La Rowe, Gavin, Ambre, Sumeet, Burgoon, John, Ke, Weimao and Börner, Katy.
(2007) The Scholarly Database and Its Utility for Scientometrics Research.
In Proceedings of the 11th International Conference on Scientometrics and Informetrics,
Madrid, Spain, June 25-27, 2007, pp. 457-462.

# The Scholarly Database and Its Utility for Scientometrics Research[1]

Gavin LaRowe, Sumeet Ambre, John Burgoon, Weimao Ke & Katy Börner

glarowe@indiana.edu, sambre@indiana.edu, jburgoon@indiana.edu, wke@indiana.edu, katy@indiana.edu
Indiana University, School of Library and Information Science, 10th Street & Jordan Avenue, Bloomington, IN 47405 (USA)

## Abstract
The Scholarly Database (SDB) at Indiana University aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale scholarly datasets. This database focuses on supporting large studies of changes in science over time and communicating findings via knowledge-domain visualizations. The database currently provides access to around 18 million publications, patents, and grants, ten percent of which contain full-text abstracts. Except for some datasets with restricted access conditions, the data can be retrieved in raw or pre-processed format using either a web-based or relational database client. This paper motivates the need for the database from bibliometric and scientometric perspectives (Cronin & Atkins, 2000; White & McCain, 1989). It explains the database design, setup, and interfaces as well as the temporal, geographical, and topic coverage of datasets currently served. Planned work and the potential for this database to become a global test bed for information science research are discussed.

## Keywords
Research database, geospatial coverage, data integration, bibliometrics, scientometrics, mapping science

## Introduction
Digitized scholarly datasets and sufficient computing power to integrate, analyze, and model these datasets make it possible to study the structure and evolution of science on a global scale (Börner, Chen, & Boyack, 2003; Boyack, Klavans, & Börner, 2005; Shiffrin & Börner, 2004). Results can be communicated via tables, graphs, geographic and topic maps. Frontiers emerging across different sciences can be discovered and tracked. Different funding models can be simulated and compared. School children can start to understand the symbiotic relationships among different areas of science.

The study of science on a global scale requires access to high quality, high coverage data and major cyberinfrastructure (Atkins et al., 2003) to process such data. Many studies require the retrieval and integration of data from different sources with differing data types. For example, input-output studies require input data, e.g., funding amounts and number of new graduates, and output data, e.g., the number of publications, received citations, awards, and policy changes. Unfortunately, the identification and inter-linkage of unique authors, investigators and inventors is non-trivial.

Contrary to other scientific disciplines where data is freely and widely shared, there are very few bibliometric or scientometric test datasets available. This makes it very time consuming (e.g., data download, cleaning and inter-linkage) or impossible (if datasets require access permissions) to replicate studies or reproduce results. Fortunately, some services and institutions, such as PubMed, CiteSeer, arXiv, and the United States Patent Office, provide free data dumps of their holdings under certain conditions. However, most bibliometric and scientometric scholars are not trained in parsing millions of XML-encoded records and very few have expertise in the setup and maintenance of multi-terabyte databases.

The Scholarly Database, online at https://sdb.slis.indiana.edu/, aims to improve the quality and reduce the costs of bibliometric and scientometric research and practice by providing easy access to high quality, comprehensive scholarly datasets.

[1] This work was supported by the National Science Foundation under Grant No. IIS-0238261, IIS-0513650, IIS-0534909, and CHE-0524661 and a James S. McDonnell Foundation grant in the area Studying Complex Systems entitled "Modeling the Structure and Evolution of Scholarly Knowledge". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. We thank Kevin W. Boyack and Russell Duhon for insightful comments on an earlier version of this paper.

**Database Architecture and Access**

The Scholarly Database comprises production, development, and research systems. The *production system* utilizes two Sun V490 servers with sixteen gigabytes of memory each serving two redundant, mirrored clusters of the Scholarly Database running PostgreSQL v8.1.4 under Solaris 10. Failover, load balancing, and replication are provided via a planned intermediary system service between each of these clusters. In addition, each instance of the database is isolated within its own zone under Solaris 10, providing increased security and failover for the system at-large. The *development system* is used for developing future production versions of various datasets. After post-processing, incorporation, and testing, new datasets receive their own data store and various schema templates are incorporated. When approved, these datasets are then pushed to one of the production clusters. The *research system* runs on a Sun V480 with thirty two gigabytes of memory providing a sandbox for researchers to develop and refine personal or proprietary datasets. Aside from being a completely isolated system, this cluster provides users with sufficient memory and disk storage for large-scale cyberinfrastructure projects involving multi-gigabyte and, in the future, terabyte-scale data and application requirements. All three systems are hosted at the Cyberinfrastructure for Network Science Center at Indiana University.

The general system architecture of the Scholarly Database has three major parts: The *Data Space* stores raw data, pre-processed data, metadata, and any other raw data artefacts. *Data Services* support data harvesting, data mining, pre- and post-processing, statistical analysis, and in the near future natural language processing, multi-agent data simulations, and information visualization services. Aside from backup and storage, *Data Provenance* is provided via the use of metadata associated with the raw data, internal database artefacts (e.g., schemas, tables, views, etc.), and user artefacts such as queries or views.

The database schema is too extensive to describe or depict here in anything but an abstract overview. The current implementation utilizes three schemas: public, base, and auxiliary. The *public schema* describes all post-processed raw data that has been loaded into the database. This data and associated metadata are rarely modified, except when new updates are received from a data provider. The *base schema* contains all foundational views found in the web interface used for search, display, and download of data. Most views are virtual, but some materialized tables are used for extremely large datasets. The *auxiliary schema* provides a space where schemas, tables, views, and functions created by the users of the system can be stored for re-use. Its ancillary purpose is to provide an area where one-time or proprietary non-public components (e.g., views or functions) can be explored for a dataset by an authorized user.

Access to the database is available via a web front-end at https://sdb.slis.indiana.edu/, a pgAdmin PostgreSQL administration and management tool, and a psql PostgreSQL interactive terminal. The front-end interface allows external users to search through all of the articles and documents in the database via author, title, or keyword for a specified year range. Results are returned showing generic fields such as journal name, title, author name, and date of publication. Each record can be further expanded to show fields specific to a given dataset. Selected datasets can be downloaded. Future services will support the extraction and download of networks such as co-author and paper-citation networks.

**Dataset Acquisition, Processing and Coverage**

The Scholarly Database is unique in that it provides access to diverse publication datasets and to patents and grant award datasets. Datasets are acquired from a wide variety of sources. Some are one time acquisitions. Others are updated on a continuous basis. Several have access restrictions. Table 1 provides an overview.

*Medline* publications provided by the National Library of Medicine (http://www.nlm.nih.gov), consists of two types of data: baseline files that are distributed at the end of each year and include all PubMed records that have been digitally encoded in XML for Medline; and newly added data for that particular year which is subsequently updated in future baseline releases. It is provided in XML format with a custom DTD. Update files are provided regularly.

Table 1: Datasets 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** | | | |

***Physical Review*** papers provided by the American Physical Society (http://aps.org) come in 398,005 XML-encoded article files covering nine journals (A, B, C, D, E, PR, PRL, RMP, PRST, and AB) over a one hundred and ten-year time span: 1893-2006. A single DTD exists for the entire collection. It encompasses all changes made throughout the history of the digital encoding of these files that were previously available in SGML format. It is a proprietary dataset that cannot be shared or used for commercial purposes.

***Proceedings of the National Academy of Sciences*** provided by PNAS (http://www.pnas.org), comprise full text documents covering the years 1997-2002 (148 issues containing some 93,000 journal pages). The dataset is also available in Microsoft Access 97 format. It 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 anybody registered for that Colloquium and who signed the copyright form. It cannot be redistributed without prior permission from PNAS. It cannot be used for commercial purposes.

***Journal Citation Report (JCR–Science Edition)*** dataset by ISI Thomson Scientific (http://www.isinet.com) comprises two datasets: (1) covers the years 1994-2004; and (2) contains cited and citing pairs records for 1974, 1979, 1984 and 1989 from the Science Citation Index – Expanded (SCI-E). Both are restricted use for the purpose of an NSF grant. This data cannot be used, distributed, or otherwise shared without prior written permission from Thomson Scientific.

***Patents*** by the United States Patent and Trademark Office (USPTO) (http://www.uspto.gov) come as XML-encoded dumps downloadable from the USPTO website (ftp://ftp.uspto.gov/pub/patdata/). This is a publicly accessible dataset of about three million records organized into ca. 160,000 patent classes.

***NSF Grants*** awarded by the National Science Foundation (NSF) (http://www.nsf.gov) support research and education in science and engineering to more than two thousand colleges, universities, and other research and education institutions in all parts of the United States through grants, contracts, and cooperative agreements. It is composed of raw text files as distributed by the NSF for the years listed above. It is a publicly accessible dataset.

***NIH Grants*** *data* from the National Institutes of Health (NIH) (http://www.nih.gov) is composed of CRISP and Awards data downloaded from the main NIH web site and the CRISP on-line search engine http://crisp.cit.nih.gov/ for the years listed above. The CRISP data includes information regarding extramural projects, grants, contracts, and so on associated with projects and research supported by the National Institutes of Health. NIH award data is composed of principal investigator and institution NIH grant award amounts concerning projects found in the CRISP data for the years listed above.

Detailed information on these datasets, and their quality and coverage, as well as available data fields, is available online at https://nwb.slis.indiana.edu/community/ (select 'Datasets'). New datasets are added on a continuous basis.

***Temporal Coverage.*** As can be seen in Fig. 1 (left), the Medline dataset has the most records per year, with about 500,000 new records each year. There are about 200,000 new USPTO patents each year, 10,000 new NSF awards, and 50,000 new NIH awards per year. The number of unique authors per year is shown in Fig. 1 (right). A concatenation of first author name and last author name was employed to identify unique authors. There is more than one 'John Smith' in the Medline dataset, and we know that some authors change names, but the graph provides a rough estimate of how many unique authors contribute to the growth of each dataset and the increase in the number of authors over time.

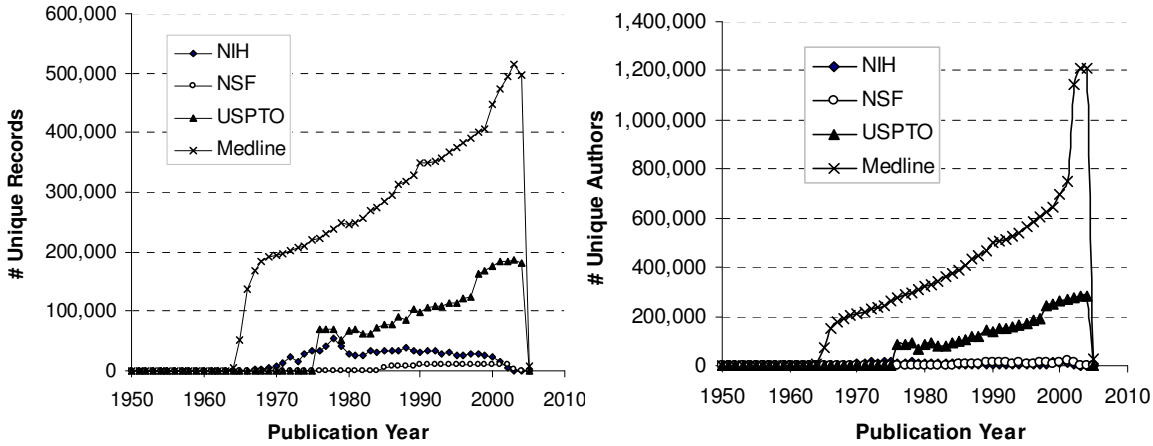


Figure 1: Number of records and unique authors per year for different datasets for 1950 to 2005.

***Geographical Coverage.*** The geographical coverage of datasets can be examined by geolocating papers, patents, and awards based on the affiliation of their authors, inventors, and awardees. It is frequently not clear how best to divide the contributions of single authors across the team. In some datasets, affiliation data is only available for the first author. Therefore, we attribute the location of the paper, patent, or award to the first author, inventor, or awardee. For each first author, inventor, or awardee we retrieved either a zip code or a city-state pair. Zip codes were matched against zip code data provided by Novak Banda at http://www.populardata.com to derive latitude and longitude coordinates. When a zip code was not available, all zip codes for the city-state pair were retrieved and a geospatially central zip code was assigned and geolocated. As we did not have access to world wide geocoding services, this analysis is restricted to US. Due to page limitations, we prototypically plot the coverage of only two datasets: Medline and NIH, in Fig. 2.

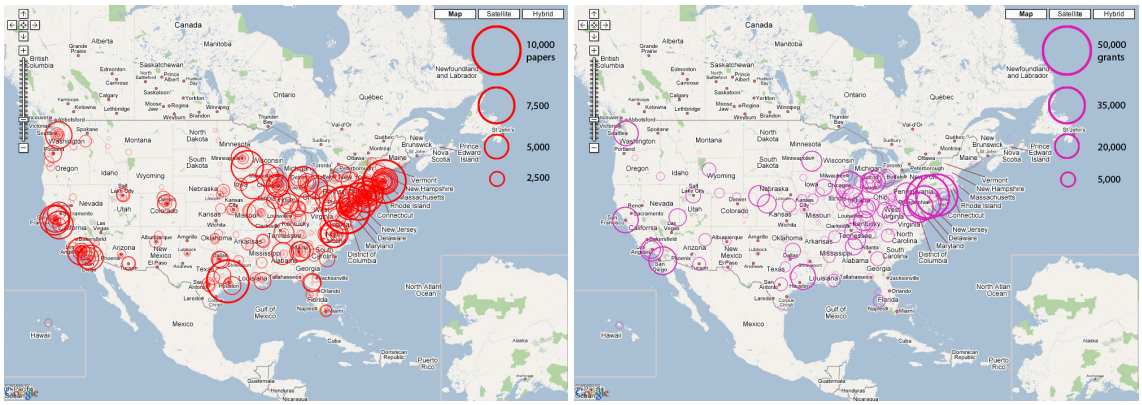


Figure 2: Number of Medline publications and NIH awards by geo-location (U.S. only)

Note that only 1,420,816 of the 13,149,741 Medline publications had an US affiliation. Out of those, only 1,036,865 had zip codes. There were 13,188 unique zip codes and 10,450 of those could be geolocated. As for NIH, 971,754 main awardees had city-state pairs that were used to identify 1,986 unique geolocations. Fig. 2 shows that major funding and publication patterns are concentrated in

urban areas where research centers, such as universities, research labs, hospitals, etc., are more prone to exist.

***Topic Coverage.*** Interested to see the topic coverage of different datasets, we tried to identify the number of journals that the different publication databases cover. In particular, we ran a query that matched Medline journals and JCR journals based on ISSN numbers. However, there were only 3,547 matches. This is partially due to the fact that journals can have multiple ISSN numbers and Medline and Thomson Scientific data might not use the same ones. Matching based on journal names is even more difficult, because abbreviations and omissions differ among the databases under consideration. Medline covers 6,991 unique journals and JCR has 9,227 unique journals from 1994 to 2004.

## 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 datasets. The value and quality of this database will depend on its adoption by the community. The more scholars and practitioners use it, the more likely it is that missing records or links will be discovered, important datasets will be integrated, the best (author/institution/country/geo-code) unification algorithms can be applied, and research studies are conducted, replicated, and verified.

The rate of adoption will greatly depend on the utility and usability of the SDB. Hence, future work aims to make the SDB easier to use and easier to extend by adding new datasets and services. Concurrent to the work being done on the Open Archives Initiative (OAI) (Bekaert & Sompel, 2006), we are working on an internal metadata framework that will encompass common relations between various scholarly datasets. This metadata framework will ease schema matching between datasets. Our solution will incorporate, where possible, any pre-existing metadata descriptions from the OAI and other standards. In order to provide reliable access to non-proprietary data, the Scholarly Database has been designed for easy mirroring in geographically distinct locations. All software used is open source and the database setup is documented in detail to ease installation.

We expect to serve ten major datasets by summer 2007 – about 20 million records. Plus, the open access parts of the database will be made available for information science research in database design, data integration, data provenance, data analysis, data mining, data visualization, and interface design. This will require close collaboration with many researchers, practitioners, and dataset providers. In return, we expect to gain access to more sophisticated data harvesting, preservation, integration, analysis, and management algorithms that are urgently needed to improve data access and management tools for scholars, practitioners, policy makers, and society at large.

## References

Atkins, D. E., Drogemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messian, P., Ostriker, J. P., & Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: National Science Foundation.

Bekaert, J., & Sompel, H. V. d. (2006). Augmenting Interoperability Accross Scholarly Repositories, *Meeting sponsored and supported by Microsoft, the Andrew W. Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation, and the Joint Information Systems Committee.* New York, NY. Retrieved from http://msc.mellon.org/Meetings/Interop/FinalReport on 2/15/2007.

Börner, K., Chen, C., & Boyack, K. (2003). Visualizing Knowledge Domains. In B. Cronin (Ed.), *Annual Review of Information Science & Technology* (Vol. 37, pp. 179-255). Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the Backbone of Science. *Scientometrics, 64*(3), 351-374.

Cronin, B., & Atkins, H. B. E. (2000). *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*: American Society for Information Science & Technology.

Shiffrin, R. M., & Börner, K. (Eds.). (2004). *Mapping Knowledge Domains* (Vol. 101 (Suppl. 1)): PNAS.

White, H. D., & McCain, K. W. (1989). Bibliometrics. In M. E. Williams (Ed.), *Annual Review of Information Science & Technology* (Vol. 24, pp. 119-186). Amsterdam, Netherlands: Elsevier Science Publishers.