Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 79, No. 2 (2009) 219–234
DOI: 10.1007/s11192-009-0414-2

The Scholarly Database and its utility for scientometrics research

GAVIN LAROWE, SUMEET AMBRE, JOHN BURGOON, WEIMAO KE, KATY BÖRNER

Indiana University, School of Library and Information Science, 10th Street & Jordan Avenue, Bloomington, IN 47405, USA

The Scholarly Database aims to serve researchers and practitioners interested in the analysis, modelling, and visualization of large-scale data sets. A specific focus of this database is to support macro-evolutionary studies of science and to communicate findings via knowledge-domain visualizations. Currently, the database provides access to about 18 million publications, patents, and grants. About 90% of the publications are available in full text. Except for some datasets with restricted access conditions, the data can be retrieved in raw or pre-processed formats using either a web-based or a relational database client. This paper motivates the need for the database from the perspective of bibliometric/scientometric research. It explains the database design, setup, etc., and reports the temporal, geographical, and topic coverage of data sets currently served via the database. Planned work and the potential for this database to become a global testbed for information science research are discussed at the end of the paper.

Introduction

The availability of digitized scholarly data sets, combined with sufficient computing power to integrate, analyze, and model multivariate data makes it possible to study the structure and evolution of science on a global scale [SHIFFRIN & BÖRNER, 2004; BÖRNER & AL., 2003] The results can be communicated via tables and graphs, but also via geographic and topic maps. Science can be observed from above. Scientific frontiers that emerge across different sciences can be discovered and tracked. Different funding models can be analyzed and simulated. School children can start to understand the symbiosis of different areas of science. Global maps of science might even provide a new means to organize, interlink, and communicate existing bibliometric and scientometric results.

To give an example, Figure 1 shows a map of the 'Melanoma' research area over the last 40 years. It comprises 53,804 unique Medline papers, 299 unique genes retrieved from the Entrez Gene database, and 367 unique proteins from the Universal Protein Resource (UniProt). All are related to melanoma. Cosine similarity based on co-occurrence of MeSH terms was used to spatially layout the papers and their associated genes and proteins, see also [BOYACK & AL., 2004].. Labeling of major research areas

Received December 5, 2007

Address for correspondence:
GAVIN LAROWE
E-mail: glarowe@indiana.edu

0138–9130/US $ 20.00

was done by hand after exploration of the map using VxInsight [BOYACK & AL., 2002]. The different research areas can be grouped into applied medical sciences (left side) and basic molecular sciences (right side). Interestingly, papers in the applied science portions of the map are less numerous than their molecular science counterparts.

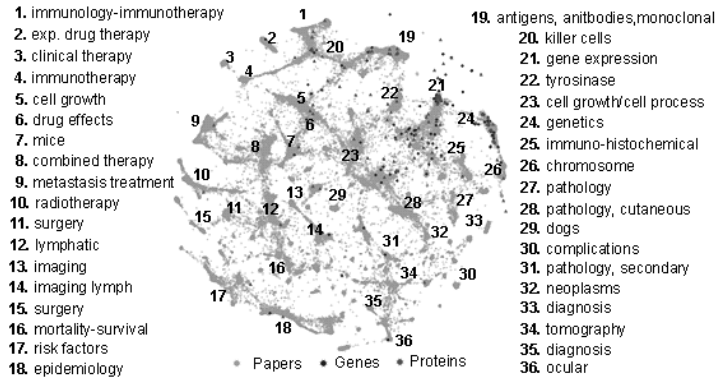

Figure 1. Snapshot of an interactive map showing melanoma related papers, genes, and proteins

The study of science on a large scale, also called *Computational Scientometrics*, by C. Lee Giles, requires a qualitatively different research approach. Instead of downloading a small number of bibliometric records from an existing digital library site and analyzing them via desktop tools, millions of scholarly records need to be interlinked and analyzed. While a co-author or paper-citation network with up to 500 nodes can be easily visualized, see for example the Author Co-citation Analysis of Information Science (reported in [WHITE & MCCAIN, 1998]), the communication of the interdependencies among thousands or potentially millions of entities poses major challenges [KLAVANS & BOYACK, 2006].

Computational biology, astronomy, or physics require major infrastructures such as large-scale biological cyberinfrastructures, the Hubble telescope, or particle accelerators. Similarly, the study of science on a macro-scale requires access to high quality, high coverage data and the computational means to process such data.[1] Any bibliometric/scientometric study is only as good as the data that it uses to confirm a hypothesis, to test a new technique, or to validate a predictive model. Moreover, many studies use data sets from different sources with often differing data types. For example, input-output studies require input data, e.g., funding amounts, number of new graduates,

[1] Note that the term 'cyberinfrastructure' was coined in a NSF blue ribbon panel report [ATKINS & AL. 2003]. It refers to advanced computer, network, and data technologies that are assumed to revolutionize scientific and engineering research.

and output data, e.g., the number of publications, received citations, awards, policy changes. Unfortunately, this data is stored in unconnected data silos. The identification and inter-linkage of unique authors, investigators and inventors is non-trivial. A number of bibliometric/scientometric scholars have created their personal databases of interlinked data sets. These databases are typically rather small and highly specific to the research interests of the researcher in question. Contrary to other scientific disciplines where data is freely and widely shared, there are very few bibliometric or scientometric test data sets available. Hence, it becomes very time consuming (in terms of data downloading, cleaning, and inter-linking) if not impossible (if data sets require access permissions) to replicate a certain study or to reproduce a given result.

Many diverse entities such as PubMed, CiteSeer, arXiv, and the United States Patent Office provide free data dumps of their holdings under certain conditions. However, most bibliometric/scientometric scholars are not trained in parsing millions of XML-encoded bibliometric records and very few have expertise in the setup and maintenance of multi-terabyte databases. Plus, data parsing, cleaning, design, integration, and setup is not really at the core of scientometrics research. If there existed a database that is jointly used by increasing numbers of scholars and practitioners, it would be more likely that missing records or links will be discovered, important data sets will be integrated, the best (author/institution/country/geo-code) unification algorithms will be applied, and prior research studies can more easily be replicated and verified.

The following sections introduce the Scholarly Database at Indiana University that aims to serve scholars in bibliometrics and scientometrics as well as information scientists with an interest to test their novel data integration or author disambiguation algorithms on large-scale, scholarly data sets. The first section describes the database design, setup, and schemas. The second section reviews the different data sets currently available in the database, together with their temporal, geospatial, and topic coverage. Finally, planned future work is discussed, emphasizing the potential of this database to become a global testbed for information science research [TICHENOR, 2006].

Database architecture and access

The Scholarly Database comprises a production, development, and research system. These systems are hosted at the School of Library and Information Science at Indiana University, Bloomington. A brief description of the system architecture and hardware setup follows.

System architecture

The system architecture of the Scholarly Database is shown in Figure 2. Raw data, pre-processed data, metadata, and any other raw data artifacts are stored and processed

in the *Data Space*. *Data Services* support data harvesting, data mining, pre- and post-processing, statistical analysis, natural language processing (NLP), multi-agent data simulations, and information visualization services among others. Aside from backup and storage, data provenance is provided via the use of metadata associated with both the raw data and internal database artifacts (e.g., schemas, tables, views, etc.) and user artifacts such as queries or views, and their associated metadata, used in published papers and current research projects.

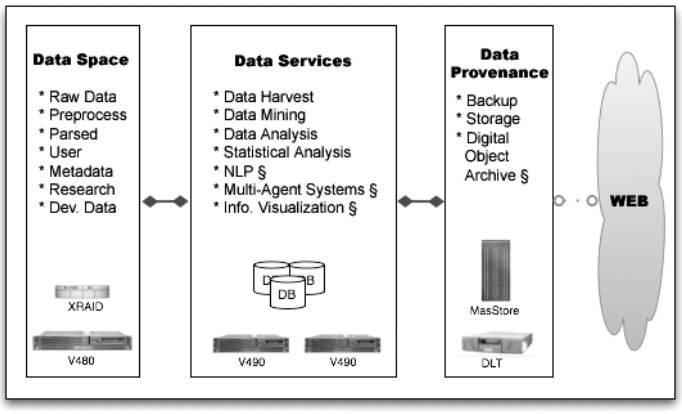

Figure 2. Overview of the Scholarly Database infrastructure (§ = future feature)

The production system for the Scholarly Database utilizes two Sun V490 servers with 16G of memory serving two redundant, mirrored clusters of the Scholarly Database running PostgreSQL v8.1.3 under Solaris 10. Failover, load balancing, and replication are provided via a future intermediary system service between each of these clusters. In addition to this intermediary system service, each instance of the database is isolated within its own zone under Solaris 10, providing increased security and failover for the system at-large.

A development cluster is used for developing future production versions of various data sets found within the Scholarly Database. After post-processing, incorporation, and testing, new data sets receive their own data stores and various schema templates are incorporated. When approved, these data sets are then pushed to one of the production clusters whereby they are replicated between each instance. The development cluster also houses research data sets which have been, or soon will be, promoted to production status.

Additionally, a research database cluster running on a Sun V480 with 32G of memory provides a database sandbox for researchers to develop and refine personal or proprietary data sets and ideas that might be used in future research. Aside from being a completely isolated system, this cluster provides users with a large-memory

environment and large disk storage for cyberinfrastructure projects involving multi-gigabyte and, in the future, terabyte-scale data and application requirements. Three large-scale cyberinfrastructure projects based at our school took advantage of this cluster as of January 2007.

The database schema is too extensive to describe or depict in this paper, but an abstract overview may be useful. The current implementation utilizes three schemas: public, base, and auxiliary. The public schema contains all of the post-processed raw data which has been loaded into the database. This data and its associated metadata are rarely modified, except when new updates are received from a data provider. The base schema contains all of the foundational views found in the web interface used for search, display, and download of data by a user. The base schema contains mostly virtual views, though for extremely large data sets, some materialized views are utilized as well. The purpose of the auxiliary schema is to provide a space where schemas/tables/views/functions created by the users of the system can be stored for re-use. Its ancillary purpose is to provide an area where certain one-time or proprietary, non-public components (e.g., views or functions) can be explored within the context of a given data set via a given user.

Database access

Access to the database is regulated via both a web-based front-end and an internal back-end client to the server. The front-end interface allows external users to search through all of the articles or documents in the database via author, title, or keyword, see Figure 3 (left). Results are returned showing generic fields such as journal name, title, author name, date of publication, etc., see Figure 3 (right). Each record can be further expanded to show fields specific to a given data set. Please see next section for information about data sets and access permissions.

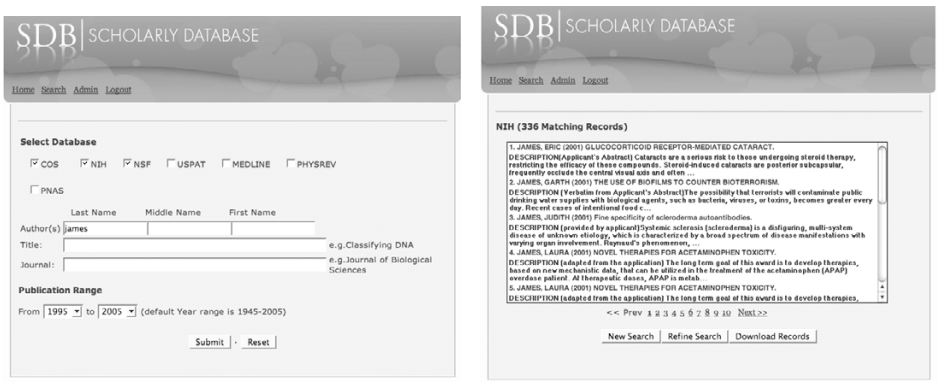

Figure 3. Database search interface (left) and initial search result interface (right)

Data set acquisition, processing, and coverage

The Scholarly Database is unique in that it provides access not only to diverse publication data sets, but also to patent and grant award data sets. Currently, four publication data sets, one patent dataset, and two grant award data sets are loaded and served with different access restrictions: Publication data comprises 40 years of Medline publications provided by the National Library of Medicine (http://www.nlm.nih.gov), a 110-year Physical Review dataset provided by the American Physical Society (http://aps.org), a 5-year dataset of the Proceedings of the National Academy of Sciences (PNAS) provided by PNAS (http://www.pnas.org), and a 30-year Journal Citation Report (JCR) dataset by ISI Thomson Scientific (http://www.isinet.com). Patents issued by the United States Patent and Trademark Office (http://www.uspto.gov) are available as well as funding award data for grants awarded by the National Science Foundation (NSF) (http://www.nsf.gov), and the National Institutes of Health (NIH) (http://www.nih.gov).

Data acquisition and processing

Data sets were acquired from different sources using different approaches. Some are one time acquisitions. Others are updated on a continuous basis. Table 1 provides an overview:

Table 1. Data sets and their properties (* future feature)

Dataset	# Records	Years covered	Updated	Restricted access
Medline	13,149,741	1965–2005	Yes	
PhysRev	398,005	1893–2006		Yes
PNAS	16,167	1997–2002		Yes
JCR	59,078	1974, 1979, 1984, 1989, 1994–2004		Yes
USPTO	3,179,930	1976–2004	Yes*	
NSF	174,835	1985–2003	Yes*	
NIH	1,043,804	1972–2002	Yes*	
Total	18,021,560	1893–2006	4	3

- *Medline* provides two types of data: baseline files which are distributed at the end of each year and include all of the Pubmed records which have been digitally encoded in XML for Medline; and newly added data for that particular year which are subsequently edited/updated in future baseline file releases. It is provided in XML format with a custom DTD which also changes year-to-year. Update files are usually provided on a monthly or quarterly basis.
- *Physical Review (PhysRev)* is composed of 398,005 XML-encoded article files covering nine journals (A, B, C, D, E, PR, PRL, RMP, PRST, AB)

from 1893–2006. A single DTD exists for the entire collection we have which encompasses all changes made throughout the history of the digital encoding of these files prior in SGML. It is an internal data set that is not searchable or available to the public-at-large. It cannot be used for commercial purposes.

- *Proceedings of the National Academy of Sciences (PNAS)* comprises full text documents from the Proceedings of the National Academy of Sciences covering the years 1997–2002 (148 issues containing some 93,000 journal pages). The data are also available on Microsoft Access 97 format. The data set was provided by PNAS for the Arthur M. Sackler Colloquium, Mapping Knowledge Domains, held May 9–11, 2003. It is available for research and educational purposes to anyone registered for the Sackler Colloquium on Mapping Knowledge Domains who signed the copyright form. It cannot be redistributed without prior permission from PNAS and cannot be used for commercial purposes.
- *Journal Citation Reports (JCR – Science Editions)* from Thomson Scientific covers two data sets: (1) one covering the years 1994–2004; and (2) the second containing Cited and Citing Pairs records for 1974, 1979, 1984 and 1989 from the Science Citation Index – Expanded (SCI-E). Both are restricted use. This data cannot be used, distributed, or otherwise without prior written permission from Thomson Scientific.
- *United States Patent Office (USPTO)* data were provided by the US Patent Office via ftp. XML-encoded dumps are regularly downloaded from ftp://ftp.uspto.gov/pub/patdata/. This is a publicly accessible data set.
- *National Science Foundation (NSF)* funds research and education in science and engineering. It does this through grants, contracts, and cooperative agreements to more than 2,000 colleges, universities, and other research and/or education institutions in all parts of the United States. Data were acquired from the National Science Foundation (NSF) web site during 2004–2005. It is composed of raw text files as distributed by the NSF for the years listed above. It is a publicly accessible data set.
- *National Institutes of Health (NIH)* data composed of both CRISP and awards data were downloaded from the CRISP on-line search engine and the NIH web site respectively The CRISP data includes information regarding extramural projects, grants, contracts, etc. associated with projects and research supported by the National Institutes of Health. NIH awards provide information on principal investigator and institution, and award amounts and can be used to augment the CRISP data. All NIH data was acquired between 2004–2006.

Detailed information on these data sets is available via http://iv.slis.indiana.edu/db. New data sets are added on a continuous basis. In the following sections, we provide information regarding the temporal coverage, geographical coverage, and topic coverage of the publicly accessible data sets.

Temporal coverage

As can be seen in Figure 4a, the Medline dataset has the most records per year, with about 500,000 unique records each year based on the baseline dumps from years prior. There are about 200,000 new USPTO patents each year, about 10,000 new NSF awards and 50,000 new NIH awards per year. The years covered by each dataset are given in Table 1.

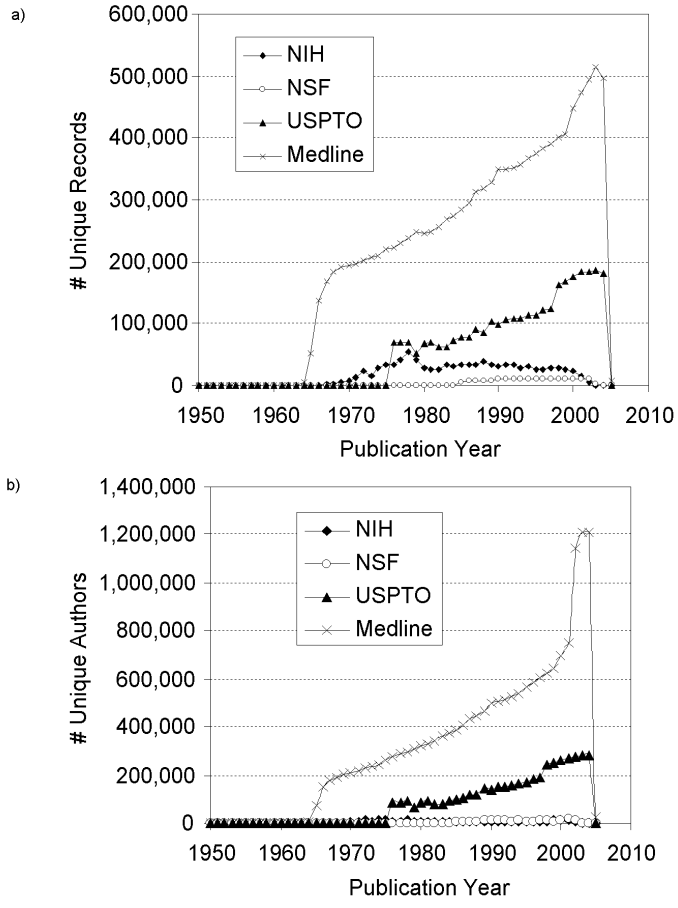

Figure 4. Number of unique records and authors per year between 1950 and 2005

The number of unique authors per year is plotted in Figure 4b. A concatenation of first author name and last author name was employed to identify unique authors. Very likely, there is more than one 'John Smith' in the Medline dataset and we know that some authors change last names, e.g., when they get married. However, the below graph can give one a quick sense of the major differences in the number of authors contributing to the growth of the different data sets. It also shows the continuous increase of the number of authors over time.

Geographical coverage

Even though information visualization has been part-and-parcel of geographical information systems for many years, recent hardware and software advances, combined with the advent of web-centric technologies and data-interchange formats (e.g., XML) have helped promote both the usage of large-scale multivariate data, as well as user-friendly web-based visual interfaces that synthesize and analyze such data. Moving from internal application-servers to web-based architectures, geographical coverage visualizations using Google Maps can quickly inform or support other geo-centric hypotheses, which is provided below for some of our data sets. As an example, the geospatial distribution of four datasets is given in Figures 5 through 8.

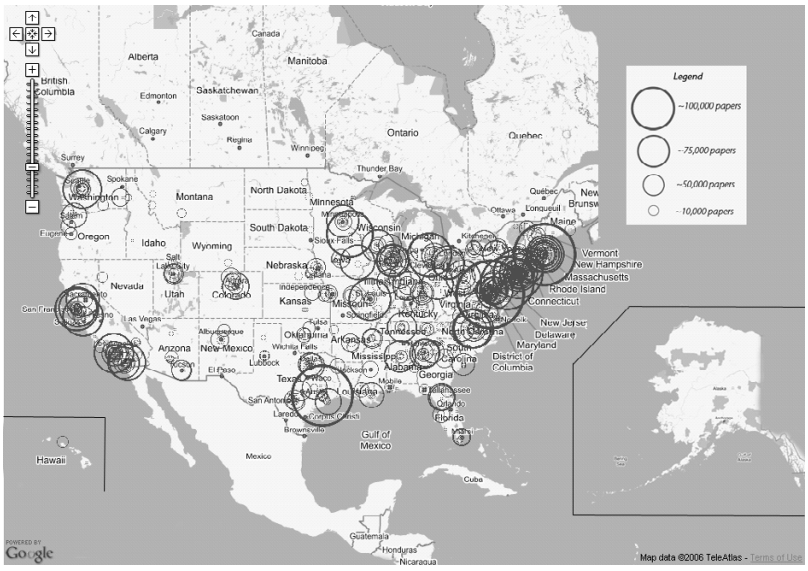

Figure 5. Total number of Medline publications per author by geo-location (U.S.)

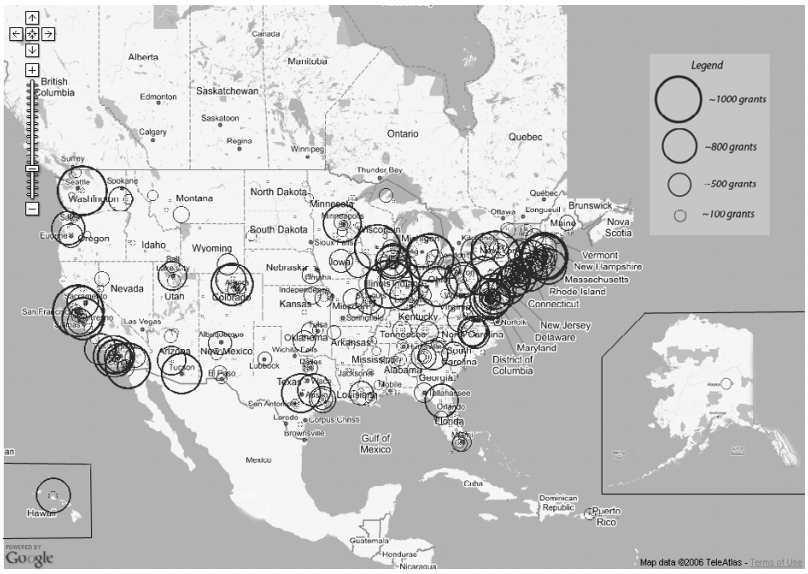

Figure 6. Total number of NSF awards per awardee by geo-location (U.S.)

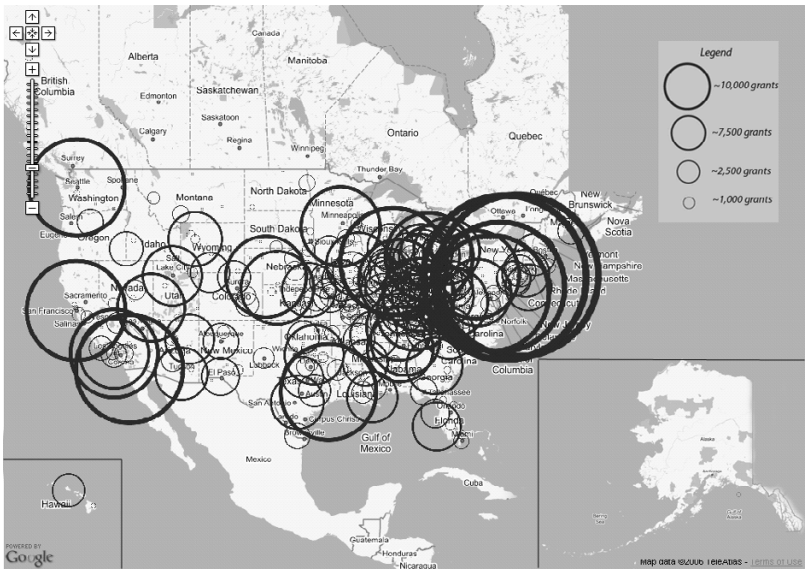

Figure 7. Total number of NIH awards per principle investigator by geo-location (U.S.)

Figure 8. Total number of U.S. patents per assignee by geo-location (U.S.)

Table 2 contains information on the geographic coverage of the Medline, NSF, NIH, and USPTO data sets. Data mining was performed to find all records where both: (1) first author (Medline), primary investigator (NSF, NIH), or assignee (U.S. Patent); and (2) either a zipcode or the city *and* state, were present. Each zipcode in these results was then correlated against the CF Dynamics zipcode database generated from the U.S. Postal Service zipcode list to determine extant geographical coordinates (i.e., *geocodes*; latitude and longitude coordinates). The zipcodes include recent additions or updates since 1999 when made available by the U.S. Postal Service [NOVAK, 2006]. When a zipcode was not found, but a city/state pair was present in a particular data set (e.g., U.S. Patent and NIH), the city and state data were correlated and grouped against the CF Dynamics data to determine and assign a central zipcode from the cluster of zipcodes in the result from the zipcode database.

Table 2. U.S. geographical coverage of publication data sets

Dataset	# Records	# Records without assigned zipcode	# Records with assigned zipcode	# Records with assigned zipcode having geo-coded location	# Records with assigned zipcode not having geo-coded location
Medline	13,149,741	1,841,437	5,654,152	5,584,991	69,161
NSF	174,835	0	174,835	47,675	127,160
NIH	1,043,804	5	1,043,797	878,112	165,685
USPTO	3,374,843	146,706	3,228,137	3,037,482	190,665
Total	17,548,310	1,988,148	10,100,921	9,548,260	552,671

Even though major funding and publication patterns are, as one might expect, concentrated in major developed areas where research centers, such as universities, research labs, hospitals, etc., are more likely to exist, the maps above can provide a quick overview of what one might term as the *scientometric* gaps, i.e. the gaps between publication metrics, such as number of publications per year or one's impact factor, and previously allocated funding. For example, in the case of NSF or NIH data, these gaps can then be examined in the context of new funding applications and policy initiatives administered by national, state, provincial, or local entities. It must be noted that the above geographic coverage maps are only the initial versions of current on-going research. More robust historical analyses, tied to news events, and predictive models are under work.

Topic coverage

The topical coverage of four databases is exemplarily given in Table 3.

Interestingly, a query that matches Medline journals and JCR journals based on ISSN numbers results in only 3,547 matches. This is partially due to the fact that journals can have multiple ISSN numbers and Medline and Thomson might not use the very same ones. Matching based on journal names is even worse as abbreviations and omissions differ among the databases under consideration.

Table 3. Topic coverage of publication data sets

Dataset	# Journals	Description
Medline (1994–2004)	6,991	Paper level data covering biomedical research.
Physical Review	9	Paper level data covering major physics research.
PNAS	1	Paper level data covering mostly biomedical research.
Journal Citation Report (1994–2004)	9,227	Journal level data covering the physical sciences.

Embedded business logic: Network science module

One of the newest features being researched and developed for the Scholarly Database involves the addition of modules containing application business logic. Though a somewhat ambiguous term, *business logic* generally refers to the functional algorithms and processing methods which handle the exchange of information between a database and a user interface under a three-tier architecture. Using a traditional design pattern or even a linear approach, it generally encompasses all of the functional programming extant between the display layer and the database layer in most common web applications.

As a majority of our research is focused on network science, we are in the alpha-stage of developing a module based on a series of stored procedures that can build output representations of networks based on currently popular formats such as edgelists, graphml, .net, etc. These representations can then be saved locally by the user for use in external programs or used from *within* the database, the latter by calling other modules that reference external programs (e.g., R, gnuplot, python, perl, etc.).

Figure 9 shows how the network science module works. As only three fields are required to build a basic network, the input to this module is simply a set of user-selected fields in our database and a string representing the type of network one wishes to build (e.g., .net, graphml). Various stored procedures which join and group the data are called depending on the particular type of network being built. As it is being generated, other methods from other modules can be called to check for such things as data integrity, errors or patterns of errors in the input data, singletons, entity disambiguations, encoding transformations, etc. The final output representation is then returned to the user, with the additional functionality of storing this data in appropriate tables, etc., in a schema in the user's account whereby they can use this data at a later time.

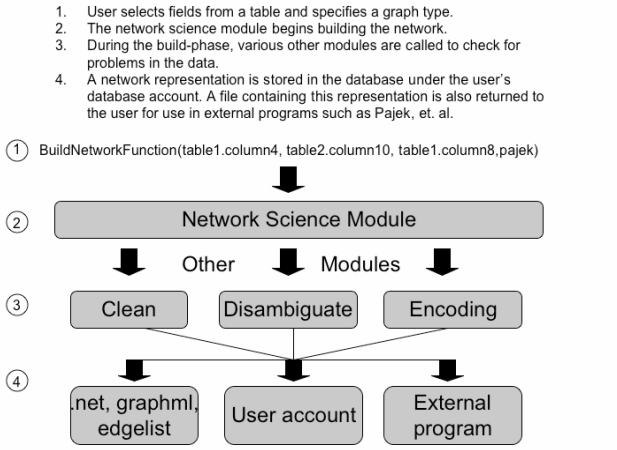

Figure 9. Diagram of how the network science module works

With the ability to also call external programs, a user has access to call all of the packages, libraries, modules, etc., available for those programs under the local installation. This provides a wealth of functionality not extant in the database itself which both the user and our staff can use to enrich one's final product. Such applications would include the ability of a user to return to the screen a graph of the degree distribution of a particular network or even apply statistical modelling methods

to their self-generated network data. Via simple controls within the web application UI, therefore, the scope of this particular line of development we feel is very exciting and promising and has the ability to empower a user to develop and embed their own desired programmatic functionality within the database instead of relying on external programmers at the middle-tier.

Distributed data testbed

One major aspect of the Scholarly Database that is being actively promoted is its use as a common, shared scholarly data testbed among researchers and organizations alike. One of the initial goals for creating the database was its common utility among researchers within a given department. These researchers would all have access to the same data and tools whereby their results and the results of their colleagues could be evaluated in a collaborative fashion. As the database has grown, external interest has also increased necessitating the need to find a way of sharing data sets across departments, other colleges and universities, and organizations alike.

By Summer 2008, we expect to serve ten major data sets – about 20 million records in total. We plan to make the open access parts of the database available as a testbed for information science research in database design, data integration, data provenance, data analysis & mining, data visualization, and interface design, to name just a few. This will require detailed documentation and close collaborations with contest organizers. In return, it is very likely to result in more sophisticated data harvesting, preservation, integration, analysis, and management algorithms that are urgently needed to provide better knowledge access and management tools for scholars, practitioners, policy makers, and society at large.

While the current version of the Scholarly Database resembles a data warehouse where all data is stored locally, this approach won't scale to capture all scholarly publications. A more distributed setup will be needed that interlinks major databases and mirrors the diverse data analysis and visualization services in preparation. Hence, another line of research will explore means to decentralize the Scholarly Database, e.g., by running it as a peer-to-peer service.

Part-and-parcel with this development, collaborations have evolved across these different organizational units whereby a shared, distributed scholarly data testbed is slowly starting to evolve. Instead of duplicating the effort of acquiring a freely given distributed data set and the implicit activities there to, various researchers see the need for a common database that can share both this type of data and its associated metadata or tools. Notwithstanding this alone, these researchers also acknowledge the value of creating arrangements to share their own, previously internal, data. This common desire has led to the need for a distributable version of the Scholarly Database that can be both used and enriched locally, but whose data and access can be shared by all and whose

development can be enhanced by all. Due to these developments, at the end of 2007 the database will no longer encompass only data specific to the United States or the English language, but will also include similar data sets from such countries as Japan and China.

Such a testbed also supports the notion of furthering the rigor and richness of the scientific pursuits mentioned in the introduction, and having a common, shared testbed can only further the ultimate goals of Science and scientists alike. As was the case with the evolution of the Internet, such a testbed will open up collaborative opportunities to not only share data, but also the researchers themselves, providing new opportunities to share in a new discovery or otherwise become aware of research outside of one's home country or known language.

Discussion and future work

The Scholarly Database addresses a central need for the large-scale study of science: access to high quality, centralized, comprehensive scholarly data sets. As discussed, bibliometrics and scientometric studies often require access to multiple data sets – either across domains, e.g., to determine the citation interaction among education and psychology, or across data types, e.g., to identify the influence of grant funding on paper and citation output [BOYACK & BÖRNER, 2003]. Unfortunately, today's scholarly data sets are kept in data silos that are hardly interlinked making it hard to retrieve all records for one person, institution, nation, etc.. In addition, one publication might be stored in diverse databases in a different state, e.g., a paper could be available via Medline, Citeseer, and Thomson Scientific ISI. Interestingly, each of the three databases has different data characteristics. Medline has annotations of genes and proteins. Citeseer provides full text and identified citation relations to other Citeseer papers. Databases provided by Thomson Scientific provide high quality citation data and citation relations to other papers within their holdings. Collectively, these issues cause major data integration and provenance problems.

Concurrent to the work being done on the Open Archives Initiative (OAI), one solution we are working on concerning data integration is the creation of an internal metadata framework that will encompass common relations between various scholarly data sets. Because a large number of fields are found to be common among scholarly data sets, the metadata framework serves as a crosswalk or lookup table for defining and describing those corresponding fields between various datasets. Our solution will incorporate, where possible, any pre-existing metadata descriptions from the OAI, as well as unique field definitions of data sets within the Scholarly Database.

Data provenance referring to the process of tracing and recording the origins of data and its movement between databases is another important issue. Some common approaches involve many data authorities, and subsequent users, to affirm the provenance of the data in question. Our current approach has been to acquire data sets

wholly from their creators, i.e., the National Science Foundation, National Institutes of Health, etc. and then verify our collections against the collections of other collaborating institutions, researchers, and OAI repositories [HITCHCOCK & AL., 2003].

In order to provide reliable access to a major subset of our collective scholarly knowledge, freely sharable data must be distributed to as many geographically diverse institutions as possible. Therefore, the Scholarly Database has been designed to be easily setup and mirrored at other host institutions, whereby local, country-specific, or proprietary data sets can be excluded or added. Such ease-of-use has the potential to provide: (1) increased access to under-represented users that may not have had prior access to such a diverse collection of scholarly data; and (2) increased accuracy and integration of data as increased access to these mirrored instances makes it easier for researchers to collaborate and identify erroneous records.

References

ATKINS, D. (2003), *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* Technical Report. Retrieved December 5, 2006, from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203

BÖRNER, K., CHEN, C., BOYACK, K. W. (2003), Visualizing knowledge domains. In: B. CRONIN (Ed.), *Annual Review of Information Science & Technology*, 37. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, pp. 179–255. http://www.infotoday.com/books/asist/arist37.shtml

BOYACK, K. W., MANE, K. K., BÖRNER, K. (2004), Mapping medline papers, genes, and proteins related to melanoma research. In E. BANISSI (Ed), *Proceedings of the Information Visualisation, Eighth International Conference on (IV'04)*. IEEE Computer Society, Washington, DC, pp. 965–971.

BOYACK, K. W., BÖRNER, K. (2003). Indicator-assisted evaluation and funding of research: visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology*, 54 (5) : 447–461.

BOYACK, K. W., WYLIE, B. N., DAVIDSON, G. S. (2002), Domain visualization using VxInsight for science and technology management, *Journal of the American Society for Information Science and Technology*, 53 (9) : 764–774.

CHEN, C. (2002), *Mapping Scientific Frontiers*. Springer-Verlag, London.

GARFIELD, E. (1955), Citation indexes for science: A new dimension in documentation through association of ideas, *Science*, 122 (3159) : 108–111.

HITCHCOCK, S., BRODY, T., GUTTERIDGE, C., CARR, L., HARNAD, S. (2003), The impact of OAI-based search on access to research journal papers, *Serials*, 16 (3) : 255–260.

KLAVANS, R., 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.

NOVAK (2006), *CF Dynamics Zip Code Database*. Retrieved December 4, 2006, from http://www.cfdynamics.com/zipbase/

SHIFFRIN, R. M., BÖRNER, K. (Eds.) (2004), Mapping knowledge domains, *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Suppl_1) : 5183–5185.

TICHENOR, S. (2006), Application software for high performance computers: A soft spot for U.S. business competiveness, *CTWatch Quarterly*, 2 (4A) : 1–6. http://www.ctwatch.org/quarterly/articles/2006/11/application-software-for-high-performance-computers-a-soft-spot-for-us-business-competitiveness/

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.