Börner, Katy. (2007). Making Sense of Mankind's Scholarly Knowledge and Expertise: Collecting, Interlinking, and Organizing What We Know and Different Approaches to Mapping (Network) Science. Environment and Planning B: Planning and Design. Vol. 34(5), 808-825, Pion.

Making Sense of Mankind's Scholarly Knowledge and Expertise: Collecting, Interlinking, and Organizing What We Know and Different Approaches to Mapping (Network) Science

Abstract

This paper discusses and compares different approaches to collect, interlink, organize, and make sense of scholarly knowledge and expertise in a comprehensive and timely fashion. 'Comprehensive' refers to the need for collecting and interlinking multi-lingual, multi-disciplinary records from multiple sources such as publications, patents, grants, and others to truly capture all relevant knowledge. By 'timely' I want to emphasize that there has to be a way to integrate the most recent, i.e., today's, publications with existing holdings of scholarly knowledge and expertise. I then discuss the advantages and limitations of using search engines to access and text mining and data mining to help extract meaning from mankind's wisdom. Next, I suggest the usage of semantic association networks as a viable and complementary alternative to interlink and make sense of scholarly knowledge and expertise. The second part of the paper exemplifies and contrasts three approaches that can be used to delineate and make sense of scholarly knowledge. The first approach uses questionnaire data, the second citation data from a major digital library, and the third personal bibliography files. These approaches are exemplified by mapping the emerging research area of *Network Science*. A particular focus is the identification of major experts, papers, and research areas and geospatial locations where network science research is conducted. The paper concludes with a summary and outlook.

1 Introduction

Today there are more researchers and scientists alive than have ever lived on this planet before. They are not only alive but are actively conducting research and publishing their results. The number of publications being produced is staggering. Some domains produce as many as 40,000 journal papers each month. At the same time, human perception and cognition capabilities remain nearly constant and our knowledge collection, access, and management tools are rather primitive. Consequently, even the smartest brain on this planet does not stand a chance of keeping up with the accelerating speed of knowledge production. The fact that each day provides only 24 hours and even more knowledge will be produced and published tomorrow does not help either.

All this leads to a quickly increasing specialization of researchers, practitioners, and other knowledge workers; a concerning fragmentation of science; a world of missed opportunities for collaboration; and a nightmarish feeling that we are doomed to 'reinvent the wheel' forever.

This is no way to run science. It becomes a major concern when scientific results are essential to enabling all humans to live a healthy, productive, and fulfilling life. We simply do need better tools to keep track, access, manage, and utilize our collective scholarly knowledge and expertise.

Recent work on knowledge domain visualizations [4, 7, 29] attempts to map science on a large scale. The resulting maps equip people with a global view of our collective knowledge and wisdom. Just like old sea charts, maps of science can help people to find places of interest while avoiding monsters. They complement local fact retrieval via search engines by providing global views of large amounts of knowledge.

Imagine scientists would not drown in the daily flood of information but could effectively navigate, access, and utilize mankind's wisdom whenever they need it. Imagine not needing to fear gaps in knowledge or missed opportunities. Instead scientists would intelligently make use of the best knowledge sources – be it a book, digital document, or expert – easily find the best collaborator, or quickly identify the best research opportunity.

Early global views or (mental) maps of science were temporarily lost due to the speed of innovation and scientific progress. Thanks to the increasing digital coverage of knowledge, algorithm development, and the compute power available today, comprehensive maps of science can be generated automatically for educational, research, and practical purposes.

This paper begins with a review of today's scholarly data collection, integration, and access. It then applies the *Semantic Association Network* (SAN) approach [3] to improve the collection, representation, and interlinkage of scholarly data. Subsequently, three approaches to mining and mapping scholarly knowledge are discussed and compared. Section 5 presents a summary and outlook.

2 Scholarly Knowledge Record Keeping and Access – State of the Art

Today, numerous scholarly databases exist. While some are quite large, none of them contains all of mankind's scholarly knowledge. Very few databases are multi-lingual. Almost none integrates different publication types such as papers, patents, grants.

As for scholarly data, the citation indexes published by *Thomson Scientific*'s *Web of Science* (http://isiknowledge.com) are some of the best sources for bibliographic entries on major, predominantly English journal publications. However, access to publications in nearly 7,600 journals, 2,000 books, as well as Web documents, e-journals, and preprints is quite expensive. *Scopus* (http://www.scopus.com) is an interesting alternative. It provides access to 15,000 titles from 4,000 different publishers including 12,850 academic journals, 750 conference proceedings, and 600 trade publications. About 245 million references interconnect 28 million abstracts. It also covers 12.7 million patent records from four patent offices. Free access to scholarly knowledge is available via *Google Scholar*, *PubMed* and *CiteSeer*. *Google Scholar* (http://scholar.google.com) indexes interdisciplinary papers, theses, books, abstracts, and

articles provided by academic publishers, professional societies, preprint repositories, universities and other scholarly organizations but also harvested from the Web. Although no concrete numbers were available, a search for 'the' conducted on Jan 2nd, 2006 returned 561,000,000 hits leading to the assumption that more than 560 million records are indexed. Recent work reports the overlap of *Google Scholar*, *Scopus*, and *Web of Science* for different subject areas [1, 23]. *PubMed Central* is the U.S. National Institutes of Health (NIH) free digital archive (http://www.pubmedcentral.nih.gov). It comprises over 15 million biomedical and life sciences journal publications dating back to the 1950's. *CiteSeer* (http://citeseer.ist.psu.edu) provides access to and advanced search for scholarly publications mostly in computer science. Publications are harvested from the Web or submitted by scholars. On Jan 2nd, 2006, 739,135 publications were searchable. The *Digital Bibliography & Library Project* (DBLP) indexed more than 800,000 articles and served several thousand links to home pages of computer scientists in October 2006 (http://dblp.uni-trier.de). The *United States Patent and Trademark Office* (UPSTO) gives free access to more than 3 million patents. Grants awarded by the *National Science Foundation* (NSF) and the *National Institutes of Health* (NIH) comprise about 180,000 and 1 million records respectively.

Unfortunately, there are major interoperability and cross-linkage problems, see also the discussion by Herbert Van de Sompel and his colleagues [35] and myself [3]. Very few of today's scholarly datasets (e.g., papers, patents, grants) are stored and integrated in a way that citation, co-author, and other linkages can be traversed. A notable exception is the *Library Without Walls* (http://library.lanl.gov/lww/) at the *Los Alamos National Laboratory* that interlinks major publication databases and supports citation-based search across different holdings. However, its development is so expensive that only few institutions can afford to participate.

In short, while digital libraries and the Web start to serve as globally distributed knowledge storage infrastructures, most of the datasets are stored in silos that are not explicitly interlinked.

2.1 Knowledge Access

Most of today's digital libraries supply search engines and reference desk librarians to help people find relevant records. Using reference desk librarians works like magic. Relying on search engines, however, can cause trouble: Companies such as *Google* and *Microsoft* try to make us believe that we can live in 'flatland' – no directory structures are necessary and data objects can have any name – yet still find everything we want, thanks to their superior retrieval software. However, the use of search engines resembles charging a needle with a search query and sticking it into a haystack of unknown size and consistency. Upon pulling the needle out, one checks the linearly sorted items that got stuck on it. This seems to work for fact-finding. However, it keeps us always at the bottom of confirmed and unconfirmed records in the haystack of our collective knowledge. We can explore local neighborhoods of retrieved records via Web and citation links, but there is no 'up' button that provides us with a more global view of what we collectively know and how everything is interlinked. Without context, intelligent data selection and quality judgments become extremely difficult.

2.2 Knowledge Mining

To summarize, group, and help make sense of large sets of scholarly records, *Data Mining* and *Text Mining* approaches are frequently applied. *Latent Semantic Indexing* (LSA) [12] or the *Topics Model* [19] are only two of many existing approaches. Mining works well if the records are written in similar styles, using similar formatting and conventions, and are of similar length, etc. Hence, important insights can be gained if mining techniques are applied to publications written using the scholarly conventions and language of a specific domain. However, if applied to make sense of interdisciplinary or multi-lingual datasets, they quickly fail. Take for example the word 'prototype' – it has a very different meanings in computer science, biology, psychology, or architecture. Or take a closer look at the scholarly papers you read. Titles like 'Everything you always wanted to know about XX', 'An unifying theory of XX', and 'Towards scalable XX' simply do not give text-based approaches a lot of meat to chew on. Keywords are used in a similarly creative manner—particularly if they are supplied by authors and not by trained

librarians. Systems like BioText (http://biotext.berkeley.edu/) which used natural language processing techniques to retrieve and synthesize information from textual descriptions of bioscience research or Arrowsmith (http://arrowsmith.psych.uic.edu) a tool for identifying links between two sets of Medline articles work in small, very well defined domains. In general, humans might simply be too uniquely creative and therefore unable to produce proper raw material that can be analyzed using existing text and data mining algorithms.

2.3 Link Traversal

Given the success of Google and the importance of social networks for knowledge dissemination, link traversal might be a viable alternative. Citation indexes published by *Thomson Scientific*, *Google Scholar*, and *CiteSeer* already support citation link traversal. For example, given a set of seminal papers or one's own papers, one can find all the papers that cite or are cited by them if they are available in one database. The *Proceedings of the National Academy of Sciences* (PNAS) online interface (http://www.pnas.org) even provides citation maps that show articles citing or being cited by a selected article. The aforementioned *Library Without Walls* project goes one step further by interlinking major publication databases and supporting citation- based search across different holdings. Some digital libraries, such as the citation indexes published by *Thomson Scientific*, *DBLP Bibliography Server* (http://www.sigmod.org/dblp/db/), and *ACM Digital Library* (http://portal.acm.org), have started to provide information on co-authorships. Services include a listing of all papers by an author, a listing of all co-authors for one author, and co-author link traversal.

All these services require the identification of unique author names. This can be done partially automatically [20, 26, 34]. To achieve high accuracy, automatic techniques have to be combined with manual efforts such as the WikiAuthors project (http://meta.wikimedia.org/wiki/WikiAuthors) or the Lattes CV system (http://lattes.cnpq.br). The latter system is used by researchers, students, managers, professionals, and other actors of the Brazilian scientific community to evaluate competences of candidates to scholarships and/or research support and to select consultants, members of committees and advising groups, among others. Scholars can update their profile at any point in time and the absence of a profile is likely to cause impediments to payments and renewals. Any manual effort should be 'seeded' by the result of automatic analyses as well as existing author databases such as MARC author records provided by the Library of Congress (http://www.loc.gov/marc) or author lists collected by scientific societies (e.g., the American Mathematical Society's Mathematical Reviews Database has kept track of individual authors since its inception in 1940, http://www.ams.org/mr- database/mr-authors.html).

3 Collecting, Interlinking, and Organizing Scholarly Knowledge – Semantic Association Networks

In prior work [3] *Semantic Association Networks* (SANs) were suggested as a novel means of using Semantic Web technology to tag and interlink scientific datasets, services (e.g., algorithms, techniques, or approaches), scholarly records (e.g., papers, patents, grants), and authors to improve scholarly knowledge and expertise management. This paper examines a subset of the SAN entity and link types so as to exemplify their utility and usage in detail.

3.1 Representing and Interlinking Scholarly Knowledge

Today, scholarly knowledge is stored in forms such as papers, patents, and grant proposals. Let's refer to those as 'records' of scholarly knowledge. Typically, each record has an associated set of authors. They might be inventors in the case of patents or principal investigators in the case of grants. Let the 'is an author of a record' relationship be denoted by an 'associated' relation. Authors that jointly publish one record are said to 'co-author'. A record that references another record is said to 'cite' the other record. Note that the 'cite' link is directed while the two other links are undirected. This leaves us with a set of two entity types and three link types as shown in Figure 1, see also [3].

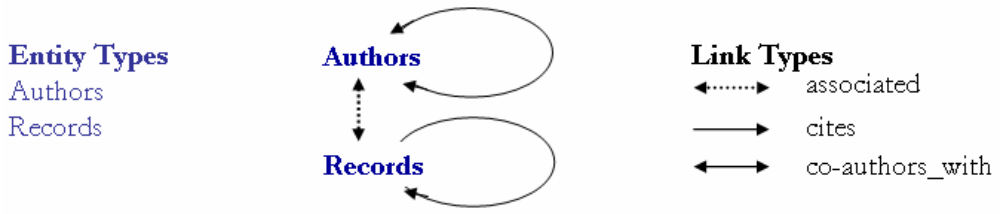


Figure 1: Entity types and link types that are commonly used to represent scholarly knowledge

The two entity types commonly have diverse attributes. For example, a record typically has a publication date, a publication type (e.g., journal paper, book, patent, grant, but see also discussion of bibliographic record types in Section 4.4), and topics (e.g., keywords or classifications assigned by authors and/or publishers). Authors have an address with affiliation and geo-location information. Because authors and records are associated, the geo-location(s) and affiliation(s) of an author can be attributed to the authors' papers. Similarly, the publication date, publication type, and topic(s) can be associated with a paper's author(s). Statistics such as the number of papers or co-authorships (over time) per author or the number of citations (over time) per paper can be derived. It even becomes possible to track changes of topics and geo-locations for authors over time.

Entities and links of different types can be represented as coupled networks. Based on the position of an entity or link in the coupled author-record network, attributes such as node degree (e.g., indicating the number of co-authors or the number of received citations) or betweenness centrality [13] (counting the number of shortest paths between nodes that pass through a node) can be computed. The latter is an indicator of a 'gatekeeper' role.

Also of interest is the grouping of authors and papers over time but also in geo-spatial and semantic space. Obviously, only authors that are currently alive can co-author, with some exceptions due to publication lags. Paper citations are commonly made between papers on similar topics. The geographic location of authors appears to have a major influence on who co-authors together and who is cited [5, 37]. Physical proximity matters – even in the Internet age [2, 6].

3.2 Collecting Scholarly Knowledge

Today, no database includes all of mankind's scholarly knowledge. Moreover, no institution seems likely to attempt one; the creation and indefinite maintenance appear too daunting. Even if such a database could be created, given its enormous and quickly growing size, how could anybody ever benefit from it?

However, it appears that with an appropriate technological setup and carefully designed incentive structures, the 'wisdom of crowds' [33] might easily solve both problems. There are several examples of how such a system might look and feel. *Wikipedia* (http://wikipedia.org), masterminded by Jimmy Wales, provides an 'empty shell' for anybody to fill with encyclopedia articles – more than 1,490,000 English articles as of October 2006. *CiteULike* (http://www.citeulike.org), conceived and implemented by Richard Cameron, provides another excellent example of an 'empty shell' that helps academics manage bibliographic entries for academic papers. It currently supports entries from many online sources. One mouse click suffices to create an entry plus a link to the original publication, into your personal *CiteULike* bibliography. Users can annotate papers, share annotations and bibliographies, and download files in EndNote or bibtex format. A similar 'empty shell' can be setup for author and record entity types and their three different types of links: associated, co-authored_with, and cites. Entity access and link traversal logs could be used as quality indicators and serve to improve access to high-demand records, authors, and linkages [35]. Obviously, there is no need to re-enter all the entity and link data already available in digital form (see Section 4.4.). Also, about 80% of data integration and linkage identification might be possible by automatic means. It is the last 20% where human input is necessary to achieve a data

quality and coverage that is truly useful as a global index to mankind's scholarly knowledge. The scholars' rewards are the many new ways of knowledge access, management, and sense-making that become possible if scholarly knowledge is collected in this manner (see Section 4).

3.3 Organizing Scholarly Knowledge

Social bookmark managers such as *BlinkList* (http://www.blinklist.com), *Connectedy* (http://www.connectedy.com), *Del.icio.ous* (http://del.icio.us), Digg (http://digg.com), *GiveALink* (http://givealink.org), Jots (http://www.jots.com), scuttle (http://scuttle.org), simpy (http://www.simpy.com), or *unalog* (http://unalog.com) and many other services support the collection, categorization and sharing of web linkages. *Flickr* (http://www.flickr.com) supports the collection, sharing, and tagging of photos. Common to all of them is the importance of tags to create folksonomies. Folksonomy, a portmanteau word combining 'folk' and 'taxonomy', refers to the organization of knowledge based on tags attached by thousands of users to millions of records. It is a rather decentralized form of classification that uses the 'wisdom of crowds' to classify records. Folksonomies are typically displayed as lists of most commonly used tags. The text font type, color, and size are frequently used to indicate the number of records that have a particular tag, the novelty of a tag, activity bursts for the usage of a tag, and so on. The interlinkage density among the diverse social record managers is astonishing. After traversing the many links, it seems that the implementation of the *Semantic Association Networks* is obvious and poses no technological challenges.

4 Making Sense of Scholarly Knowledge – Mapping Science

Scholarly knowledge collected, represented, interlinked, and organized as discussed in Section 3 can be analyzed and mapped in diverse ways to support sense-making. Depending on the information need, different subsets of entities and links become more relevant. Nodes and links can be analyzed at the local level (e.g., to identify highly cited papers or all authors that have a gatekeeper role) or at the sub-network level (e.g., to determine all sub-areas in a domain of science). Sometimes, the properties of the complete network need to be computed (e.g., to map all of science). Subsequently, there will be a brief introduction into mapping science locally and globally. Following this I review, exemplify, and contrast three approaches to access and make sense of 'local' scholarly knowledge. These local science maps can be used as 'high resolution inserts' in global science maps [7].

4.1 Mapping Science Research

Research in *Bibliometrics* and *Scientometrics* [4, 8, 10, 38] aims to analyze, map, and study science by scientific means. Recently, scientometric techniques have been extended to deal with very large datasets [7, 16, 21, 24, 25] creating the new research area of *Computational Scientometrics*. Here, advanced data mining and information visualization techniques are applied to interlink and analyze papers, patents, grants on a large scale. The resulting visualizations can be utilized to objectively identify, for example, major research areas, experts, institutions, publications, journals, or grants in a research area of interest. In addition, they can identify interconnections, the import and export of research between fields, the dynamics of scientific fields (e.g., speed of growth, diversification), scientific and social networks, and the impact of strategic and applied research funding programs. This knowledge is not only interesting for funding agencies but also for companies, researchers, and society.

Most studies have aimed to visualize established domains—among others, *Information Science* [39], *Geography* [30, 31], *Animal Behavior* [28]. These studies are commonly conducted on publication datasets downloaded from major digital libraries or online sources. In order to map a specific discipline, searches for relevant keywords are run or cited reference searches are used to retrieve all papers that cite or are cited by a set of seminal papers.

There are two major problems with this approach. First, relatively few individuals have access to high quality publication data as provided by Thomson Scientific or Scopus. Second, in this age of increasing

disciplinary specialization, it is very hard if not impossible to identify appropriate search phrases or the complete set of seminal papers that should be used to retrieve all relevant papers. This is an issue particularly for newly emerging or highly interdisciplinary research areas, such as *Nanotechnology* or *Network Science*. These areas interlink and draw from diverse sources. Research typically starts out from unconnected or loosely interconnected islands of research. Results are published in diverse venues. It is only after a while that research strands interconnect, joint conferences bridging multiple scientific disciplines are held, and journals are created.

In the case of *Network Science*, research is currently conducted in *Mathematics, Statistics, Graph Theory, Computer Science, Information Science, Scientometrics, Biology, Physics,* and *Economics*, to name just a few areas. While major network science experts and seminal papers can potentially be identified for a specific research area, it is hard if not impossible to identify and compare the entities (records and authors) from all contributing domains.

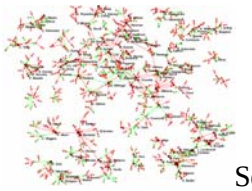
In the next subsections, three different approaches to analyzing and mapping *Network Science* are introduced. The first is based on expert consultation via questionnaires; details are reported in a *National Research Council* study [9]. The second applies standard bibliometric techniques and tools to citation data downloaded from Thomson Scientific's database. The third uses personal bibliography files to make sense of this emerging, highly interdisciplinary domain.

The focus here is particularly on social and scholarly collaboration networks [17, 18] – last but not least, it is people that contribute, consume, and hopefully benefit from the fruits of science.

4.2 First Approach: Expert Consultation via Questionnaires

As part of the *National Research Council* study on *Network Science* [9], Weimao Ke and myself conducted a bibliometric analysis of the social networks and expertise of network science researchers. The data were contained in 499 completed questionnaires that reported 923 self-identified 'collab_with' links (linking to collaborators) and 376 'invite' links (identifying people that should be invited to fill out questionnaires). Details on the questionnaire design can be found in [9]. In total, the names of 1,241 unique network science researchers were identified. Matching based on email addresses was used to ensure that these names are truly unique. Examination of the data revealed that 'collab_with' links were primarily made to researchers in spatial or thematic proximity. According to comments by colleagues who had filled out a questionnaire, 'invite' links were typically made to people who were expected to be potentially missing in the dataset, i.e., not to 'major players' or 'gatekeepers'.

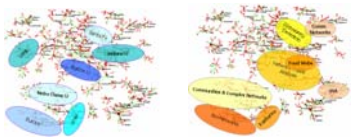
The resulting dataset was used to identify collaboration and invitation networks. Figure 2 shows all components with at least 10 nodes plotted using *Pajek* [11]. Each of the 630 network science researchers and practitioners is denoted by a node. Nodes that have a high betweenness value or were mentioned frequently in the dataset are labeled by the name of the researcher/practitioner. The size of the node corresponds to the number of times the name was mentioned in the dataset. Node color indicates whether the person submitted questionnaire answers (orange) or not (dark red). Red links denote 'collab_with' relations, and green links denote 'invite' relations.



See Figures file.

Figure 2: Major components of the collaboration and invitation network based on questionnaire data

Figure 3 shows institutional affiliations and research areas overlaid on the network in Figure 2. The institutional affiliations and research areas were independently identified by eight researchers who conduct network science research at Indiana University.



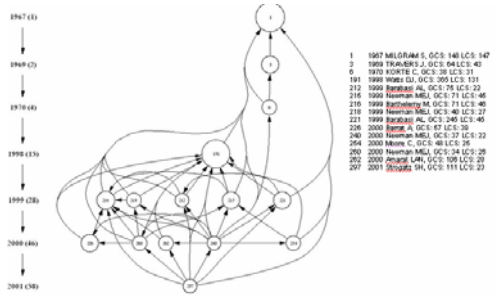
See Figures file.

Figure 3: Collaboration and invitation networks overlaid with affiliation data (left) and research topics (right).

Figure 3 merely demonstrates the potential of visualizing the interplay of affiliations, topical themes, and social interrelationships of researchers/practitioners in a certain domain of research. Compared to other maps of scientific disciplines, this network clearly exhibits the characteristics of a new and emergent research area: It consists of many unconnected networks of collaborating network science researchers and a rather heterogeneous coverage of research topics.

4.3 Second Approach: Citation Data

The second attempt applies standard bibliometric techniques and tools to citation data downloaded from Thomson Scientific's database. All papers that use "Small World" in their titles or cite Stanley Milgram's 1967 *Psychology Today* paper on that topic were downloaded. The 412 papers were written by 482 authors in 1967-2002 and comprise major network science papers. As the dataset contains citation linkages, the paper citation graph can easily be plotted using the *HistCite* software [14]. The citation graph for a local citation score (LCS) of at least 20 (i.e., each of the papers was cited at least 20 times by any of the 412 papers in the set) is also shown in Figure 4. The interactive version of this graph is available online at http://garfield.library.upenn.edu/histcomp/small_world/ (select 'Graphs' and then 'LCS >= 20').



See Figures file.

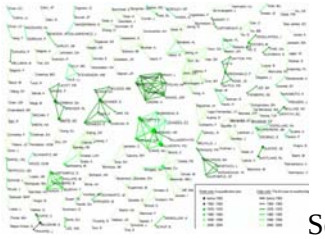
Figure 4: HistCite paper-citation graph of network science publications.

The graph comprises 15 papers interlinked by 52 citations that are ordered in time, see timeline on the left. The number of papers published in a year is given in parentheses. Each node is numbered. Node details are provided at the right including publication year, first author, global citation score (GCS) (i.e., the number of citations the paper has in the Thomson Scientific database) and LCS. Node size roughly corresponds to the total number of citations received. The paper with the highest citation count is Watts and Strogatz' "Collective dynamics of 'small-world' networks" – a paper published in *Nature* in 1998. It received more than 1,700 citations from papers in the Thomson Scientific database within the first 8 years after publication. In comparison, Milgram's seminal paper received 148 citations within the first 45 years.

Closer examination quickly reveals that most papers in this 15 node graph are written by physicists. This can be attributed partially to the fact that "no one descends with such fury and in so great a number as a pack of hungry physicists, adrenalized by the scent of a new problem," as Duncan J. Watts described the invasion of network science by physicists [36]. However, it also reflects the very different dynamics of publication patterns across different domains of science. Physicists use e-print archives to quickly and

effectively disseminate their results. Papers easily attract hundreds of citations within a few years – a number unreachable throughout the entire lifetime of most publications in for example the humanities.

Using code provided by Weimao Ke and documented at http://iv.slis.indiana.edu/lm/lm-kdvis.html, the co-author network of the very same Thomson Scientific dataset was extracted. It comprises 255 author nodes (excluding authors that published single author papers exclusively) and 521 co-authorship links. It is rendered in *Pajek* in Figure 5. The node area corresponds to the number of papers an author has published. Author nodes are color-coded based on average publication year. Edges are color-coded based on the first year of the co-authorship. See figure legend for details.



See Figures file.

Figure 5: Complete co-author network based on citation data.

Compared to the co-author networks in Figures 2 and 6, this network is rather unconnected. Only 12 components have 5 or more authors. The largest component shown at left in Figure has 13 authors and is discussed in Section 4.5.

4.4 Third Approach: Personal Bibliographies

The third attempt uses personal bibliography files such as bibtex or EndNote files. Interested in mapping the area of 'network science,' bibliography files were invited from nine major network scientists: Albert-László Barabási, Noshir S. Contractor, Loet Leydesdorff, José F. F. Mendes, Mark E. J. Newman, Mike Thelwall, Alessandro Vespignani, Duncan Watts, and Stanley Wasserman. The details of the data acquisition, cleaning, and analysis are reported in [27]. In sum, 13 files in EndNote, bibtex, and free-text format containing over 7,000 references were parsed. Extracted were 5,425 unique article records and their 5,330 unique authors. The dataset covers many different disciplines over the years 1637-2005. Information on which expert submitted what record was preserved so that the 'popularity' of a record can be studied, along with the coverage and overlap of personal bibliographies. Diverse statistics such as the number of articles per publication year per expert, or the number of article records per author, were derived and are reported in [27]. Here, the focus is on the analysis, mapping, and interpretation of the co-author network.

The 5,330 unique authors belong to 266 components. There are 14 components that have at least 10 nodes, with the largest component having 131. Figure 6, rendered in *Pajek*, shows the complete co-author network with triads and dyads removed. 579 authors are shown. Nodes for authors with article counts of 10 or higher are labeled with the author's name. As in Figure 5, node size corresponds to the number of papers published. Also like Figure 5, author nodes are color-coded based on average publication year; edges, based on the first year of the co-authorship (see the Figure 5 legend for details). The edge width is based on the number of co-authorships.


See Figures file.

Figure 6: Components with size larger than three of the co-author network based on bibliography data.

The author with the most papers in this bibliography dataset is *Leydesdorff*. This is particularly surprising as his bibliography files could not be parsed automatically and hence are not included in the dataset. Other experts must have included many of his papers in their personal bibliography files. The node with the highest degree represents *Jeong*, who co-authored with 25 other authors in this data set. *Dorogovtsev* and *Mendes* co-authored most often – 29 times – according to this dataset.

Compared to the co-author network extracted from the citation dataset in Section 4.3, this network is much more interconnected. This can be partially attributed to a larger sample (5,330 as opposed to 255), in which only components larger than three are shown. A comparison of the giant components of the co-author network based on citation and bibliography data reveals interesting matches, as will be shown in section 4.5.

While personal bibliography files do not provide information on paper citation interlinkages, they are a readily available source of high quality bibliographic data. The record type specification possible via *bibtex* and *EndNote* files supports record type tagging equal or superior to professional databases.

Bibtex supports standard entry types such as:

@article
@book
@booklet
@conference
@inproceedings
@inbook - a chapter, section, etc.
@incollection
@manual
@mastersthesis
@misc
@phdthesis
@proceedings
@techreport
@unpublished
@collection
@patent.

EndNote 7.0 supports reference types such as:

Journal Article
Book
Book Section
Manuscript
Edited Book
Magazine Article
Newspaper Article
Conference Proceedings
Thesis
Report
Audiovisual Material
Film or Broadcast
Artwork
Map
Patent
Hearing
Bill
Statue
Case
Figure

Personal Communication	Chart or Table
Computer Program	Equation
Electronic Source	Generic

Hence, scholarly records of different types, e.g., papers and patents, can be interlinked. Book chapters can be interlinked at the chapter level instead of at the book level. EndNote even supports links to specific figures, tables, or equations.

A specific research community can be invited to submit their bibliographies. Bibliographic entries can be tagged with research topics. This seems to provide an easy, comprehensive, and timely means to delineate interdisciplinary domains. The resulting datasets can be used to identify major publication records (based on number of times submitted) but also major experts (based on the number of papers published or number of co-authors as well as their positions in the network). They can be used to examine the domain knowledge of the experts that submitted the files. Plus, the collected bibliographic data can be used to analyze the coverage of existing databases.

Obviously, the collected data does not provide any information on paper citation linkages. Yet high-quality bibliographic data constitutes a great starting point to harvest supplemental information such as citation linkages or citation counts from *Google Scholar, PubMed, CiteSeer,* or other databases.

Personal bibliographies are more subjective and 'local' compared to the presumably more objective, 'global' collections produced by publishers. However, personal sampling biases can be exploited to identify the value of a publication record in a certain local context. The more people, either in a certain domain of research or across different research domains, who include a paper in their personal bibliography files the more valuable a certain paper might be either locally and/or globally.

Note that this is not a suggestion to store all of mankind's scholarly knowledge in one database. Instead, a unifying index to mankind's scholarly knowledge and a set of effective knowledge management tools that use this index should be created. While efforts like *CiteULike* (http://www.citeulike.org) aim to help academics share, store, and organize their academic papers, we must go one step further. We need to help academics identify the best records and experts in a domain of interest, to interlink them in meaningful ways, and to design tools that help us manage and utilize our collective knowledge and expertise.

4.5 Comparison

Subsections 4.2-4.4 reviewed three very different approaches to mapping (network) science. Here a qualitative rather than quantitative comparison of these approaches is attempted.

Table 1 presents an overview of the data types and attributes made available by the different approaches (see also discussions in Subsections 4.2-4.3).

Questionnaires or wikipedia-like environments can be used to acquire data discussed in Section 3.1. This approach is restricted mainly by the time and money available to acquire and analyze questionnaire data.

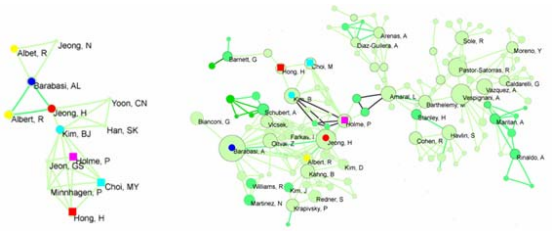
Citation data, particularly when harvested from the Thomson Scientific citation databases, provides all entity and link information. This approach is limited by the coverage of the database used and our ability to issue relevant search queries. The latter is a non-trivial undertaking for interdisciplinary or emerging research areas.

Personal bibliographies are easy to acquire for any scientific discipline and in any language. This approach is limited mainly by missing affiliation and geo-location for authors and missing citation linkages. However, bibliography records, when perfect, make excellent queries for downloading missing data from professional and other databases.

Approach	**Entities**					**Links**		
	Authors		Records					
	Affiliation	Geo-location	Publication Date	Publication Type	Topic(s)	associated	cites	co-authors_with
Questionnaires	yes	no	no	no	yes	no	no	yes*
Citation Data	yes	yes	yes	yes	yes	yes	yes	yes
Personal Bibliographies	no	no	yes	yes	yes+	yes	no	yes

Table 1: Entity and link types and attributes available/used in different approaches.
* We report information on collaborations here assuming that collaborations result in joint papers.
+ Information on topics can be collected by asking experts to identify research topics when they submit bibliography files. Supplied topics can then be used to tag bibliographic records.

The three approaches differ in the sampling and the number of records used. Yet, the extracted co-author networks show interesting commonalities. As an example, Figure 7 shows the giant components of the co-author networks based on citation data, see Figure 5, and bibliography data, see Figure 6.



See Figures file.

Figure 7: Comparison of the giant components of the co-author network based on citation data (left) and bibliography data (right).

When author names in the left network are compared with those in the right, eight out of 12 names can be matched, see node shape and color-coding in Figure 7. Interlinkage patterns among the major nodes are similar as well. Furthermore, many of the eight author names can also be identified in the collaboration and invitation network, based on questionnaire data shown in Figure 2.

5 Summary and Outlook

This paper started with an overview of today's scholarly knowledge collection, access, and management approaches. Then an argument for a re-conceptualization of the way we collect, access, and make sense of scholarly knowledge was presented. Instead of relying on textual records and text-based search and mining approaches, it appears to be very feasible and highly beneficial to improve and help

users traverse and mine the interlinkage among scholarly entities. Applying the semantic association network approach introduced in [3], the representation of scholarly knowledge by two entity types and three link types and their attributes was discussed. Based on this representation, improved means of collecting, organizing, and making sense of scholarly knowledge become possible. The network-based scholarly knowledge representation was also used to review three approaches that aim to map science to facilitate sense-making. The approaches were exemplified by mapping the scholarly networks of network science researchers and practitioners. Obviously, any of the three approaches can as well be used to map other domains of research. In fact, the replication of the comparison conducted here is a high priority item for future work.

In this paper, the representation of scholarly knowledge was restricted to a subset of the entity and link types introduced in [3]. An obvious next step is the full implementation of semantic association networks in a domain of research. *Network Science* is a good candidate, since the recently funded 'Network Workbench' (NWB) project (http://nwb.slis.indiana.edu) would definitely benefit from it. The project aims to provide a unique distributed environment for large-scale network analysis, modeling, and visualization. The envisioned data-code-computing resources environment will be a one-stop online portal for researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

One of the first domain-specific portals that the NWB infrastructure will support is a science mapping service that will soon be available at http://scimaps.org/biblio. Anybody will be able to enter the portal site and select a topic area or upload a personal bibliography file. The selected or uploaded data will then be used to extract basic statistics, e.g., major authors, highly cited papers (if data permits), and number of publications per year (per author). Additionally, the site can be used to plot the structure and evolution of paper-citation and co-authorship networks or simply to merge and clean a set of personal bibliography files. I hope you will check it out. Ideally, it will improve the way we collect and make use of scholarly data across disciplinary and language boundaries in an incremental and scalable fashion.

Up until today, major attempts to map all of science have used data provided by the *Institute of Scientific Information* (ISI), known today as *Thomson Scientific*. These efforts originated with the pioneering *Atlas of Science* published by ISI in 1981 [22]. Work by Eugene Garfield [15] and Henry Small [32] followed. Today, Kevin Boyack and Dick Klavans are at the forefront of mapping all of science [7, 24]. However, it would not be surprising if the first comprehensive and timely map of science—one that truly covers all of mankind's scholarly knowledge—is generated based on data supplied by the very scholars who produce this knowledge.

Acknowledgements

I would like to thank Weimao Ke for writing and running diverse software used in this study. Thanks go to Eugene Garfield and Soren Paris who gave us access to the HistCite software and data used in this paper. Michael Thelwall, Andre Skupin, Ron Day, July M. Smith, Stacy Kowalczyk, and the anonymous reviewers provided expert comments on the presented work. Albert-László, Noshir S. Contractor, José F. F. Mendes, Mark E. J. Newman, Loet Leydesdorff, Mike Thelwall, Alessandro Vespignani, Duncan Watts and Stanley Wasserman were kind enough to share their personal bibliography files with us. This research is supported by the National Science Foundation under IIS-0513650, CHE-0524661, and a CAREER Grant IIS-0238261 as well as a James S. McDonnell Foundation grant in the area Studying Complex Systems.

References

1. Bakkalbasi, N., Bauer, K., Glover, J. and Wang, L. (2006) Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3* (7).
2. Batty, M. (2003) The geography of scientific citation. *Environ Plan A*, *35*. 761-765.

3. Börner, K. (2006). Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. in Geroimenko, V. and Chen, C. eds. *Visualizing the Semantic Web*, Springer Verlag, 183-198.
4. Börner, K., Chen, C. and Boyack, K. (2003). Visualizing Knowledge Domains. in Cronin, B. ed. *Annual Review of Information Science & Technology*, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, 179-255.
5. Börner, K., Penumarthy, S., Meiss, M. and Ke, W. (in press) Mapping the Diffusion of Scholalry Knowledge Among Major U.S. Research Institutions. *Scientometrics*.
6. Börner, K., Penumarthy, S., Meiss, M. and Ke, W. (2006) Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. *Scientometrics, 68* (3). 415-426.
7. Boyack, K.W., Klavans, R. and Börner, K. (2005) Mapping the Backbone of Science. *Scientometrics, 64* (3). 351-374.
8. Chen, C. (2002) *Mapping Scientific Frontiers*. Springer-Verlag, London.
9. Committee on Network Science for Future Army Applications - National Research Council (2005) *Network Science*. The National Academies Press, Washington, D.C.
10. Cronin, B. and Atkins, H.B.E. (2000) *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. ASIST.
11. de Nooy, W., Mrvar, A. and Batagelj, V. (2005) *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, Cambridge.
12. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41* (6). 391-407.
13. Freeman, L.C. (1977) A set of measuring centrality based on betweenness. *Sociometry, 40*. 35-41.
14. Garfield, E. (2004) Historiographic mapping of knowledge domains literature. *Journal of Information Science, 30* (2). 119-145.
15. Garfield, E. (1998) Mapping the world of science. in *the 150 Anniversary Meeting of the AAAS*, (Philadelphia, PA), The Scientists.
16. Giles, C.L. and Councill, I.G. (2004) Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. *Proceedings of the National Academy of Sciences, 101* (51). 17599–17604.
17. Glanzel, W. (2001) National characteristics in international scientific co-authorship relations. *Scientometrics, 51*. 69-115.
18. Glänzel, W. (1995) International scientific collaboration in a changing Europe. A bibiometric analysis of coauthorship links and profiles of 5 East-European countries in the sciences and social sciences 1984-1993. *Science and Science of Science, 4*. 24-31.
19. Griffiths, T.L. and Steyvers, M. (2004) Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl_1)*. 5228-5235.
20. Han, H., Giles, C.L., Zha, H., Li, C. and Tsioutsiouliklis, K. (2004) Two Supervised Learning Approaches for Name Disambiguation in Author Citations. in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*, 296-305.
21. Huang, Z., Chen, H., Chen, Z.-K. and Roco, M.C. (2004) International Nanotechnology Development in 2003: Country, Institution, and Technology Field Analysis Based on USPTO Patent Database. *Journal of Nanoparticale Research, 6* (4). 325-354.
22. ISI (1981) *ISI atlas of science: Biochemistry and molecular biology, 1978/80*. Institute for Scientific Information, Philadelphia, PA.
23. Jacso, P. (2005). Comparison and analysis of the citedness scores in web of science and Google Scholar. in *Digital Libraries: Implementing Strategies and Sharing Experiences*, Springer Verlag, Berlin, 360-369.
24. Klavans, R. and Boyack, K.W. (2006) Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57* (2). 251-263.

25. Lawrence, S., Giles, C.L. and Bollacker, K. (1999) Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, *32* (6). 67-71.
26. Malin, B. (2005) Unsupervised Name Disambiguation via Social Network Similarity. in *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security, in conjunction with the SIAM International Conference on Data Mining*, (Newport Beach, CA), 93-102.
27. Murray, C., Ke, W. and Börner, K. (2006) Mapping Scientific Disciplines and Author Expertise Based on Personal Bibliography Files. in *Information Visualisation Conference*, (London, UK), 258-263.
28. Ord, T.J., Martins, E.P., Thakur, S., Mane, K.K. and Börner, K. (2005) Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases. *Animal Behaviour, 69*. 1399-1413.
29. Shiffrin, R.M. and Börner, K. (eds.). (2004) *Mapping Knowledge Domains*. PNAS.
30. Skupin, A. (2002) A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications, 22* (1). 50-58.
31. Skupin, A. (2004) The World of Geography: Mapping a Knowledge Domain with Cartographic Means. *Proceedings of the National Academy of Sciences, 101, Suppl. 1*. 5274-5278.
32. Small, H. (1999) Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50* (9). 799-813.
33. Surowiecki, J. (2004) *The Wisdom of Crowds*. Doubleday.
34. Torvik, V.I., Weeber, M., Swanson, D.R. and Smalheiser, N.R. (2003) A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology, 56* (2). 140-158.
35. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C. and Warner, S. (2004) Rethinking Scholarly Communication: Building the System that Scholars Deserve. *D-Lib Magazine, 10* (9).
36. Watts, D.J. (2003) *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company.
37. Wellman, B., White, H.D. and Nazer, N. (2004) Does Citation Reflect Social Structure? Longitudinal Evidence from the 'Globenet' Interdisciplinary Reserach Group. *Journal of the American Society for Information Science and Technology, 55* (2). 111-126.
38. White, H.D. and McCain, K.W. (1989). Bibliometrics. in Williams, M.E. ed. *Annual review on information science and technology. Volume 24*, Elsevier Science Publishers, Amsterdam, Netherlands, 119-186.
39. White, H.D. and 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.