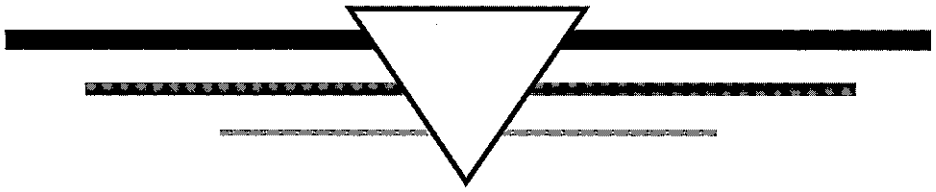
Visualizing Library and Information Science concept spaces through keyword and citation based maps and clusters

Åström, Fredrik

Published in:
Emerging frameworks and methods: CoLIS4

2002



Visualizing Library and Information Science Concept Spaces Through Keyword and Citation Based Maps and Clusters

Fredrik Åström
Inforsk, Department of Sociology
Umeå University
901 87 Umeå, Sweden
Fredrik.astrom@soc.umu.se

ABSTRACT

Co-citation analysis has been widely accepted as the foremost method for bibliometric mapping of research fields, whereas analyses based on keywords have been discussed, without gaining any overall acceptance. There are, however, advantages with keywords such as being understandable by others than those immediately connected to the field analyzed. This study aims at testing the relation between keyword and citation based analyses, and showing the significance of journal selection while mapping scientific fields. The preliminary study is based on 1135 *Social Science Citation Index (SSCI)* records from nine library and information science journals with descriptors added from the *Resources Information Center Database (ERIC)* database. Three maps are compared: one based on co-citations, one on keyword co-occurrences, and one merging citations and keywords. The mappings show the same basic structures, and when

merged, cited authors and keywords form corresponding relations. In comparison with earlier bibliometric studies, the wider journal selection makes it possible to identify a library science research area within library and information science.

1. INTRODUCTION

The outcome of the bibliometric analysis is dependent on what information within the database is being analyzed. Another aspect with strong implications is the journal selection. Both of these are questions related to types of information being used in the analysis. Citations and keywords are examples of two types of analyzable bibliographic information, but how can they be used in bibliometric analyses? What are the similarities and differences between bibliometric mappings based on keywords or citations, and how does journal selection affect how the field is described? These questions are important in the case of analyzing cognitive structures of a research field. Library and information science (LIS) is an example of a field that has been subject to various attempts at definitions: by bibliometric analyses (e.g., Persson, 1994; White, and Griffith, 1981; White, and McCain, 1998), in essays (e.g., Saracevic, 1999), in research papers taking a theoretical standpoint (e.g., Hjørland, 2000; Ingwersen, 1992; Vakkari, 1994), and in historical studies (e.g., Hahn, and Buckland, 1998). These texts come to different conclusions on the nature of LIS, even on fundamental issues such as basic sub-fields of the discipline and relations between library science and information science.

The most comprehensive bibliometric study of LIS so far was made by White and McCain (1998). The analysis identifies two sub-disciplines—information retrieval (IR) and studies of aspects of literature and communication—and 11 research specialties. This was done by an author co-citation analysis based on the 120 most cited authors in 12 key LIS journals during the years 1972–1995. One issue that requires comment is the selection of journals, where there is a strong emphasis on information science journals, whereas library science journals are few and only cover issues on library automation.

Co-citation techniques were introduced by Small and Griffith (1974), and further developed by Small and Sweeney (1985). A development of co-citation analysis for mapping of research fields is author co-citation, where the relationships between authors and their *oeuvres* are analyzed (McCain, 1986; White, and Griffith, 1981). Methods for co-word analysis have been developed to measure the associative strength of terms representing publications and documents (Callon, Courtial, and Laville, 1991; Courtial, 1994; Courtial, and Law 1989; Law, and Whittaker, 1992; Whittaker, 1989). But word and co-word analyses in bibliometrics have also been criticized due to their instability: words change, both in terms of usage and meaning. This renders the networks representing the cognitive spaces likely to destabilize over time, as opposed to citation analysis, where stability is assured by the use of whole documents (Leydesdorff, 1997).

The use of combinations of different analyses has been tested by, for example, Braam, Moed and Van Raan (1991a; 1991b), who investigated the structural and dynamical aspects of a combined co-citation and word analysis to map scientific structures. They combined a co-author analysis with a frequency analysis of index terms and classification codes, keywords from abstracts and so on. The results indicate a complimentary role of the two analyses. Relations between subject and citation indexing from an information retrieval point of view have been analyzed in a number of studies. McCain (1989) compared descriptor and citation retrieval in 11 search topics and evaluated these for relevance and novelty. Both search strategies gave a good result, even though the recall ratio was larger for the descriptor retrieval, and the overlap between the different strategies was moderate. The complimentary role of citation retrieval has also been confirmed by Pao (1993) and Pao and Worthen (1989). A database was constructed with document representations consisting of both semantic information and information on cited documents. Also here the semantic retrieval led to higher recall, both strategies provided good precision, and the overlap was relatively small.

2. AIM

This paper aims at comparing and analyzing different indexing methods—keywords and citations—to see how they affect bibliometric analyses of LIS. Do the different kinds of indexing bring on major changes in the perception of fields of research, or do they show the same structures when used as a base for mapping of cognitive fields? This will be examined by analyzing one set of records according to three modes of co-occurrence analysis. The first two will provide results that can be compared. The third analysis will merge the two first analyses, to see if there is any level of concordance. A second aim is to test an alternative journal selection for an analysis of LIS, to see what differences can be found in comparison to previous analyses of LIS. The main questions in this paper are: what different kinds of bibliometric analyses can be used for describing the cognitive structures of LIS? What differences or similarities can be found in different bibliometric analyses on one set of records? What are the implications of different journal selections when analyzing LIS?

3. METHOD

This paper compares three different bibliometric analyses on one set of records. Articles from nine LIS journals were selected, based on ranking in *the Institute for Scientific Information (ISI) Journal Citation Reports*. The four highest ranked general information science journals, and the five highest ranked library science journals were selected. Searches for research articles from the years 1998–2000 from the nine journals were made and the result was downloaded

from the *Web of Science, Social Science Citation Index* (*SSCI*), and from the *Resources Information Center Database* (*ERIC*) SilverPlatter 1992–2001/03 edition (Table 1). The search criterion for finding research articles were made by using the document type selection, choosing the "articles" alternative, in *Web of Science*. After downloading the results, the files were merged into one, adding the descriptors from the DX-field in the *ERIC* database to the *SSCI* records, except for the cases where a match could not be made.

Table 1. Number of articles downloaded from *SSCI* and *ERIC* 1998–2000.

Journal name	*SSCI*	*ERIC*	Analyzed
Journal of Documentation	69	54	
Journal of the American Soc. for Info. Sci.	324	278	
Information Processing & Management	124	127	
Journal of Information Science	101	30	
College & Research Libraries	111	104	
Information Technology and Libraries	87	38	
Library Quarterly	38	36	
Library Trends	136	100	
Journal of Academic Librarianship	145	78	
Total	1135	845	797

One of the aims of the journal selection was to avoid the strong emphasis on information science that can be found in for example the White and McCain (1998) material. The reason for using fewer information science related journals is to level the uneven amount of documents in the respective journals. The use of multiple databases also affected the journal selection, where some top ranked information science journals were not indexed in the *ERIC* database. The use of two databases also led to a decrease in documents to analysis, since there are some discrepancies in the number of documents indexed in the two databases. The analyses are based on the documents where a match could be made between the *SSCI* and *ERIC* records. The reason for combining the two databases is of course that *SSCI/ISI* does not add keywords to all the records in their databases. Problems concerning completeness in coverage in different databases, and the characteristics and use of the index terms in the case of LIS, have been discussed by authors such as Harter, Nisonger and Weng (1993).

The three different bibliometric analyses are co-occurrence of keywords, co-citations and a merged citation and keyword co-occurrence analysis. The short time span of citing journals made the co-citation analysis seem more appropriate—even though author co-citation analyses are usually preferred for scientific mapping—to get a more recognizable set of cited authors. The analysis was made by coupling co-occurring keywords and/or authors. These couples

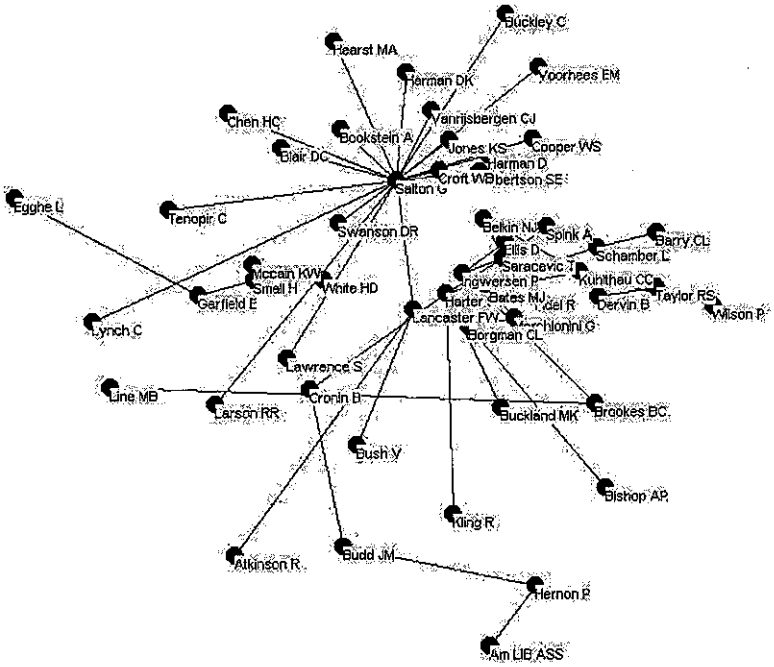
form a matrix that can be visualized by mapping techniques such as multidimensional scaling (MDS). By using the *Bibexcel* software, the downloaded data is processed, analyzed according to the different bibliometric analyses, formed into a matrix, and visualized in the form of a map through MDS. The analysis is further clarified by the use of a clustering routine suggested by Persson (1994).

The maps can be qualitatively analyzed by comparing the different structures formed by the different analyses. Comparisons can be made in terms of how authors or research areas relate to each other, and finding different places on the map. The third analysis, the merged map of citation and keywords, is also a way to compare the behavior of the different forms of indexing, seeing if authors and keywords are placed in a way that corresponds with what topics the authors are writing about. These comparisons of the maps can also be amplified by cluster analysis, to make the structures within the maps clearer. This makes it easier to compare the maps, since the clusters enhance the main trends in the maps.

4. RESULTS

For the first part of the analysis, the 52 most cited authors were selected and coupled. These couples formed a matrix that was processed through a MDS routine, resulting in a map containing the authors and their relations to each other (Figure 1).

Figure 1. Co-citation map of nine LIS journals 1998–2000. Included are the 52 most cited authors.



The couples were also clustered, to help form the lines connecting the names on the map, and amplify the structure of the map. Three clusters were formed (Table 2).

Table 2. Clusters formed by the 52 most cited authors in nine LIS journals 1998–2000.

Cluster1: Hard IR	Cluster2: Soft IR	Cluster3: Bibliometrics
SALTON G	HARTER SP	SMALL H
LANCASTER FW	SARACEVIC T	WHITE HD
BLAIR DC	BATES MJ	GARFIELD E
BUCKLEY C	KUHLTHAU CC	EGGHE L
BUSH V	INGWERSEN P	MCCAIN KW
CHEN HC	SCHAMBER L	
COOPER WS	HERNON P	
CROFT WB	BELKIN NJ	
HARMAN D	BUDD JM	
HARMAN DK	CRONIN B	
HEARST MA	DERVIN B	
JONES KS	BROOKES BC	
BOOKSTEIN A	MARCHIONINI G	
LARSON RR	BORGMAN CL	
LAWRENCE S	BARRY CL	
LYNCH C	ELLIS D	
ROBERTSON SE	KLING R	
ATKINSON R	AM LIB ASS	
SWANSON DR	LINE MB	
TENOPIR C	FIDEL R	
VANRIJSBERGEN	BUCKLAND MK	
VOORHEES EM	BISHOP AP	
	SPINK A	
	TAYLOR RS	
	WILSON P	

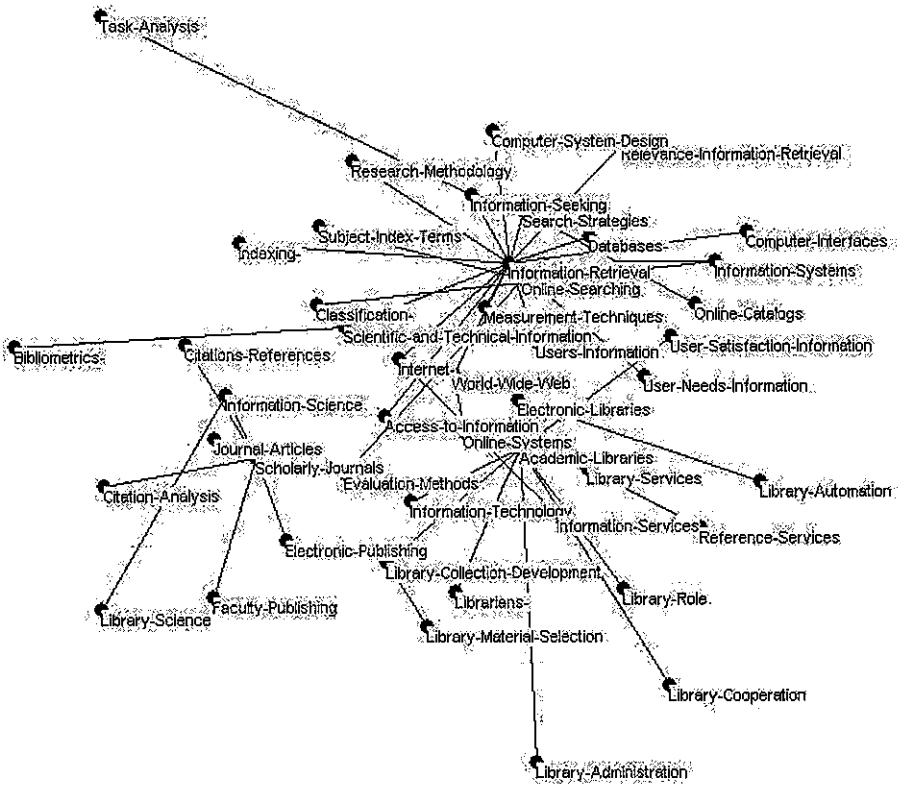
The structure of the map is essentially the same as in Persson (1994; 2001) and White and McCain (1998). Distinguishable groups are the two sub-fields, information retrieval and bibliometrics, and aspects of literatures and communication. The IR field can be further divided into one "hard" area centered around Salton, dealing with algorithm and system development; and one "soft" area dealing with user-system interaction. Here the center is not as obvious, but key figures are Harter, Bates, Saracevic, Kuhlthau and Ingwersen. These areas are also supported by the clusters.

A closer look at the map also reveals smaller groups not noticeable in the clusters. This raises questions as to whether they actually form their own research areas, or remain within the general structure because of the small amount of representation. One is the group that includes Dervin and Kuhlthau, where the question can be raised if the "information needs and uses" research is really part of the IR research, or an independent research area. Another group, perhaps of stronger importance, is one with for example Hernon and Budd, representing a

library science oriented research area. Library science is only connected to the IR area by links, while the location on the map suggests separation from the IR research. That this area does not show in the clustering is probably due to the fact that library science authors are represented by so few names that they cannot form a cluster on their own. A further consideration, in comparison with the White and McCain analysis, is the differences in names of cited authors. Some, such as the inclusion of Peter Ingwersen, are probably due to the different time span. But other names, such as John M Budd, are probably related to the inclusion of library science journals.

The second part of the analysis followed the same basic procedure as the first. Instead of cited authors, the 47 most frequent occurring keywords were selected and coupled, resulting in a map containing the keywords and their relations to each other (Figure 2).

Figure 2. Co-occurrence map of *ERIC* keywords from nine LIS journals 1998–2000. Included are the 47 most frequent occurring keywords.



Here, the couples were also clustered (Table 3). In this map, the separation between different areas is not as clearly distinguishable as with the cited authors. But the same structures and areas can still be found, with the same location on

the map. What can be noted is the remaining sub-area of "information needs and uses" keywords being somewhat separated from the rest of the IR area, while the IR area in general seems to have become less divided and the library science area is more visible.

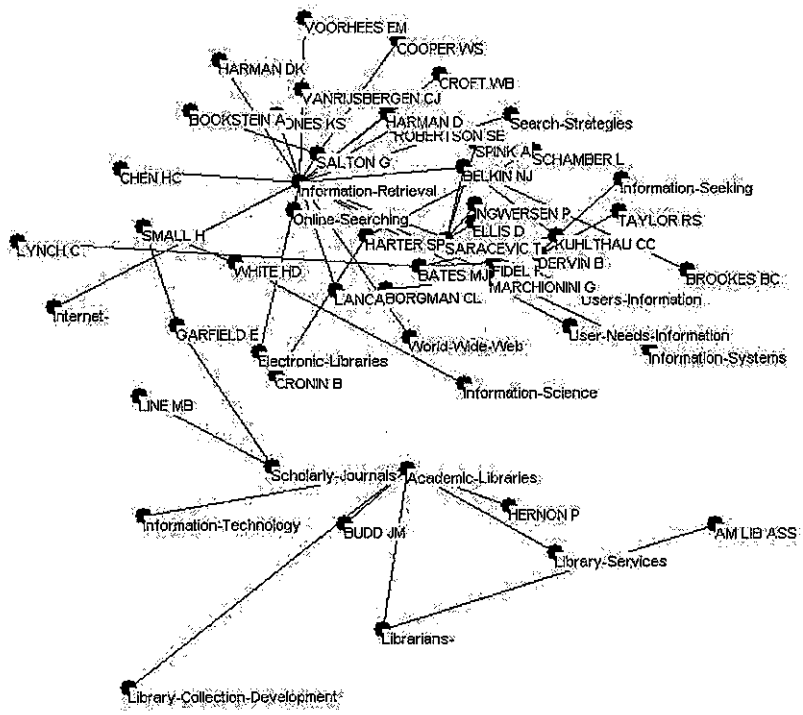
Table 3. Clusters formed by the 47 most frequent occurring ERIC keywords in nine LIS journals 1998–2000.

Cluster 1: Library Sci.	Cluster 2. IR	Cluster 3: Bibliometrics
Academic-Libraries	Information-Retrieval	Scholarly-Journals
Electronic-Libraries	Online-Searching	Information-Science
Library-Services	Internet-	Electronic-Publishing
Libr.-Collection-Dev.	Information-Systems	Faculty-Publishing
Library-Administration	Sci.-and-Tech.-Info.	Citation-Analysis
Library-Automation	World-Wide-Web	Journal-Articles
Librarians-	Information-Seeking	Library-Science
Library-Cooperation	Search-Strategies	Citations-References
Libr.-Material-Sel.	Classification-	
Library-Role	Evaluation-Methods	
Info.-Technology	Information-Services	
Reference-Services	Computer-Interfaces	
User-Satisfaction-Info.	Bibliometrics-	
	Measurement-Tech.	
	Online-Catalogs	
	Indexing-	
	Online-Systems	
	Relevance-Info.	
	Res.-Methodology	
	Comp.-System-Design	
	Access-to-Information	
	Subject-Index-Terms	
	Task-Analysis	
	User-Needs-Info.	
	Users-Information	
	Databases-	

The trends of the map are also supported by the clustering, where the IR related keywords have merged into one cluster, and the library science keywords have formed a separate cluster. The merging of the IR field can probably be explained by the indexing. Although the articles deals with different aspects of IR, the similarities are big enough in terms of subject description for them to form one cluster instead of two. More interesting to note is how library science has evolved into what could be described as a sub-field of its own within LIS. Although some reservations should be made for the quality of the indexing (e.g., Harter, Nisonger, and Weng, 1993), the different and wider journal selection has important implications for how LIS is perceived.

In the third part of the analysis the keywords and citations were merged and ranked, and the 53 most frequently occurring keywords and authors were selected, coupled, mapped and clustered (Figure 3). The structure of this map is basically the same as in the former two analyses.

Figure 3. Co-occurrence map of *ERIC* keywords and cited authors from nine LIS journals 1998–2000. Included are the 53 most frequent occurring keywords and authors.



The clustering also gives the same results as the keyword based analysis, with three main areas: library science, IR and bibliometrics (Table 4). On the map, the distinction between "soft" and "hard" IR is still visible, even though the clustering does not show the same trends, and the "information needs and uses" sub-field is also slightly distanced from the main IR area. The main difference is that the library science area is now clearly separated from the IR field, both in location and by links. Another, very interesting, feature is how well the keywords and the authors correspond to each other. The connection between authors and keywords is very close in terms of actual research interests of the authors.

5. DISCUSSION

The starting point of this paper was the general, and perhaps obvious, statement that the outcome of bibliometric analyses are dependent on the presentation of the data, and what kind of data is being analyzed. The first aim of the paper was to compare and analyze different methods of indexing to see how they affect bibliometric analyses of LIS. This aim raised questions about which different types of bibliometric analyses can be used for describing cognitive structures and what differences can be found in different kinds of bibliometric

analyses. This was done by first producing separate maps containing either citations or keywords, and then a map with both. The results indicate quite strong relations between citations and keywords. The two separate maps formed the same basic structures, identifying the same central areas in LIS: bibliometrics, information retrieval and a library science oriented sub-field. These tendencies were also supported by the close fit when mapping and clustering combinations of cited authors and keywords. The respective location on the map largely corresponds to what topics the cited authors have addressed in their writings, which is also evident in the clustering.

Table 4. Clusters formed by the 53 most frequent occurring *ERIC* keywords and cited authors in nine LIS journals 1998–2000.

Cluster 1: IR	Cluster 2: Library Sci.	Cluster 3: Bibliometrics
Information-Retrieval	Academic-Libraries	GARFIELD E
SARACEVIC T	Librarians-	SMALL H
KUHLTHAU CC	AM LIB ASS	WHITE HD
BATES MJ	HERNON P	Scholarly-Journals
BELKIN NJ	Info.-Technology	Information-Science
HARTER SP	BUDD JM	LINE MB
SALTON G	Libr.-Coll.-Dev.	
DERVIN B	Library-Services	
MARCHIONINI G		
BROOKES BC		
Electronic-Libraries		
ELLIS D		
CROFT WB		
HARMAN D		
HARMAN DK		
CRONIN B		
BORGMAN CL		
Information-Seeking		
Information-Systems		
INGWERSEN P		
Internet-		
JONES KS		
FIDEL R		
LANCASTER FW		
LYNCH C		
BOOKSTEIN A		
Online-Searching		
ROBERTSON SE		
CHEN HC		
COOPER WS		
SCHAMBER L		
Search-Strategies		
SPINK A		
TAYLOR RS		
User-Needs-Info.		
Users-Information		
VANRIJSBERGEN CJ		
VOORHEES EM		
World-Wide-Web		

The only significant variation detected in the material was in the clustering: the cited author analysis formed three clusters containing "hard" IR, "soft" IR and bibliometrics; whereas the keywords and the keywords plus author analysis formed three clusters containing IR, bibliometrics and library science. This is probably caused by the author selection, where library science authors were under-represented in the citation based ranking. One reason for this might be publication patterns in library science, where a lot of the research is published in books, and also in regional or national journals, channels of communication that are not covered by, for instance, *SSCI*. It should however be noted that in the keyword based map, the different research areas are not as clearly defined as in the citation and citation/keyword maps. And it should be kept in mind that the citation analysis is based on the documents per se, which makes the connection to the research itself closer, while the keywords are created to represent texts and make them searchable. This can to some extent explain the dissappearance of the distinction between "hard" and "soft" IR. While the distinction is visible as research areas, keywords for IR purposes do not need to make the same distinction.

What is interesting with the relation between the author and keyword maps and clusters is of course that they show such a similar structure. This would suggest that keywords are a good starting point for scientific mapping, or that they at least can serve as complimentary to the co-citation maps. That they in some way reflect research is quite obvious, but the results here support the premise for analyzing co-occurrence of keywords, that similar documents are represented by the same keywords or indexing terms (Whittaker, 1989), and thus, that statistical aggregates of keywords also represent cognitive spaces. Another interesting aspect of this is the analyses of the keyword-citation relations in information retrieval situations (e.g., McCain, 1989; Pao, 1993; Pao, and Worthen, 1989), suggesting a complimentary role with a relatively small overlap. The maps in this analysis suggest a closer relation between citations and keywords than seen in those studies.

The second aim of this paper was to investigate if and how a changed journal selection would change the perceived structure of LIS. This was done in comparison with the White and McCain (1998) analysis. The analysis showed some substantial differences in comparison with White and McCain, introducing library science as a sub-field of LIS on the level with IR and bibliometrics. While the White and McCain analysis is based on a journal selection with a strong emphasis on information science, this analysis sets out to make a more balanced selection of information science and library science journals. The findings in the analysis suggest that the journal selection does affect how research fields can be perceived and defined. When changing the journal selection, the analysis shows a different structure of the field. This is important, since studies such as the White and McCain analysis has been used as empirical basis for further discussion on the nature and cognitive structure of LIS (e.g., Saracevic, 1999).

The selection of nine journals and a time span of only three years is obviously not enough to draw any wider conclusions on the cognitive structure of LIS. But it has been useful for making some obsverations on the usefulness of

the methodology suggested. The selection of both library science and information science journals shows differences in comparison with earlier mappings of LIS, highlighting journal selection as an important factor when mapping research fields. This can be developed further. For instance, even though there are five library science journals and only four information science journals, the number of articles analyzed still shows a majority of information science research articles. Another issue of methodological interest is differing citation behavior in library and information science, respectively. This has implications for the results of the analyses: when citing a lot of diverse sources the internal links get weaker and it gets harder to identify a distinct set of highly cited authors. A third issue is the relation between citations and keywords, two different concepts with different functions. The differences are visible when comparing the two maps based on either citations or keywords, where the structure of LIS is more clearly defined in the citation based map. While citations represent some sort of cognitive relation between documents, keywords are representations of documents assigned to them for information retrieval purposes. One way of solving this problem, or at least to come closer to the actual documents, is by using keywords derived directly from the text. Despite these problems, when merging the keywords and citations into one map they largely correspond with each other, resulting in a map that is easier to interpret than a map showing only author names, for example.

REFERENCES

Braam, R. R., H. F. Moed, and A. F. J. van Raan. (1991a) Mapping of science by combined co-citation and word analysis: I: Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–51.

Braam, R. R., H. F. Moed, and A. F. J. van Raan. (1991b) Mapping of science by combined co-citation and word analysis: II: Dynamical aspects. *Journal of the American Society for Information Science, 42*(4), 252–66.

Callon, M., J.-P. Courtial, and F. Laville. (1991) Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 153–203.

Courtial, J.-P. (1994) A co-word analysis of scientometrics. *Scientometrics, 31*(2), 251–60.

Courtial, J.-P., and J. Law. (1989) A co-word study of artificial intelligence. *Social Studies of Science, 19*, 301–11.

Hahn, T. B., and M. K. Buckland, eds. (1998) *Historical studies in Information Science.* Medford, NJ: Information Today.

Harter, S. P., T. E. Nisonger, and A. Weng. (1993) Semantic relationships between cited and citing articles in Library and Information Science journals. *Journal of the American Society for Information Science, 44*(9), 543–52.

Hjørland, B. (2000) Library and Information Science: Practice, theory, and philosophical basis. *Information Processing & Management, 36*, 501–31.

Ingwersen, P. (1992) Information and Information Science in context. *Libri, 42*(2), 99–135.

Law, J., and J. Whittaker. (1992) Mapping acidification research: A test of the co-word method. *Scientometrics, 23*(3), 417–61.

Leydesdorff, L. (1997) Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science, 48*(5), 418–27.

McCain, K. W. (1986) Cocited author mapping as a valid representation of intellectual structure. *Journal of the American Society for Information Science, 37*, 111–22.

McCain, K. W. (1989) Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science, 40*(2), 110–14.

Pao, M. L. (1993) Term and citation retrieval: A field study. *Information Processing and Management, 29*(1), 95–112.

Pao, M. L., and D. B. Worthen. (1989) Retrieval effectiveness by semantic and citation searching. *Journal of the American Society for Information Science, 40*(4), 226–35.

Persson, O. (1994) The intellectual base and research fronts of JASIS 1986–1990. *Journal of the American Society for Information Science, 45*(1), 31–38.

Persson, O. (2001) All author citations versus first author citations. *Scientometrics, 50*(2), 339–44.

Saracevic, T. (1999) Information Science. *Journal of the American Society for Information Science, 50*(12), 1051–63.

Small, H. G., and B. C. Griffith. (1974) The structure of scientific literatures: I: Identifying and graphing specialties. *Science Studies, 4*, 17–40.

Small, H. G., and E. Sweeney. (1985) Clustering the Science Citation Index. *Scientometrics, 7*, 391–409.

Vakkari, P. (1994) Library and Information Science: Its content and scope. In I. P. Godden (ed.). *Advances in librarianship: Volume 18* (pp. 1–55). San Diego, CA: Academic Press.

White, H. D., and B. C. Griffith. (1981) Author cocitation: A literature measure of intellectual structure, *Journal of the American Society for Information Science, 32*, 163–72.

White, H. D., and K. W. McCain. (1998) Visualizing a discipline: An author co-citation analysis of information science, 1972–1995, *Journal of the American Society for Information Science, 49*(4), 327–55.

Whittaker, J. (1989) Creativity and conformity in science: Titles, keywords, and co-word analysis. *Social Studies of Science, 19*, 473–96.