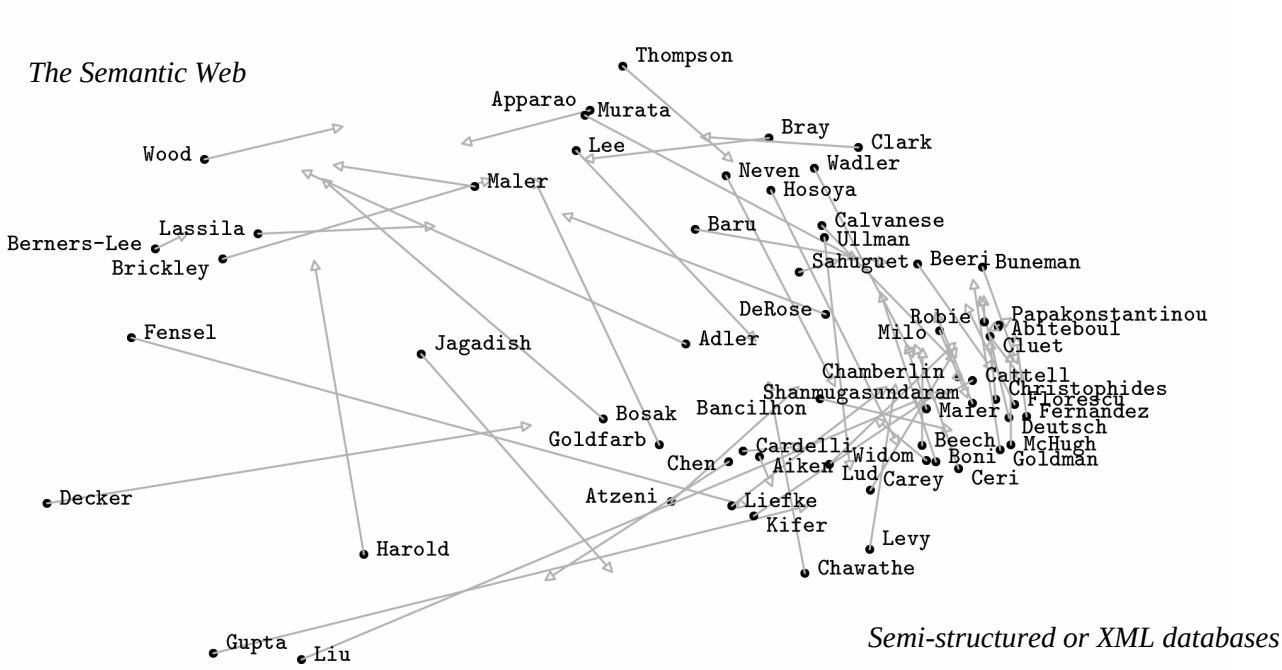

Figure 1: Regrouping of highly cited authors who are common to both media
(Moves are indicated by arrows pointing from positions on the map from *ResearchIndex* to those from the entire *SCI* database)

Similarities

There are two aspects in which congruence was observed between results from *ResearchIndex* using straight counts (Table 1) and those from *SCI* — both computer science journals only (Table 2) and the entire database.

The first one is that *XML (or, semi-structured) data management* is the most active research area in the XML research field. About forty percent of authors in the analyses are placed in this research area. Authors working in this area form a single group in the results from the *SCI* data while, in those from *ResearchIndex* data, this research area splits into two, which was also observed in the pilot study (Zhao & Logan, 2002).

It is not clear why the two database groups in the results from *ResearchIndex* data keep merging into one single

group in the results from *SCI* data. One possible explanation is that the intellectual difference between the two groups is blurred or weakened in the print world by such factors as "diplomatic citing." As Edge (1979, p. 120) observed, "adding a list of references to a paper is often a last-minute chore: colleagues, 'trusted assessors', referees and editors all contribute suggestions as to authors and papers that 'ought' to be included somewhere." These citations are usually not among the core set of documents that directly contributed to the writing of a citing paper, and would therefore most likely widen its intellectual scope from the point of view of a citation analysis. Since editors or referees often come into play after the papers have been published on the Web, this type of citations would obviously happen more frequently in the print world, which may have pulled the two database groups together. This appears to be supported by the data from the present study: as shown at the beginning of the *Results and discussion* section, the reference lists in print journals were at an average about 20% longer than those on the Web (18.1 vs. 14.7 references per paper).

The second congruence is that the dominant structural dimension of the XML research field in both sets of results is the degree to which database technology is involved. This can clearly be seen from the MDS maps. Although the individual maps from the two data sources are not presented here, we still can see this, to some degree, on the map in Figure 1. From left to right, the involvement of database technology becomes more pronounced with scholars working on *The Semantic Web* at the left end that has little to do with databases and those on *semi-structured or XML databases* at the far right in which the database technology is the focus.

Conclusions

New data sources and tools for scholarly communication research increasingly available on the Web have opened up the possibility of various studies that may develop new methods and lead to new theories (Borgman & Furner, 2002; Zhao, 2003). The present study explored this possibility through an author co-citation analysis of the intellectual structure of the XML research field using data from both print journals as indexed by *SCI* and the Web as indexed by *ResearchIndex.*

Findings from this study indicate that 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 or specifications and XML database design and implementation, research on XML applications is a focus only in journals, and research into the Semantic Web and programming for and processing of XML data is better represented on the Web. It appears that while emerging specialties such as *The Semantic Web* are better represented on the Web, new trends in well-established specialties such as *Software agents* are quite visible in journals.

This reinforces some of the findings from our study of the visibility of XML scholars (Zhao & Logan, 2003), and can shed more light on issues of both citation analysis and scholarly communication in transition.

Citation analysis

Findings from this study clearly show that citation analysis using either one of the two data sources alone would not reveal the complete communication structure of the XML research field. In other words, in order to gain a complete picture of the scholarly communication patterns in the XML research field, multiple data sources should be used rather than only the *SCI* databases or *ResearchIndex*.

The importance of using multiple data sources is also suggested by the very different results from different citation counting methods and by the increasing importance of publications on the Web.

Studies (e.g. Garfield, 1979; Lindsey, 1980; Zhao, 2003; Zhao & Logan, 2003) have shown that different citation counting methods can result in divergent author rankings and different pictures of the specialty structure of a research field. Data sources that support various citation counting methods such as *ResearchIndex* should 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 would be achieved, and a clearer and more complete intellectual structure obtained.

Moreover, the rapid development of information technology is revolutionizing the way that information is produced and exchanged. As a result, the scholarly communication system is 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.

Scholarly communication in transition

As discussed earlier, the differences in research focus in the two media suggest that research published on the Web is perhaps more at a research front than that in print journals in the XML research field. Actually, it has become very common in some fields such as mathematics and computer science that scholars put on the Web the papers they have just finished and immediately send a link to the papers to those people in their field who may be interested while the papers make their way to either conferences or journals, which can take years. In other words, research in

these fields is now largely being initially reported on the Web to obtain priority and fast recognition and then gradually distributed 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."

All this seems to provide further evidence of 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; 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 most likely to continue to be dominated by journals (van Raan, 2001, p. 61). As suggested by the present study, the first tier primarily serves as an information distribution medium to make the informal communication, on which scholars have relied heavily to obtain the information they need for their research, more effective and efficient. And the second tier primarily serves 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, and 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 visibility of XML scholars (Zhao & Logan, 2003), if this system evolves, journals that currently do not accept papers already 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 would significantly improve the efficiency of scholarly communication.

Acknowledgements

I wish to thank Dr. Gary Burnett and Dr. Elisabeth Logan for their stimulating advice and guidance, and Dr. Andreas Strotmann 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

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. *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.

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 8, 2004, from http://www.isinet.com/media/presentrep/acropdf/impact-oa-journals.pdf

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

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. (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/

Pringle, J. (2004). Do Open Access journals have impact? Retrieved May 20, 2004, from http://www.nature.com/nature/ focus/accessdebate/19.html

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

Turnbull, D. (2000). *Bibliometrics and the World-Wide Web*. Retrieved April 13, 2003, from http://www.ischool.utexas.edu/ ~donturn/research/bibweb-abstract.html

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. & 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.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29(1)*, 59-66

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

Zhao, D. & Logan, E. (2003). A comparative citation analysis study of Web-based and print journal-based scholarly communication: a look at author visibility in the XML research field. *Proceedings of the 9th International Conference on Scientometrics & Informetrics* (pp 378-392), August 25-29, 2003, Beijing, China

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