

## Using global mapping to create more accurate document-level maps of research fields

Richard Klavans<sup>a</sup> & Kevin W. Boyack<sup>b</sup>

<sup>a</sup> SciTech Strategies, Inc., Berwyn, PA 19312 USA ([rklavans@mapofscience.com](mailto:rklavans@mapofscience.com))

<sup>b</sup> SciTech Strategies, Inc., Albuquerque, NM 87122 USA ([kboyack@mapofscience.com](mailto:kboyack@mapofscience.com))

### Abstract

We describe two general approaches to creating document-level maps of science. To create a local map one defines and directly maps a sample of data, such as all literature published in a set of information science journals. To create a global map of a research field one maps ‘all of science’ and then locates a literature sample within that full context. We provide a deductive argument that global mapping should create more accurate partitions of a research field than local mapping, followed by practical reasons why this may not be so. The field of information science is then mapped at the document level using both local and global methods to provide a case illustration of the differences between the methods. Textual coherence is used to assess the accuracies of both maps. We find that document clusters in the global map have significantly higher coherence than those in the local map, and that the global map provides unique insights into the field of information science that cannot be discerned from the local map. Specifically, we show that information science and computer science have a large interface and that computer science is the more progressive discipline at that interface. We also show that research communities in temporally linked threads have a much higher coherence than isolated communities, and that this feature can be used to predict which threads will persist into a subsequent year. Methods that could increase the accuracy of both local and global maps in the future are also discussed.

### Introduction

There is a relatively long tradition in creating document-level maps of different areas of science. Although there are many ways to categorize such studies, for the purposes of this article we use the notions, introduced by Small (1997, 1999), of ‘local maps’ and ‘global maps’. Local maps are created by identifying and mapping a particular target literature, typically defined at the field or specialty level. Once the sample of literature is obtained, one can choose to map the source documents using, for example, bibliographic coupling or co-word analysis, or the reference documents using co-citation clustering. In any case one typically creates a square relatedness matrix that shows the relationship between the documents, and then generates a layout of the documents (i.e., the map) from that relatedness matrix. Further work is often done to assign documents to clusters in order to provide a more legible and understandable map. Alternately, documents are first assigned to clusters, and then a visual layout or map is created showing the relative positions and relationships between clusters.

By contrast, global maps seek to map the global structure of science rather than a field-based sample; the goal is to map all scientific concepts. Global maps are created using the same mapping methodologies – co-citation clustering, bibliographic coupling, text analysis, etc. – that

are used for local maps. Ideally, a global map would not use a subsample of a database (such as the Web of Science or Scopus), but in practice global maps are often limited to single-year samples. Once the global map is created, the particular target literature can be located within this broader context of all scientific concepts. Small (1981) provides perhaps the earliest example of how one might generate a global map of a particular field – in his case, information science.

In this article we first review relevant literature, and then present the claim that global maps should inherently have more accurate partitions or clusters than local maps, providing a thought experiment to illustrate why this should be the case. For balance, we also provide methodological reasons why a local map might be more accurate than a global map in practice. We then test our claim by creating both local and global maps of information science using Scopus data, and comparing the accuracies of the two maps using a textual coherence measure. We conclude by pointing out the unique insights gained from a global map and discuss future directions for improving the accuracy of document-level maps of science.

## Maps of information science

**Local maps:** Local maps form a majority of the science maps that have been published. Many such maps have been generated of the field of information science, most likely because it is easier to interpret or evaluate the validity of a map in one's own field than in another (Persson, 1994; White & McCain, 1998). Among the earliest of these studies was that by White & Griffith (1981) in their introduction of author co-citation analysis (ACA). Using a set of 39 authors, they found 5 major groupings within Information Science – *scientific communication*, *bibliometrics*, *generalists*, *information retrieval*, and *precursors*. Persson (1994) performed a similar analysis of 62 authors several years later, but also mapped the research front (51 articles from JASIS) using shared cited authors (a form of bibliographic coupling) as the basis for mapping of current documents. He found that both maps were quite similar conceptually to the earlier map of White & Griffith, and showed the two major branches of information science – *bibliometrics* and *information retrieval*, where the *bibliometrics* branch combined the *communication* and *bibliometrics* groups from the White & Griffith map. White & McCain (1998), in their landmark ACA study, solidified the two-camp view of information science (called *retrievalists* and *citationists*), while enumerating 12 specialties within the field.

Nearly all of the studies that have mapped information science have done so from the starting point of a list of information science journals. White & McCain (1998) used a set of 12 journals, and the majority of studies since that time have used the same list of journals, give or take one journal (Chen, Ibekwe-SanJuan, & Hou, 2010; Klavans, Persson, & Boyack, 2009; Persson, 2001, 2010; White, 2003; Zhao & Strotmann, 2008a, 2008b, 2008c). Two studies have used a subset of the 12 journals – Janssens et al. (2006) use five, and van den Besselaar & Heimeriks (2006) use seven and add the *Canadian Journal of Information Science*. Two other studies use the bulk of the White & McCain journal set, and add another 9 journals each (Åström, 2007; Moya-Anegón, Herrero-Solana, & Jimenez-Contreras, 2006). While some of these studies map authors and others map documents (with one mapping journals as well), and while each finds a different number of detailed specialties, their results all correlate well with the view that there are two dominant camps in information science – the information seeking/retrieval camp, and the citation/bibliometrics/informetrics camp. There are some caveats to this generalization. The

author pair co-citation study by Klavans et al. (2009) showed information seeking and information retrieval as separate camps. Several studies (Åström, 2007; Janssens, Leta, et al., 2006; van den Besselaar & Heimeriks, 2006) show the emergence of web-based studies or webometrics within the informetrics camp, but have not shown that topic to have fully emerged as a separate camp.

Cronin & Meho (2008) opine that most such studies are too small in scale, being limited in time window, specialty or sub-discipline, or number of works examined. They counter this situation by using a set of 275 information science journals to explore the exports from information science to other fields. A smaller set of 80 journals was used to examine imports from other fields. They find that *Lecture Notes in Computer Science* (and computer science in general) is a heavy importer from information science, and that information science is drawing more on the computer science and management disciplines with the passage of time. However, based on their numbers, information science still seems a rather import-resistant discipline – 105 of the 204 top cited journals from 1997-2006 are from within the discipline.

Chen et al. (2010) is the most recent and comprehensive exemplar of a local map of information science, combining ACA and a document co-citation analysis (DCA) with advanced clustering metrics and labeling capabilities to give multiple perspectives enabling a more robust interpretation of the structure of a field. Their static ACA map (120 authors, 2001-2005) compared very favorably to the results of a similar map by Zhao & Strotmann (2008c). Their progressive (time-dependent) ACA clustered 633 authors into 40 clusters, while their progressive DCA clustered 655 cited references into 50 clusters.

**Global maps:** Few have attempted to map all of science at the document level. Earlier work at the Institute for Scientific Information (Griffith, Small, Stonehill, & Dey, 1974; Small, 1981, 1999; Small, Sweeney, & Greenlee, 1985) and the Center for Research Planning (CRP) (Franklin & Johnston, 1988) mapped all of science in a representative fashion using thresholded document sets. Recent work at Sandia National Laboratories and SciTech Strategies has created maps of complete file years (all citing documents) of the ISI (Boyack, 2009; Klavans & Boyack, 2006) and Scopus (Klavans & Boyack, 2010) databases.

Only one global map of information science has been reported in the literature. Using a three-year set of the SSCI database, Small (1981) generated a representative map of the entire SSCI by 1) limiting the set of cited references to those cited 10 or more times (24,954 references), 2) calculating co-citation frequencies and normalized strengths between pairs of cited references, 3) limiting the pairs to those with a normalized strength of 0.22 or higher, and 4) clustering the remaining pairs. Using this process, a set of 2,095 co-citation clusters was generated. Citing (source) papers were then assigned to the co-citation clusters using their reference lists. These clusters formed the model of the SSCI for his study. Small then identified a set of 50 journals representing information science, and found those co-citation clusters where at least 10% of the source papers came from the 50 journal set, resulting in a set of 22 co-citation clusters that were dominated by or related to information science. The benefit of this approach was that it not only identified topics dominated by information science, but also identified strong links between information science and various topics in the social sciences.

## **A global map should be more accurate than a local map**

We claim that, all other things being equal<sup>1</sup>, a global map including a particular target literature will be inherently more accurate than a local map of that same target literature. Increased accuracy comes from two sources: a more accurate estimator of concept relatedness and a more accurate clustering from the spatial relationships between concepts.

The reasoning behind these two accuracy claims can be illustrated using the following thought experiment. Assume we are in a world that is completely defined by 100 concepts, that we have identified 10 of those concepts that are thought to comprise a particular field or specialty, and that we want to create a map of that field. There are two major options for doing this: one can generate a local map of the 10 concepts, or one can generate a global map of the 100 concepts, and then locate the 10 concepts in the global map. Assume that we do have access to full information for the 100 concepts; thus access to data is not a limiting factor.

For both the local and global maps in this example, we assume the use of co-occurrence values between pairs of concepts, and then form a matrix of normalized values using the cosine index. We also assume the use of a standard layout technique, such as a force-directed algorithm, to provide a 2-D layout of the concept space (i.e., a concept map) based on the relatedness matrix, and will use the spatial locations of the concepts to determine clusters of concepts.

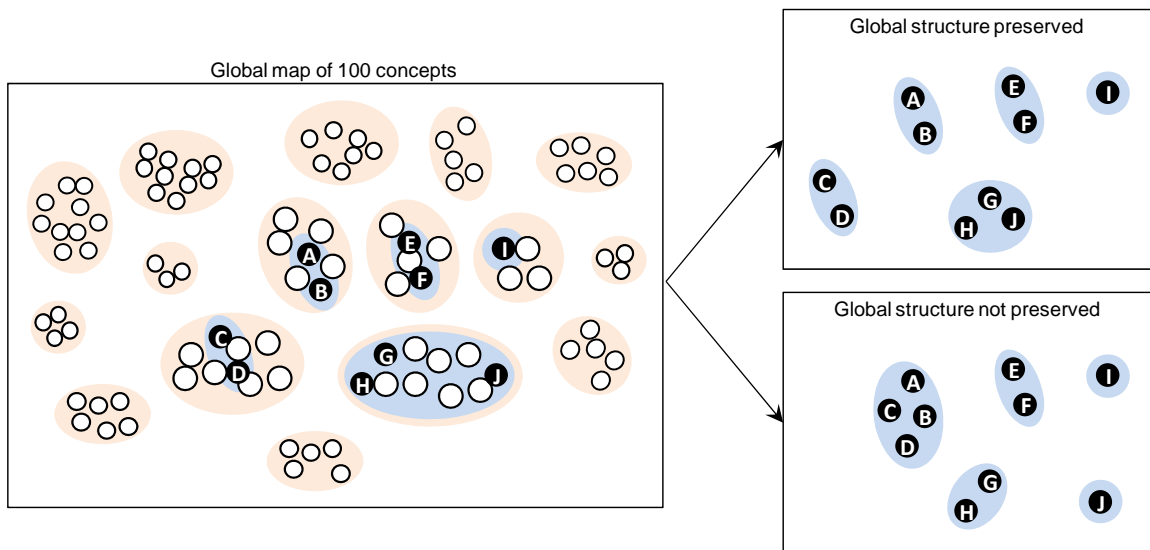
For the case of the global map, we use all of the available data and generate a 100x100 matrix. Each relatedness value in the matrix is based on full information because the row and column sums used in the normalization were based on the full co-occurrence matrix. Using the assumption that full information is always better than partial information when computing the similarity between any two concepts, the full matrix would thus enable calculation of a best estimate for the actual similarity values between two concepts.

For the case of the local map one would identify the 10 concepts to be mapped, and would then generate a 10x10 matrix. Using the same methodology used for the global map, each relatedness value in the matrix would be based on the co-occurrence value between the pair of concepts, and would then be normalized by the row and column sums. For any cell in this local (10x10) matrix, the value is likely to differ from the value in the same cell in the global (100x100) matrix because of two things. First, the co-occurrence values for the local and global cases might be different. For example, in the case of document co-citation, if the source documents are limited to a small journal set, then any co-citations from documents outside that journal set will not be accounted for in the local matrix. This can have a significant effect on the relatedness values, especially if documents are heavily cited from outside the source journal set. Second, the row and column sums in the local matrix will in most cases be far less than those in the global matrix, and will thus impact the relatedness values. Using the assumption that the full matrix gives the best estimate for the actual similarity value between two concepts, any deviation from the global value in the local matrix can be considered as a potential source of error. Using this argument it can clearly be seen that similarity values used in local maps, which are based on sampled or biased subsets of data, can never be more accurate than those used in a global map which is based on full information.

---

<sup>1</sup> Using the same source database, mapping approach, similarity measure, layout algorithm, clustering technique, etc.

This brings us to our second accuracy claim – that the clusters in a global map should inherently be more accurate than the clusters in a local map. Clustering is dependent upon relatedness values as well as upon the number of concepts. Any differences in clustering between a local and global map are thus related to both factors. Figure 1 shows an example of a global map generated from a 100x100 relatedness matrix, along with two different local maps generated from a 10x10 subset of concepts. The global map highlights the locations of the 10 local concepts, which appear in five clusters (AB, CD, EF, GHJ, and I). Of the many possible local maps, we show two. The local map in the upper right shows a case where the relationships between the 10 concepts are the same in the local map as they are in the global map. In this case one could say from a purely visual standpoint that one map is not more accurate than the other. However, the global map does include much more information.



**Figure 1: Local maps can preserve or change the clustering from a global map.**

However, given the differences in relatedness measures between the global and local cases that are likely to occur, as detailed above, it is perhaps more likely that the local map will not preserve the relationships from the global map. The local map at the lower right shows this case. Concepts ABCD have been joined in one cluster in the local map, while the global map shows that with full information they are split into two clusters. In this case the concepts were likely joined in the local map because the information from outside the field that would have placed them in different clusters was not available. The opposite case is also possible, as shown by concepts GHJ. These are joined in the global model, but are split into two clusters in the local model. In this case the local model was missing information from outside the field that linked these concepts more fully.

We posit that local maps will not preserve the structure of local concepts within a global map whenever the relatedness values between the local set (10 concepts in this case) and the environmental set (90 concepts) are greater than the relatedness values within the local set of concepts. This should be true not only at the summed level, but at the individual concept level as well. Each concept that is more related to a concept outside the local map than it is to the

concepts in the local map will likely be clustered differently in the local map than in the global map. Stated differently, one can expect a local map to be less accurate if boundary forces (between the local and environmental concepts) are stronger than core forces (among and between the local concepts). A local map can only maintain the integrity of the structure of a global map if core forces are clearly dominant.

Based on this thought experiment, we therefore argue that, all other things being equal, the clusters in a map based on a biased subsample of the literature can never be more accurate than the clusters associated with that subsample within a global map. A local map cannot be more accurate than a global map if boundary forces are significant. Local maps can be of equal accuracy if the concept relatedness values within the biased sample are perfectly correlated with the relatedness values derived from full information, and if the boundary relationships are weaker than the core local relationships.

### **Why a local map might be more accurate than a global map**

Despite the arguments presented above, there may be practical reasons for expecting a local map to be more accurate than a global map. Most notable are (a) size limitations and (b) database limitations.

***Size limitations:*** The sizes of early science maps were constrained to the capabilities of the computational resources and algorithms that were available at the time. For example, White & McCain (1998) explicitly mention one algorithmic barrier, pointing out that the early maps of information science were limited to 40 or fewer concept symbols due to the limitations imposed by MDSCAL (from SPSS). ALSCAL raised the limit to around 100 concept symbols; White and McCain mapped 120 authors in their 1998 study. Small (1997, 1999), while mapping tens of thousands of documents in his global maps, wrote software that would display around 40 clusters at any particular level and branch within a multi-level cluster solution.

Generation of a complete global map is still out of reach using current resources. Although many researchers use layout and clustering algorithms capable of dealing with tens of thousands of documents, and we use DrL<sup>2</sup> (Martin, Brown, Klavans, & Boyack, 2011) to routinely generate maps of up to 2.5 million documents<sup>3</sup> (Boyack & Klavans, 2010; Klavans & Boyack, 2010), the Thomson Reuters and Scopus databases each contain records on tens of millions of documents. Thus, any global map created today must still rely on filtering or thresholding, both of which are sampling biases. Given the arguments raised in our thought experiment about the effect of sampling bias on accuracy, one could expect that a global map based on thresholding and a local map based on filtering might both be subject to the same sorts of inaccuracies due to bias, and that the local map might be more accurate. Until we have algorithms that are capable of mapping tens of millions of documents, we cannot claim that we are generating a fully specified global map.

---

<sup>2</sup> Sandia has renamed DrL to OpenOrd, which is available at <http://www.cs.sandia.gov/~smartin/software.html>.

<sup>3</sup> DrL running on a desktop PC with 4GB RAM can calculate the layout for a map of up to 2.5 million nodes and 20 million edges without overrunning memory. Nodes and edges both contribute to memory usage. We are unaware of any other system in use in the science mapping community that can work with larger maps. Thus, 2.5 million nodes with 20 million edges can be considered as a current practical limit for the size of science maps.

**Database limitations:** Local maps have an advantage over global maps in that they can collect the literature from a variety of database sources. As such, they can have greater local coverage of a target literature. To date, global maps have only been created using either the Thomson Reuters or Scopus databases. Global maps are therefore limited in local areas by the depth of the literature in a database. For example, it is a well-known fact that computer science and many engineering subfields are dominated by conference papers rather than journal articles (Boyack, 2009). In such cases a local map that includes the conference literature compiled from many data sources might be more accurate than a global map because of the inherent bias in a single database against a particular local area.

## Methodology

To examine our claims about the relative accuracies of local and global maps of science, we generate both local and global maps of the field of information science and then compare the quality of clusters in both maps using a textual coherence measure. Cluster quality based on textual components has been used in many studies as a proxy for accuracy (Boyack & Klavans, 2010; Braam, Moed, & Van Raan, 1991; Glänzel & Czerwon, 1996; Janssens, Quoc, Glänzel, & de Moor, 2006; Janssens, Zhang, De Moor, & Glänzel, 2009; Jarneving, 2007), a tradition we continue here. For the local map we re-create the document co-citation map recently published by Chen, Ibekwe-SanJuan & Hou (2010) using Scopus data. For the global map we create a new nine-year global model of science and then locate the target literature upon which the local map was based within that model. The processes by which the two maps are generated are detailed below.

Before detailing the map generation processes, it is important to clarify some points regarding methodology. First, both the local map and the global map are generated from Scopus data using co-citation clustering with cosine-normalized relatedness values, thus going a long way toward satisfying the “all other things being equal” basis of the thought experiment above. Second, we wish to emphasize that the full co-citation process can create two sets of document clusters. The first (and foremost) set is clusters of co-cited references, which will be referred to as the *intellectual base*. The second set of clusters contains the current papers that are or can be assigned to these intellectual bases through their reference lists, and are called the *research front*. Co-citation analysis differs from other citation methods (such as bibliographic coupling and direct citation) in that it is often used to create these two sets of document clusters<sup>4</sup>.

**Local map of information science:** Chen et al. (2010) generated a document co-citation map of information science using the following process. First, they identified a set of 12 information science journals to use as the citing source material for their map. This was the same list of journals, with one change, used by White & McCain (1998) in their pioneering study of information science. After downloading the source records for the 12 journals from the Web of Science for the period 1996-2008, they divided the source data into yearly slices and identified the 100 most highly cited references from each of the 13 annual time periods. These highly cited

---

<sup>4</sup> Bibliographic coupling and direct citation identify research fronts directly, but do not identify an intellectual base. An additional step would be necessary (i.e., specifying the intellectual bases that correspond to the research fronts) for these methods to be directly comparable to a co-citation model.

references were then combined into a superset of 655 unique cited references. These 655 references were then mapped using the CiteSpace software and clustered into 50 clusters, forming the intellectual base referred to by Chen as specialties.

A list of the 655 references by specialty was graciously provided to us by Dr. Chaomei Chen. Rather than recreating Chen's map from scratch, we simply searched for the references in our Scopus data.<sup>5</sup> The *intellectual base* of our local map thus consists of the same 655 references in 50 specialties as reported by Chen et al. (2010). We then assigned current papers (from 2000-2008) that cite these 655 references to the 50 intellectual bases using their reference lists to form the *research fronts* for the local map. Current papers can be fractionally assigned to multiple intellectual bases if their references cite to more than one base. Each specialty in the local map thus has an intellectual base and a corresponding research front. From this point forward, we no longer refer to Chen's map, but rather to our re-creation of his local map of information science.

**Global model of science:** To generate a global map of information science, we first generate a global model of science, and then create a global map of information science using components from the global model of science. Each of these processes is explained below.

Figure 2 shows the major steps used to build a nine-year global map of science. The data from each source publication year is considered separately so that annual models can be built for each year. To create an annual model from a single year's data, we use the following method:

- An age-dependent threshold is applied to identify a set of highly cited references. A relatively low threshold is used – references are included if cited five or more times in a single year. Recent references have a lower threshold – those published within 3 years must have at least  $age+1$  cites, while those of age 0 must have been cited twice. This threshold maximizes the number of cited references (~2.5M) that are retained in the calculation, subject to the current limits of our clustering approach (footnote 3).
- Co-citation counts are calculated for each pair of the selected references, and a full relatedness matrix is created using the K50<sup>6</sup> measure of relatedness (Klavans & Boyack, 2006).
- Given the current practical limits of the DrL layout algorithm (footnote 3), we filter the relatedness matrix, keeping only the top- $n$  most related references (and their relatedness values) for each reference. We vary the  $n$  in top- $n$  from 5 to 15, and scale its value on  $\log(\text{degree})$ , where degree is the number of other references to which the reference is linked through co-citation.
- We then cluster the references with DrL using the detailed process explained in Boyack & Klavans (2010). Briefly, DrL is run 10 times with different starting points in order to reduce possible errors in final layout. Reference pairs that are highly proximate in 6 or more of the 10 maps are then selected and clustered. Using this method and criteria the clusters are extremely well defined and one can use single link clustering without

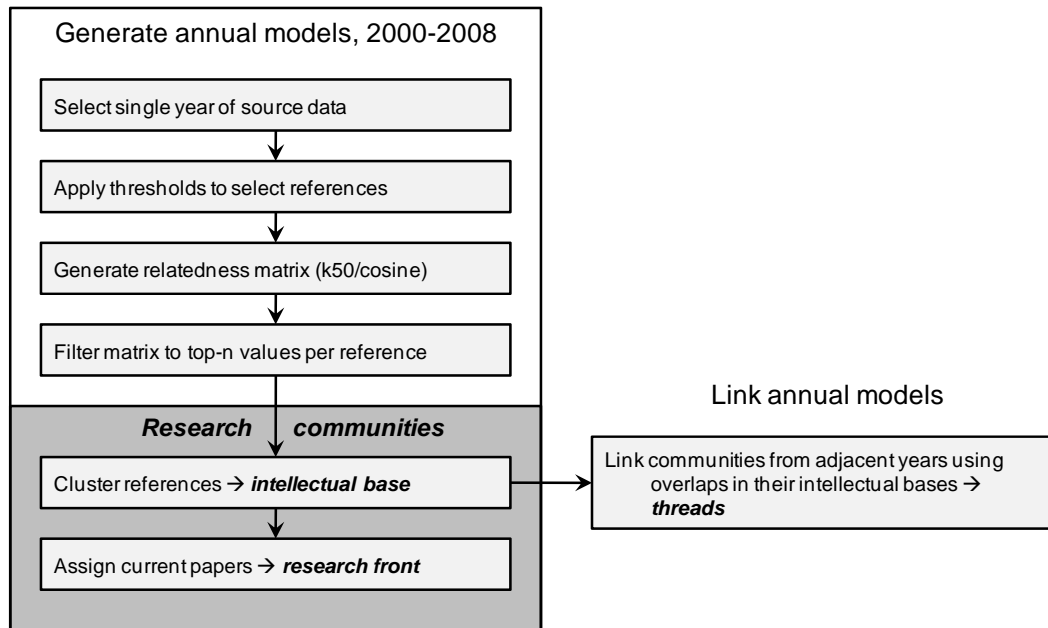
---

<sup>5</sup> We were able to obtain Scopus article IDs for all but 17 references. Of those that we could not locate, 11 were books, and all but 2 were published prior to 1996. Note, however, that we were able to obtain titles for all but 3 references through web searches. Titles are used later in the coherence calculations.

<sup>6</sup> K50 is the cosine index minus the expected value of the cosine index; expected values are so small ( $\sim 10^{-6}$ ) in a large document-document matrix that the cosine index can be used in place of K50.

experiencing chaining effects. These clusters of references represent the *intellectual base* for the annual model.

- Current papers are then fractionally assigned to the intellectual base using their reference lists to form the *research fronts* of the annual model. A single research front and its corresponding intellectual base is called a *research community*.



**Figure 2: Flowchart showing general steps used to create the nine-year global model of science.**

Once the nine annual models have been generated we link the annual models into a multi-year global model of science by linking research communities from adjacent model years together using overlaps in their intellectual bases (references). Each of the annual models contains on the order of 2 million unique references in its intellectual base. Unfortunately, roughly 50% of these are not in the intellectual base of the prior or subsequent annual models. This is almost solely due to the use of a threshold for selecting references. About half of the references that meet the minimum threshold of 5 cites in one year will drop below this threshold in the following year. Almost all, however, are still present but at a lower citation frequency. We temporarily add these ‘missing’ references back into the intellectual bases of the annual models to provide more information on the actual overlaps between the bases from year-to-year.

For example, there may be 2 million references in the intellectual bases of the 2007 and 2008 annual models, of which only 1 million appear in both models. The 1 million references that are in the 2007 model but not in 2008 model are temporarily added to the 2008 model by calculating the relatedness between these 1 million ‘adds’ and the existing 2008 intellectual bases. These references are added in without changing the structure of the 2008 model. A similar procedure is used for the 2007 model. Both models now contain around 3 million references in their bases, nearly all of which are overlapping. This procedure essentially triples the signal that can be used to link models from 1 million to 3 million references.

To link annual models, a simple cosine index was calculated for pairs of intellectual bases in adjacent years using the number of overlapping references, and the numbers of references in the two bases. Bases, and their corresponding communities, were linked if the cosine index was 0.3 or higher.<sup>7</sup> These linked research communities are called *threads*.

Threads provide an insight into the persistence of the intellectual base of a research community or set of communities. For communities of roughly similar size, the 0.3 cosine index threshold can be interpreted as meaning that a community only needs to retain 30% of its intellectual base from year-to-year to persist as a thread. Our nine-year (2000-2008) global model of science consists of 425,000 unique research communities (over the nine years) that are linked into 97,000 threads lasting two years or more. Some of the threads persist over the entire nine year period while others are shorter. In addition, there are 243,000 research communities that could not be linked into threads – they did not have any linkage with a community in an adjacent year with an overlap of 0.3 or higher. These single communities can be thought of as one-year threads or they can be labeled as *isolates*, or discontinued experiments in the realm of scientific endeavor.

The practical limit of mapping up to 2.5 million documents mentioned earlier plays a significant role in how this nine year global model was generated. Rather than selecting 2.5 million references for the entire nine year period and generating a single set of intellectual bases covering all nine years together, we chose to generate annual models, each of which could contain up to 2.5 million references. In effect, this choice reduces the bias in the global map because we are using roughly the top 12% of cited references each year rather than the top 2-3% of cited references over a nine-year period. However, it is not clear that use of the top 12% of cited references is good enough. Although the citation thresholds used here are low, it is very likely that there are many perfectly correlated sets of references below the threshold that, if included, would increase the accuracy of the model. We hypothesize that coherence of a global model will increase as citation thresholds are loosened, but leave testing of that hypothesis to a future study. For the time being we simply do not know how the underlying document relatedness values would change if one were to use the top 1%, top 10%, top 20%, top 50%, or all of the available references.

***Global map of information science:*** Once the global model of science has been generated, a global map of information science is created by a) locating the research communities in the global model that contain the target literature, and b) mapping those research communities. The target literature as specified by Chen et al. (2010) consists of 12 information science journals on the source side and 655 highly cited references that form the intellectual base. Of the 12 journals listed by Chen, only 11 are applicable to the 2000-2008 time frame of this study, and all 11 are available in Scopus (see Table 1).

We identified all research communities in the nine-year global model of science that met at least one of the following criteria:

- at least one of the intellectual base papers are from the 655 target references,
- at least two of the research front papers are from the target literature,
- at least 5% of the research front papers are from the target literature.

---

<sup>7</sup> This threshold was chosen using a large amount of data, scree plots, and thread size distributions. A slightly lower threshold leads to chaining and giant components. A slightly higher value forms few threads.

**Table 1. Source literature used to define the field of information science from Scopus, 2000-2008.**

Citing source title	Abbrev	Records	Refs	Cites	CPP
Annual Review of Information Science and Technology	ARIST	105	17133	1093	10.41
Electronic Library	EL	596	7086	757	1.27
Information Processing and Management	IPM	661	20668	4939	7.47
Information Technology and Libraries	ITL	239	4267	618	2.59
Journal of Documentation	JOD	317	12718	2004	6.32
Journal of Information Science	JIS	439	14544	2049	4.67
Journal of the American Society for Information Science and Technology	JASIST	1322	44124	10088	7.63
Library and Information Science Research	LISR	256	8433	1047	4.09
Library Resources and Technical Services	LRTS	205	6589	410	2.00
Program	PROG	207	2755	334	1.61
Scientometrics	SCM	948	21245	5459	5.76
<i>TOTALS</i>		5295	159562	28798	5.44

1,701 research communities within our global model met at least one of these criteria. 944 of these research communities were isolates. The remaining 757 research communities were in 395 threads. The 395 threads contained additional research communities that did not meet our original criteria for inclusion, but since they were part of the threads, and thus related in that way, they were included. This added an additional 1,005 research communities. Thus, our global map of information science is based on a total of 2,706 research communities, of which 944 are isolates, and the balance (1,762) are in 395 threads.

A global map of information science was created from an analysis of the relationships between the 395 threads mentioned above. We calculated the normalized (cosine index) overlap between the intellectual bases in the threads<sup>8</sup>, doing an initial layout in DrL using a default edge cutting setting, and displaying the resulting reduced similarity file (after DrL had pruned edges) in Pajek using the Kamada-Kawai layout.

**Textual coherence:** In this study we compare the relative accuracies of both the intellectual bases and the research fronts of the local and global models of information science using textual coherence. In a recent study using this measure, Boyack & Klavans (2010) compare the textual coherence of three different citation-based methods – co-citation analysis, bibliographic coupling and direct citation analysis – on a corpus of over 2 million documents. This was a fair test because the text-based method for assessing coherence was independent of the citation-based methods used to cluster documents.

The quantity we use to measure textual coherence is the Jensen-Shannon divergence (JSD), which is used to quantify the distance (or divergence) between two (or more) probability distributions. JSD is calculated for each document from the word probability vector for that document, and from the word probability vector for the cluster in which the document resides as:

<sup>8</sup> Since a reference paper can occur in one community per year, it can be a part of different threads in different years.

$$JSD(p, q) = \frac{1}{2} D_{KL}(p, m) + \frac{1}{2} D_{KL}(q, m)$$

$$\text{where } m = (p+q)/2 \text{ and } D_{KL}(p, m) = \sum (p_i \log (p_i/m_i))$$

and  $p$  is the frequency of a word in a document,  $q$  is the frequency of the same word in the cluster of documents, and  $D_{KL}$  is the well-known Kullback-Leibler divergence. JSD is calculated for each cluster as the average JSD value over all documents in the cluster.

JSD is a divergence measure, meaning that if the documents in a cluster are very different from each other, using different sets of words, the JSD value will be very high, or close to 1.0. Clusters of documents with similar sets of words – a less diverse set of words – will have a lower divergence. JSD also varies with cluster size – larger clusters will naturally be more divergent than smaller clusters. We normalize for this by calculating JSD for random samples.<sup>9</sup> The coherence value for cluster  $i$  is defined as:

$$Coh_i = JSD(rand)_i - JSD(actual)_i$$

where  $JSD(rand)$  is the random divergence for the particular cluster size. The average coherence value for an entire map is then calculated as a weighted average

$$Coh = \sum n_i * Coh_i / \sum n_i .$$

summed over all clusters  $i$  where  $n_i$  is the number of items analyzed in cluster  $i$ .

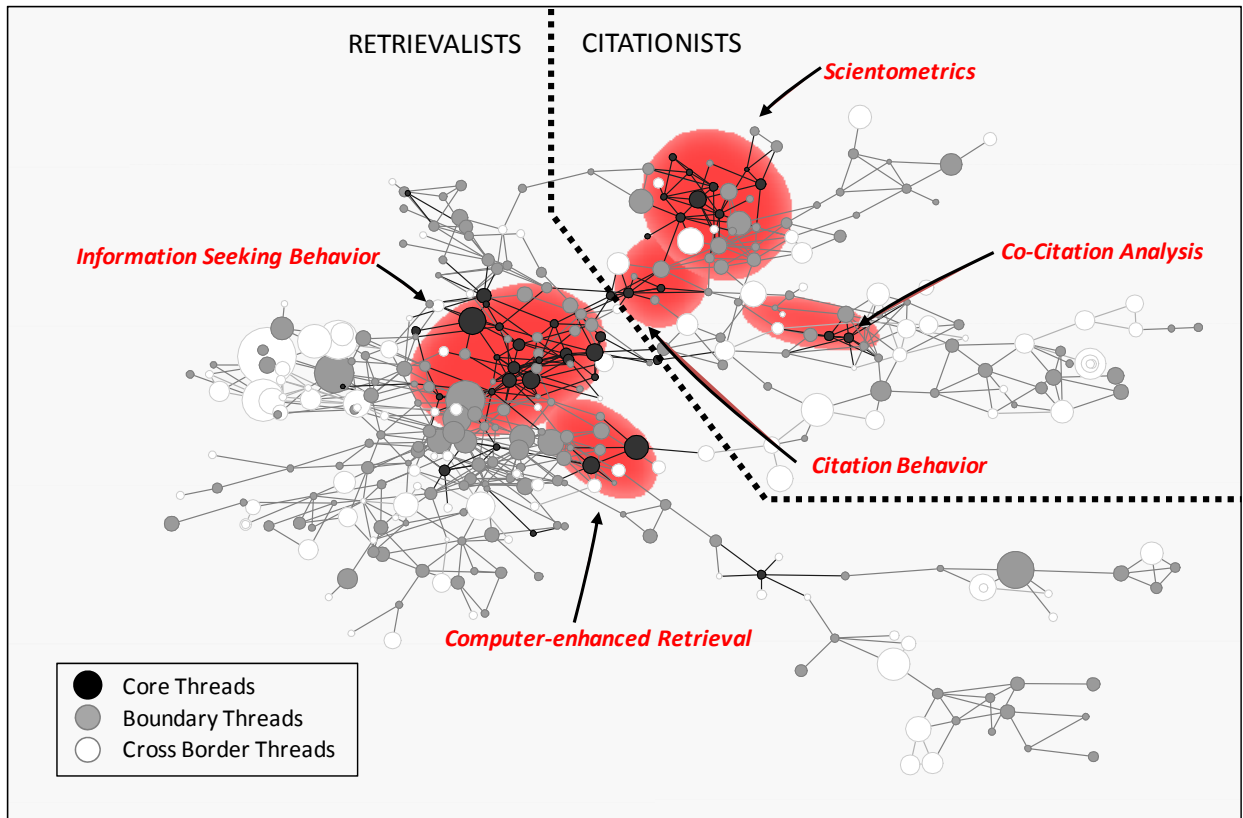
## Results

**Global map of information science:** Figure 3 shows the major network relationships between the 395 threads in the global map of information science where each node is a thread. Node sizes reflect the total number of research front papers by thread. We identify three different types of threads in this map, each of which is indicated by different shading: *core* threads are those whose research fronts are comprised of at least 40% target literature, *boundary* threads are those with 10-40% of their research front from the target literature, and *cross-border* threads are those with less than 10% membership in their research fronts from the target literature. These distinctions are used both in our descriptions of the global map and in the coherence analysis given later.

**Core areas:** Two broad areas containing networks of core threads are highlighted and separated by a dashed line in Figure 3. The lower left area deals with information retrieval, while the upper right deals with citation analysis. The map shows a clear separation between the two areas, which is consistent with the two-camp view that has been highlighted in many local maps of information science; we label the two camps as *retrievalists* and *citationists* in keeping with the labeling used in White & McCain (1998).

---

<sup>9</sup> In principle one should normalize with the maximum entropy rather than the random entropy for a particular cluster size. However, for very large populations (e.g. 2,000,000 articles) there is little difference between the two for small random samples (e.g. 10 or 100 articles). Thus we use the easily calculable random coherence here.



**Figure 3: A global map of 395 threads related to information science. Five core areas are highlighted.**

We have also labeled the substructures within these two core areas. Labeling of these areas was based on three separate analyses: correspondence with the intellectual bases of specialties in the local map, term ranking by thread, and journal ranking by thread. Correspondence was based on matching the references in the intellectual base of the specialties to the cumulative intellectual bases of the threads. Most of the local map specialties could be unambiguously assigned to one of the core areas in Figure 3. For phrases, two-word phrases were extracted from the titles and abstracts of the research front papers in each thread and ranked using log-likelihood ratios. Journal ranking was based on the summed totals of fractionalized paper assignments by thread. Matched specialties, terms, and journals from each of the core areas of Figure 3 are shown in Table 2.

Within the retrieval area, information seeking behavior is the larger network (with more core threads), and computer-enhanced retrieval is the smaller network with just two relatively large core threads and a number of boundary threads. Journals from the target literature dominate the information seeking behavior area; however, they do not dominate the computer-enhanced retrieval area. This area is dominated by computer science, and by one source title in particular, *Lecture Notes in Computer Science* (LNCS). We note that LNCS is also the #2 publication in the information seeking behavior area. This one source appears to have a significant overlap of research interests with the discipline of information science. We will return to this observation in a subsequent discussion.

**Table 2. Labels for core areas in Figure 3. Numbers in parentheses show the number of papers per journal; journals marked with an asterisk are from the target journal set.**

Core area	Specialties	Terms	Journals
Information seeking behavior	18 – User information problems during interactive information retrieval 43 – Academic web 12 – Classroom resources 15 – Using everyday life information 17 – Information behavior 14 – Info science philosophical bias 45 – Open access 41 – Reading scholarly literature 9 – Dialectic approach / activity theory	Information seeking Information retrieval Search engine Digital libraries Information sources Information literacy	(189) – JASIST * (169) – LNCS (87) – IPM * (76) – Proc. ASIST (68) – JOD * (57) – SCM
Computer-enhanced retrieval	13 – Information retrieval / probabilistic models 11 – Web search 8 – Noun phrasing 47 – Contextual media 16 – Multitasking interplay 6 – Cross-language 25 – Automatic survey coding	Information retrieval Search engines Natural language Machine learning Web pages Query expansion	(850) – LNCS (109) – IPM * (101) – JASIST * (92) – LNAI (85) – Int. Conf. Inf. Knowl. Management (CIKM) (68) – SIGIR Forum
Scientometrics	2 – H index 30 – Power law 21 – Lotka’s law 26 – Journal impact 46 – Socio-bibliometric mapping	Scientific journals Scientific production International journals Citation counts Bibliometric indicators	(167) – SCM * (48) – JASIST * (29) – J INF (26) – LNCS (25) – MED LIBR ASSOC
Co-citation analysis	1 – Co-citation analysis	Multinational corporations Technological knowledge Intellectual capital Life cycle Technological innovation	(61) – J. Intellectual Capital (55) – Research Policy (47) – SCM * (29) – PICMET (25) – RD MANAGE
Citation behavior	35 – Citation behavior	Academic writing Academic discourse Bibliographic Coupling Author co-citation	(14) – SCM * (10) – Knowledge Organization (10) – JASIST * (8) – LNCS

As shown in Table 2, many of the threads in the ‘citationist’ area of Figure 3 correspond to specialties within scientometrics. The co-citation analysis area overlapped very closely with a single specialty of the same name. Yet, while co-citation may be the label used in this area, the phrases and journals imply that the emphasis is much more on the application of these metrics for evaluating R&D, and that co-citation is really an applied technology rather than an area of scientific research.

The third citationist area, citation behavior, is relatively small, and again correlates quite closely with a single specialty of the same name from the local map (specialty 35). We will use this area



there may be threads that start in 2008 that cannot be detected until one builds a model of 2009 and links it back to 2008. The three threads in the gray section of Figure 6 show the remaining threads that are linked to specialty 35. These three threads were all short-lived (two years each), but have a strong overlap in their reference bases and are shown proximate to each other in Figure 4.

Regarding the isolated communities in Figure 4, one can view them as experiments – socio-cognitive groups that form around an intellectual base that is not a continuation of one currently in existence. The experiment is successful (i.e., the thread persists) if the research community is able to establish its definition of the intellectual base on the broader scientific society. Figure 4 shows two research communities in 2000, one that persisted (non-prolific authors) and one that did not (disproportionate citation). We do not know if the isolated research community in 2000 is a continuation of a thread from 1999 since a 1999 model was not generated. There were eight isolated communities between 2001 and 2007; these can be considered as experiments that were not continued. By contrast, there were 11 experiments that succeeded (the eleven threads) to varying degrees.

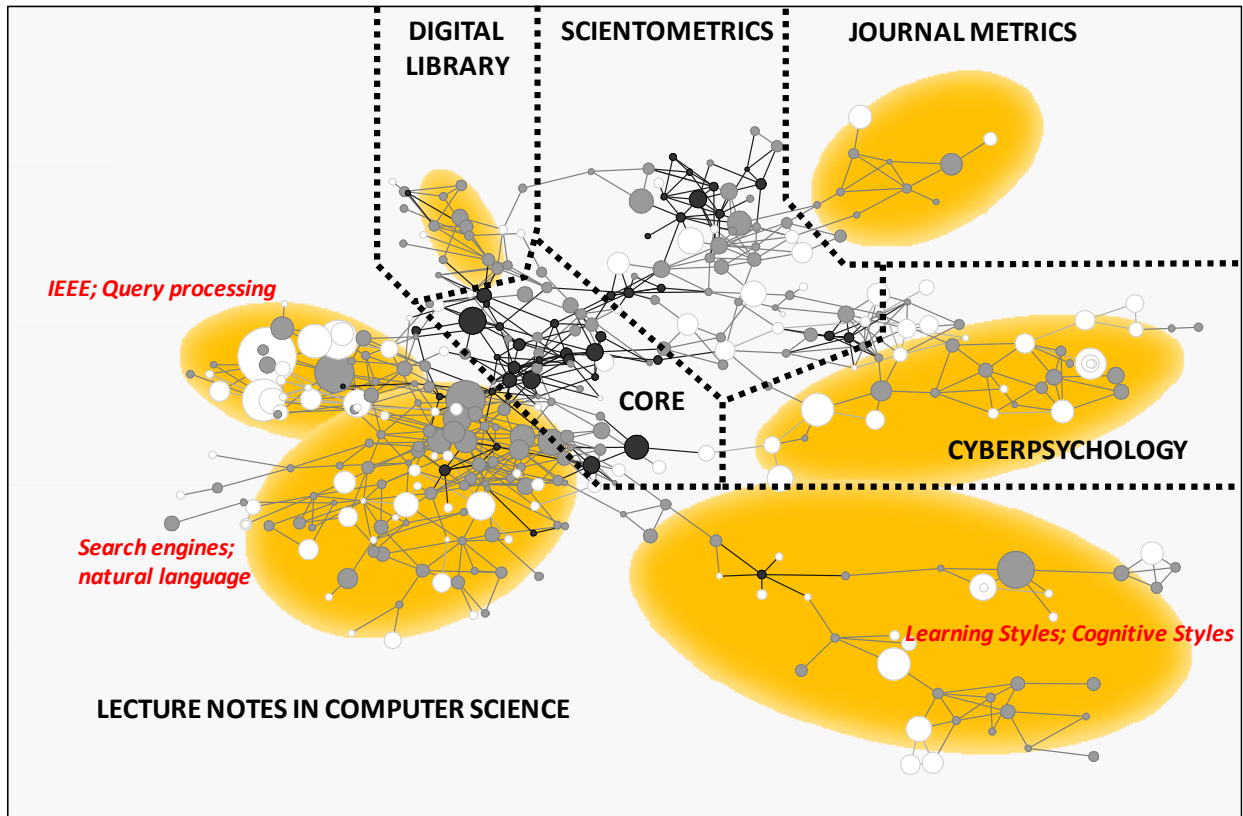
As shown in Figure 4, we can characterize specialty 35 as being composed of 11 successful experiments (multi-year threads), eight discontinued experiments (isolates from 2001 to 2007) and four current experiments (isolates in 2008). Interestingly, two of the isolates from 2008 appear as if they could have been continuations of threads that ended in 2007. For instance, one of the 2008 isolates and one of the 2007 communities are both labeled ‘author co-citation’. However, despite a common theme from a term standpoint, these communities did not have a common reference structure above the 0.3 threshold, or they would have been linked in a thread.

We stress that fragmentation of the sort shown in Figure 4 should not be considered to be ‘bad’. We suspect that some of the fragmentation may be signaling a re-evaluation of the intellectual base to support investigation in new topic spaces, which we view as a positive feature in science. Further investigation will be needed to explore this idea.

To be fair, we note that this is only one example, and one with admittedly short threads. Some threads in other parts of the map last the entire nine years and thus show a different dynamic. The point is that the global model enables us to see and understand the micro-dynamics in an area in a way that is not visible in a longitudinal map that averages results over many years. This detail provides useful insights; linking of communities from annual models shows clearly the persistence (or lack thereof) in the activities of a research community, and the degree to which there is experimentation with new socio-cognitive frames or rearrangements of the intellectual bases to explore new problem spaces.

Boundary areas: We now turn from a micro view of threads back to the macro view of the boundaries of information science. There are four major boundary areas in the global map of information science (see Figure 5). The smallest boundary area deals with journal metrics (upper right). The leading journal in this area is *Journal of Applied Measurement*, followed by *Scientometrics*. Another small boundary area is labeled digital libraries; the lead publication is *Digital Library Magazine*, followed by *Lecture Notes in Computer Science*. The medium sized boundary area on the right deals with cyberpsychology, and is led by a journal named

*Cyberpsychology and Behavior*. The largest boundary area, comprising the entire lower portion of the map, is dominated by one source – *Lecture Notes in Computer Science*.



**Figure 5: The global map of information science with boundary areas highlighted.**

To illustrate the competitive relationship between source titles<sup>10</sup>, we have calculated the publication shares for each source for each of the 2,706 research communities in our global map from 2000 to 2008. This allows us to determine the number of communities in which each source title was the leader, and the average community year by source for those communities in which it was the leader. Table 3 shows that the three source titles that were leaders in the largest number of communities are *LNCS*, *Scientometrics*, and *JASIST*. Of these, *LNCS* was the leader in roughly 7 times as many communities as either of the information science journals, and also had the highest mean community year. Thus, *LNCS* was the leader in the more recently occurring research communities, suggesting that in the core and boundary areas of information science *LNCS* is more influential in terms of currency than is *JASIST*.

To put a sharper point on this issue, we also found the sources that were #2 in each community. These are also listed in Table 3 as a function of the leading source title. The major information science overlaps with computer science appears to be by way of *IPM* and *JASIST*. *IPM* is the #2 source in 140 research communities where *LNCS* is the leader, while *JASIST* is the #2 source in

<sup>10</sup> Source titles include journals, conference proceedings, and book series. The distinction is useful in this study since *LNCS* is a very large source than can be considered as a book series, with each issue typically containing papers from a different conference in computer science. As such, *LNCS* has much broader topical coverage than most journals.

129 of the research communities where *LNCS* is the leader. *LNCS* is the #2 source title in 19 of the research communities where *JASIST* is the leader. The data also suggest that *Scientometrics* is not directly competing with computer science. The major competitors to *Scientometrics* are *JASIST* and two journals that are more oriented to R&D decision-making – *Research Evaluation* and *Research Policy*. *LNCS* is #2 in only 11 of the research communities where *Scientometrics* is the leader.

We note that the dominant role of computer science is not consistent with Small’s (1981) global map of information science. We did not determine if this disagreement represents a change in the competitive environment (i.e., computer science wasn’t as active in the 1970’s) or is simply due to database bias. We note that Small used a three-year set of the SSCI database to specify his global model. This database included very little, if any computer science literature as citing sources.

**Table 3. Competitive environment for the top three source titles.**

Leader	LNCS	Scientometrics	JASIST
# research communities	634	85	82
Mean community year	2005.76	2004.4	2003.6
#2 source (number of communities)	(140) – IPM (129) – JASIST (108) – LNAI (101) – CIKM	(35) – JASIST (26) – Research Eval. (21) – Research Policy (20) – IPM	(24) – Scientometrics (24) – IPM (19) – JOD (19) – LNCS

To summarize, in general the global map of information science makes the same distinctions within the information science literature as found in many local maps. There is a broad distinction between retrievalists and citationists with further delineation within each of those two main areas. The global map also breaks up the larger clusters in the local map. Specialty 35 from the local map corresponded with loosely linked threads and isolates that are viewed as separate topics and experiments in the global model. From the data presented thus far one cannot conclude that breaking up a large cluster into smaller clusters is inherently a more accurate description of the phenomena. Our claim of increased accuracy requires further analysis.

**Comparison of the local and global maps:** Table 4 shows summary statistics for the specialties and communities in the local and global models. Analysis of the local model involves the 655 references and the 18,358 citing articles that co-cite at least two of these references. Analysis of the global model of information science involves 45,019 unique references and a set of 68,320 citing articles. The values in Table 4 do not sum to these values for communities since references can be part of multiple intellectual bases and current papers are fractionally assigned. Titles and abstracts for articles published since 1996 and covered by Scopus were obtained from Scopus data, giving us titles for roughly 50% of the articles in the intellectual bases, and full coverage for the research fronts. Additional web-based searches were done to obtain titles (and abstracts where available) for the references in the 50 specialties to provide as much textual information for the local model as possible.

**Size characteristics:** The numbers of document clusters in the local models are very similar if one considers the temporal characteristics of the specialties. The local map contains 50 specialties, of which only 33 contained articles from the target literature during each year from

2000-2008. The global model contains 287 core research communities that each have a duration of a single year, by definition. The global model thus contains, on average, 32 research communities per year over the nine years, which compares quite well with the 33 specialties in the local model that persist for all nine years. Another way to think of this is that the core portion of the global model is very similar to a splitting up of the activity of the local model into annual research communities.

**Table 4. Summary statistics for the local and global maps of information science.**

Property	Local map	Global map		
		Core	Boundary	Cross-border
# of specialties	50			
# of research communities		287	1,393	1,026
BASE – # of references	655	5,803	31,264	29,273
BASE – average size of bases	13.1	20.2	22.4	28.5
BASE - # of references with titles	652	3,355	14,817	14,793
FRONT – # of articles	18,358 <sup>†</sup>	4,404 <sup>††</sup>	27,068 <sup>††</sup>	22,457 <sup>††</sup>
FRONT – average size of fronts	367.2	15.3	19.4	21.9
FRONT – # of articles from target	2,823	1,446	1,688	244

<sup>†</sup> Current papers that cite at least two of the 655 references and have a fractional assignment to the specialty of at least 0.5.

<sup>††</sup> Current papers that have a fractional assignment to the community of at least 0.5.

The global map has roughly twice as many references in each intellectual base as does the local model. However, the research fronts are much larger in the local model than in the global model. This is mostly due to the use of highly cited references in the local map that were not specific to the target literature. Many of these references could likely have been called perfunctory (Moravcsik & Murugesan, 1975) in that they represent a highly popular book or study that is relatively old. These are cited outside of the target literature far more than they are cited by the target literature, by a 6.5:1 margin. It is also interesting to note that the sizes of the intellectual bases and research fronts increase as one moves from core to cross-border.

We also note that the global model has higher coverage of the target literature. There are only 2,823 target articles in the local map using a reasonable assignment policy (current papers must cite at least two of the 655 references in the local model). By contrast, there are 3,378 target articles in the global map. But less than half of the target literature is located in the core. The majority is located in the boundary amidst a large number of articles that (as mentioned previously) are mostly in computer science. Over half of the target literature is assigned to research communities where the target literature is not dominant.

*Coherence:* To assess the relative accuracies of the clusters in the local and global maps, we calculated coherence values of each intellectual base and each research front from their titles and abstracts, and then calculated weighted averages for each map. Table 5 shows the average coherence values for the intellectual bases and research fronts in the two maps. Note that the number of specialties where one could calculate a coherence value was only 40 for the intellectual base and 47 for the research front. Insufficient textual information (e.g., two or three short titles with no abstracts) was available for the remainder of the specialties to calculate

coherence. There is a similar reduction in the number of intellectual bases in research communities with sufficient textual information to calculate coherence. More textual information was available for research fronts than for intellectual bases, leading to larger numbers of observations.

The coherence of research communities in threads in the global map is significantly higher than that of the global map isolates or the local map specialties in both the intellectual bases and research fronts. The large difference in coherence between communities in threads and isolates supports the interpretation provided earlier that isolates can be considered as discontinued experiments. Newly formed research communities will persist if they present a coherent picture of their topic space, and are more likely to be discontinued if they don't. The coherence of the isolates is not statistically different from the coherence of the local map.

**Table 5. Weighted average coherence values for the local and global maps of information science. Numbers of observations are in parentheses.**

	Local map	Global map: isolated communities	Global map: communities in threads
Intellectual base	0.0353 (n=40)	0.0382 (n=751)	0.0656 (n=1,572)
Research front	0.0328 <sup>†</sup> (n=47)	0.0377 <sup>††</sup> (n=940)	0.0571 <sup>††</sup> (n=1,757)

<sup>†</sup> Based on current papers that cite at least two of the 655 references and have a fractional assignment to the specialty of at least 0.5.

<sup>††</sup> Based on current papers that have a fractional assignment to the community of at least 0.5.

Table 5 also suggests that the coherence of the intellectual base is higher than the coherence of the research front. This is not surprising because co-citation analysis is designed to directly identify the intellectual base. The research front is based on a secondary assignment process.

Table 6 shows the results of a regression analysis where the effects of five independent variables on coherence (the dependent variable) were tested using all of the observations from Table 5. The five independent variables were converted to simple binary (0,1) indicators where possible so that the regression coefficients can be interpreted as marginal increases (or decreases) in coherence. The five variables along with their codings in the regression analysis are:

- persistent – coded as 1 if persistent, 0 if not
- global map – coded as 1 if from a global map, 0 if from a local map
- log(titles) – variable ranging from 0 to 1 based on number of titles
- intel base – coded as 1 if intellectual base, 0 if research front
- info sci – coded as 1 for specialties, 1 if core, 0.5 if boundary, and 0 if cross-border

The regression constant (0.0223) represents the level of coherence that one starts with. The most important factor is whether the document cluster is persistent (increasing coherence by 0.0175). 94% of specialties were persistent, meaning that there were research fronts assigned to the specialties for multiple years. 65% of the research communities were persistent in that they were members of threads. The second most important factor is whether the document cluster was in the global model. The global model increases coherence by 0.0135 over the local model. This finding supports the central claim of this paper that global models are more accurate than local models, and is robust in that alternative hypotheses for explaining the variation in coherence are included in the analysis. We also adjust for cluster size (log-titles), intellectual base vs. research

front (intel base), and the association of the cluster with information science (info sci). For these variables, coherence increases with an increase in cluster size, is higher for the intellectual base than for the research front, and is lower in information science than in surrounding areas. All coefficients are significant at the  $p=0.001$  level.

**Table 6. Regression results showing the effects on coherence of five variables.**

Source	SS	df	MS	Number of obs = 5107		
Model	.670846224	5	.134169245	F( 5, 5101) =	129.72	
Residual	5.27578471	5101	.001034265	Prob > F	= 0.0000	
				R-squared	= 0.1128	
				Adj R-squared	= 0.1119	
Total	5.94663093	5106	.001164636	Root MSE	= .03216	

coherence	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
constant	.022315	.004668	4.78	0.000	.0131635	.0314661
persistent	.017549	.0011512	15.24	0.000	.0152918	.0198054
global map	.013456	.0037455	3.59	0.000	.0061128	.0207983
log titles	.020214	.0045989	4.40	0.000	.0111984	.0292295
info sci	-.011233	.0016826	-6.68	0.000	-.0036329	-.0019836
intel base	.004663	.0009716	4.80	0.000	.0027584	.0065678

*Predicting community persistence:* We also test the ability of the coherence value to predict whether a research community will persist into the next time period. We have already established that the coherence of isolates is much lower (by 0.027 for intellectual bases, see Table 5) than the coherence of communities in threads. What is even more interesting is that the coherence of the last research community in a thread – the community that causes the death of the thread – is lower than the coherence of the communities in the thread that persist. The difference is significant (0.047 vs. 0.053).

Table 7 presents two regression analyses that explore the ability of the coherence value to predict whether a research community will persist into the next time period. Regression analyses were done independently for communities in the intellectual base (model #1) and research front (model #2). Persistence, the dependent variable, is coded as a binary variable. The two dependent variables are the coherence value and whether the community is an isolate (coded as 1) or not (coded as 0). The analysis was only done with data from the global map to test at the community level, and is limited to the 2001 to 2007 time period because one cannot tell if the isolated communities in the boundary years of 2000 and 2008 are in threads or not.

Since persistence is a binary variable, we can interpret Table 7 in much the same way as we did Table 6. The constants (0.80 and 0.78 for models #1 and #2, respectively) suggest that one starts with roughly an 80% chance of persistence. The average coherences across all communities (from Table 5) are 0.057 and 0.050 for intellectual bases and research fronts, respectively, with standard deviations of about .035. If a research community has average coherence, the persistence is raised to 0.84 (using the coefficient of 1.1 or 1.2). If the research community is very coherent (2 standard deviations above the norm), the persistence increases to about 0.91. Highly coherent research communities are extremely likely to persist. However, if a research

community is an isolate, persistence is reduced by 0.83 (the coefficient of ‘isolate’). Even the best research communities will only have an estimated persistence value of 0.065 (i.e., it’s likely to die). Whether the research community is an isolate is therefore the dominant factor with a coefficient of 0.83, while the coherence of a research community creates a much more narrow swing in the likelihood of persistence (0.14 if we use two standard deviations).

**Table 7. Predicting persistence in research communities from coherence.**

Model #1 (base) R-square=.658 Number of obs= 1810				
	Coef.	Std. Err.	t	P> t
persist				
constant	.796414	.014386	55.36	0.000
coh_intbase	1.107997	.1941783	5.71	0.000
isolate	-.835269	.0156015	-53.54	0.000

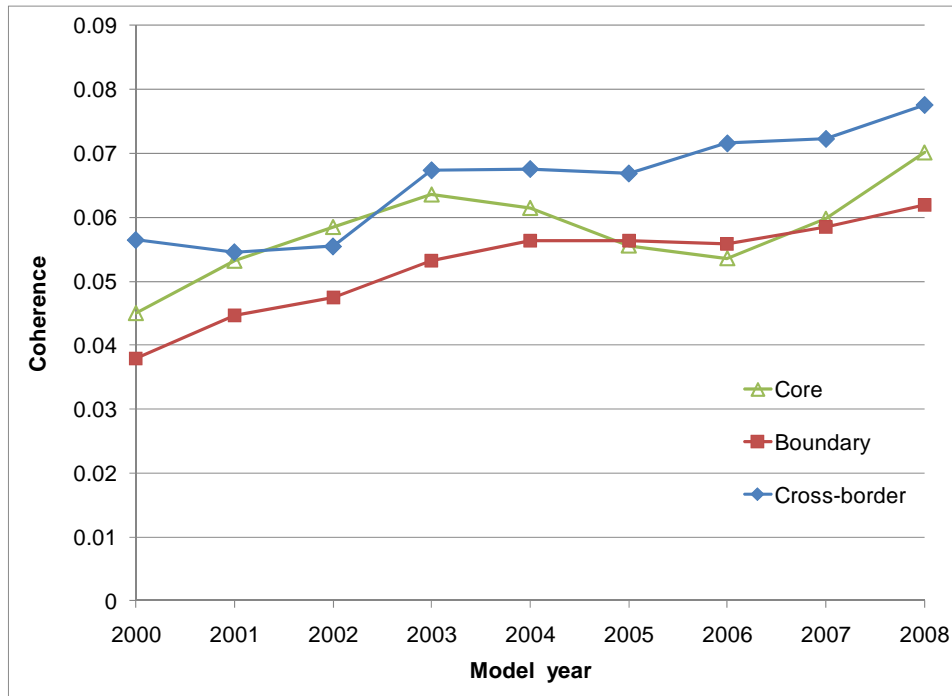
  

Model #2 (front) R-square=.644 Number of obs= 2107				
	Coef.	Std. Err.	t	P> t
persist				
constant	.777614	.0143028	54.37	0.000
coh_front	1.212083	.2073647	5.85	0.000
isolate	-.8249927	.014338	-57.54	0.000

The results in Table 7 are quite strong. 65% of the variance of the data can be explained by two variables, coherence and isolate status. The coefficients are essentially the same in both models. Annual calculations of coherence (see next section) and knowing whether the research community is new (i.e., not linked to an existing thread) is an exceptionally strong indicator of whether a research community will be linked to another research community in the next year.

*Coherence over time:* Our last analysis looks more closely at the change in coherence over time for information science and the environment surrounding information science. We focus specifically on the coherence of those research communities that are part of a thread in the core, boundary and cross-border sectors. We have excluded the research communities that are isolates because isolates are not represented in the cross-border sector.

Figure 6 shows that, for almost every year, coherence in the cross-border sector is superior and increasing. If we characterize this cross-border sector as consisting mostly of computer science, we have to conclude that in areas near the boundary between information science and computer science, computer scientists are pursuing a more coherent research strategy. Coherence in areas where information science is dominant (the core sector) are more variable (rising, falling and then rising) and generally remains in between the cross-border and boundary values. The boundary area has the lowest coherence. This may correlate with the fact that boundary research is often equated with being multidisciplinary. This raises the intriguing hypothesis that multidisciplinary research will have lower coherence than expected from the contributing disciplines.



**Figure 6: Temporal change in coherence in the communities populating core, boundary, and cross-border threads.**

## Discussion

Our final discussion focuses on two questions: What does the global map tell us about information science that cannot be gleaned from a local map? What have we learned that might enable us to increase the accuracy of future maps?

Global maps are unique in that they locate a target literature in context. In the global map of information science we find that the context is dominated by computer science. Previous document-level maps have not pointed this out. Small's (1981) global map of information science placed it within the context of the social sciences, but included little or no computer science literature due to database bias. None of the many local maps of information science published in the past two decades have included computer science literature. And while the relationship between computer science and information science at the journal level has been explored (Cronin & Meho, 2008), information science is seen as more of an exporter to computer science than an importer. Our analysis does not support this interpretation. The global map shows that computer science and information science are participating in (or, one might say, competing over) some of the same research areas. Computer science is larger and is dominating the boundary sector in terms of publication share. Within the global map, computer science is focused in areas of research with higher textual coherence, while information science is working in areas that have lower textual coherence.

The global map suggests that the intellectual base for a research community is undergoing constant change; local maps simply do not address this issue. Specialty #35 on citation behavior (Figure 4) provides an illustration of the value of tracking the intellectual base each year. This specialty had a very large intellectual base that one might assume did not change over time. We

would have seen a few very long threads in the global model if the underlying concept matrix had been stable. This was not the case. The intellectual base was linked to dozens of research communities that were either isolates or in very short threads. The intellectual bases did not persist for long periods of time. We therefore call into question the methodological choice behind most local maps of using multi-year averages without further proof that there is stability or drift in the underlying conceptual space.

The global map allows one to establish whether there is a causal relationship between coherence and persistence. This is, to our knowledge, the first time that this relationship has been shown to exist. And while the results only apply to this one environment, the results are sufficiently robust that we expect this relationship to hold for many, if not all, areas of science.

Finally, the global model also enables examination of coherence trends in different sectors. The cross-border sector started with the highest coherence and remained dominant for seven of the eight subsequent years in this study. The core sector had lower coherence that is more variable over time. The boundary sector had the lowest coherence. This raises hypotheses that are worth testing in other domains. Specifically, it would be interesting to explore the possibility that boundary areas will have lower coherence because they are multidisciplinary. We have only established this relationship for one disciplinary area. Identifying counter examples, disciplines where the boundary research is more coherent, would be a promising area for future research.

***Improving the accuracy of local and global maps:*** There are relatively few studies in the literature that try to objectively evaluate whether one map is more accurate than another. A review of that literature can be found in Boyack & Klavans (2010). We find this gap in the literature quite troublesome, and make the following suggestions as avenues for future research.

We posit that local maps will be more accurate if a different document sampling strategy is used. The 13-year document sampling strategy used for the Chen et al. (2010) local map seems excessive, and tends to hide fundamental shifts in the conceptual structure of the field. We point to recent local maps that use shorter time periods (5 to 10 years) and overlap these time periods in order to show evolution (Upham, Rosenkopf, & Ungar, 2010a, 2010b). We also point out the value of creating local maps each year so that one can pick up recent trends in coherence.

It is not necessary to limit a local map to the target literature. We point to a recent local map that started with a target literature and then expanded the target based on cited references (Greene, Freyne, Smyth, & Cunningham, 2010); the expanded map picked up application spaces that were not covered in the target literature. Expanding the target to include the domain of citing and cited disciplines would have picked up, and probably shown quite accurately, the role of computer science in information science.

We posit that global maps will become more accurate when the existing technical barrier of mapping 2,500,000 concepts is pushed to  $10^8$  concepts. This will allow one to create a fully specified global map based on 2-5 years of data. This map should be more accurate than a partially specified global map that stitches together annual maps.

The effect of different thresholds on the accuracy of the resulting models should also be explored. It would be useful to know what is gained in terms of accuracy as one goes from a top 1% threshold (currently used by Thomson Reuters ScienceWatch research front maps) to an effective top 12% threshold<sup>11</sup> to a top 20% or more threshold. It would be useful to know if there is a significant gain in accuracy if one excludes the older references – a policy that has already been incorporated in ScienceWatch. Other policies, such as excluding perfunctory references, might increase the accuracy of local and global maps.

A fully specified model need not be limited to references. Using text as well as citations to identify the structure of an intellectual base should be better than using either text or citations alone. The use of hybrid relatedness measures is a promising area of research that has received recent attention (Ahlgren & Colliander, 2009; Boyack & Klavans, 2010; Janssens et al., 2009; Liu et al., 2010).

Finally, the fact that the coherence of the research fronts was lower than the coherence of the intellectual bases in this study is, in our opinion, a reflection of the lack of attention spent on how current papers have been (or could be) assigned in co-citation analysis. The earliest approaches were quite simplistic; they either assigned a paper to one (and only one) intellectual base or fractionally assigned the paper based on partitioning of references (Franklin & Johnston, 1988). The historical ISI research front database made even simpler assignments; current papers were assigned based on the number of overlapping references. We suggest that more sophisticated ways of assigning research fronts to intellectual bases have the potential to increase their accuracy, and see that as a potentially fruitful area for future study.

## Conclusion

In summary, we have presented both a theoretical argument claiming that a global map should be more accurate than a local map, and a practical example showing that a global map of information science is both more accurate and more useful than a local map of information science. Although the coherence numbers and regression results in this study are strong, additional case examples in different disciplines are needed to further substantiate this claim.

In addition, we have shown that the research communities in temporally linked threads have a much higher coherence than isolated communities. This phenomenon was shown to have predictive value – the persistence of a research community from one year to the next can partially be predicted from the coherence value.

We have also suggested many ways in which accuracy of both local and global maps might be increased. We see many of these suggestions as potentially fruitful avenues for future research and hope that many others will join us in seeking the most accurate methodologies for science mapping.

---

<sup>11</sup> This threshold and process for creating an annual global map are used in Elsevier's SciVal Spotlight product.

## Acknowledgements

We thank Dr. Chaomei Chen of Drexel University for providing us with data on his recent map of information science for this comparison study, Henry Small for his extremely helpful comments on an earlier draft of this article, and anonymous reviewers for their many useful comments and suggestions.

## References

- Ahlgren, P., & Colliander, C. (2009). Textual content, cited references, similarity order, and clustering: An experimental study in the context of science mapping. *12th International Conference of the International Society for Scientometrics and Informetrics*, 862-873.
- Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. *Journal of the American Society for Information Science and Technology*, 58(7), 947-957.
- Boyack, K. W. (2009). Using detailed maps of science to identify potential collaborations. *Scientometrics*, 79(1), 27-44.
- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, in press.
- Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined cocitation and word analysis I. Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233-251.
- Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7), 1386-1409.
- Cronin, B., & Meho, L. I. (2008). The shifting balance of intellectual trade in Information Studies. *Journal of the American Society for Information Science and Technology*, 59(4), 551-564.
- Franklin, J. J., & Johnston, R. (1988). Co-citation bibliometric modeling as a tool for S&T policy and R&D management: Issues, applications, and developments. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 325-389). North-Holland: Elsevier Science Publishers, B.V.
- Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional, and institutional level. *Scientometrics*, 37(2), 195-221.
- Greene, D., Freyne, J., Smyth, B., & Cunningham, P. (2010). An analysis of current trends in CBR research using multi-view clustering. *AI Magazine*, 31(2), 45-62.
- Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). Structure of scientific literatures. 2. Toward a macrostructure and microstructure for science. *Science Studies*, 4(4), 339-365.
- Janssens, F., Leta, J., Glänzel, W., & de Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management*, 42(6), 1614-1642.
- Janssens, F., Quoc, V. T., Glänzel, W., & de Moor, B. (2006). Integration of textual content and link information for accurate clustering of science fields. *International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006)*, 615-619.

- Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, 45, 683-702.
- Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, 1(4), 287-307.
- Klavans, R., & Boyack, K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475-499.
- Klavans, R., & Boyack, K. W. (2010). Toward an objective, reliable and accurate method for measuring research leadership. *Scientometrics*, 82(3), 539-553.
- Klavans, R., Persson, O., & Boyack, K. W. (2009). Coco at the Copacabana: Introducing co-cited author pair co-citation (coco) analysis. *12th International Conference of the International Society for Scientometrics and Informetrics*, 265-269.
- Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology*, 61(6), 1105-1119.
- Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout, *Conference on Visualization and Data Analysis 2011*. San Francisco, CA.
- Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86-92.
- Moya-Anegón, F., Herrero-Solana, V., & Jimenez-Contreras, E. (2006). A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research. *Journal of Information Science*, 32(1), 63-77.
- 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), 339-344.
- Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, in press (doi:10.1016/j.joi.2010.03.006).
- Small, H. (1981). The relationship of information science to the social sciences: A co-citation analysis. *Information Processing & Management*, 17, 39-50.
- Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, 38(2), 275-293.
- Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799-813.
- Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 8(5-6), 321-340.
- Upham, S. P., Rosenkopf, L., & Ungar, L. H. (2010a). Innovating knowledge communities: An analysis of group collaboration and competition in science and technology. *Scientometrics*, 83(2), 525-554.
- Upham, S. P., Rosenkopf, L., & Ungar, L. H. (2010b). Positioning knowledge: Schools of thought and new knowledge creation. *Scientometrics*, 83(2), 555-581.
- van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, 68(3), 377-393.

- White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434.
- White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163-171.
- 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(4), 327-356.
- Zhao, D., & Strotmann, A. (2008a). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics*, 2, 229-239.
- Zhao, D., & Strotmann, A. (2008b). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology*, 59(13), 2070-2086.
- Zhao, D., & Strotmann, A. (2008c). Information science during the first decade of the web: An enriched author cocitation analysis. *Journal of the American Society for Information Science and Technology*, 59(6), 916-937.