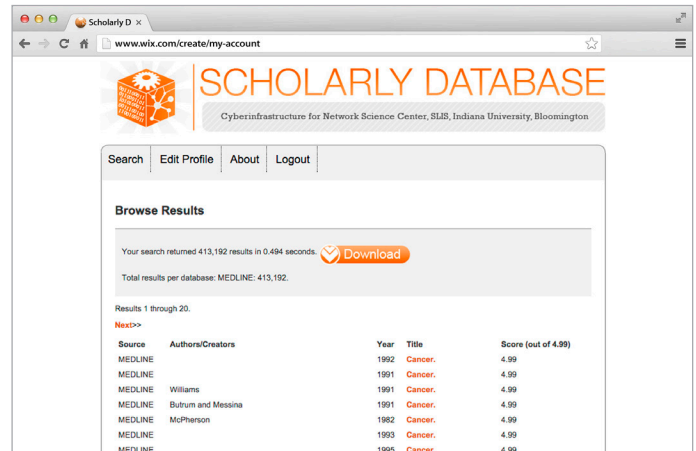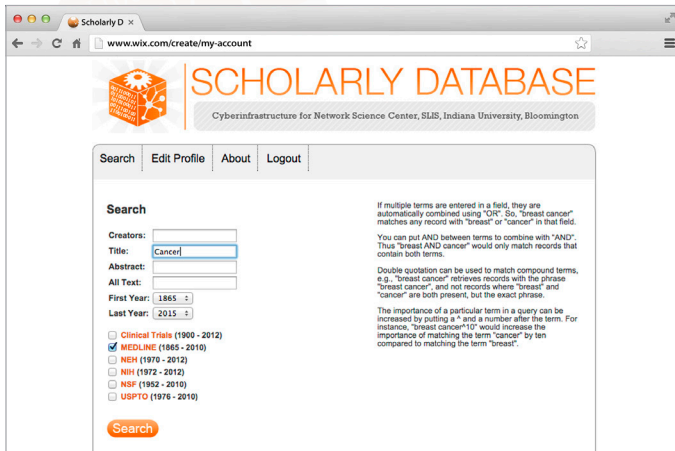# Scholarly Database

Investigator
Dr. Katy Börner
(katy@indiana.edu)
Indiana University

Project Lead
Nianli Ma
(nianma@indiana.edu)
Indiana University

Developer
Russell Duhon
(rduhon@indiana.edu)
Indiana University

Introduction

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. The online interface at provides access to four datasets: Medline papers, U.S. Patent and Trademark Office patents (USPTO), National Science Foundation (NSF) funding, and National Institutes of Health (NIH) funding – over 20 million records in total. Users can register for free to cross-search these databases and to download result sets as dumps for scientometrics research and science policy practice.

Getting Started

New users should register at **sdb.cns.iu.edu**. Approved users should go to **sdb.cns.iu.edu** and login with their email and password.

Functions

Search: SDB supports search across paper, patent, and funding databases. To search, enter search term(s) in creators (author/awardee/inventor), title, abstract, full text (keywords and other text) fields, select a year range and database(s).

Downloading: Search results retrieved from different databases can be downloaded as data dump in csv file format. Note that each database has a different download format. The download limit can be found in your personal profile. If you need to download a larger number of records, please contact the SDB team at **cns-sdb-dev-l@indiana.edu** with details of your intended usage.

Edit Profile: Please update your personal information regularly. The profiles are used to report the utility of the SDB to private and federal sponsors which fund the development, deployment, and maintenance of the SDB.

System Architecture

All data is stored in a PostgreSQL (**postgresql.org**) database. Full-text search is supported using Solr (**lucene.apache.org/solr**) to index the contents of the database. Solr is an industry-standard, open source search server that can scale to very large amounts of data using replication and sharding. The online interface was developed in Django (**djangoproject.com**). Django is a web framework written in the Python (**python.org**) programming language with particularly good support for content-oriented web applications.

Get Invovled

If you would like to be added to the Scholary Database listserv, send an email to **listserv@indiana.edu** with a blank subject line and in the body type "subscribe cns-sdb-l@listserv.indiana.edu".
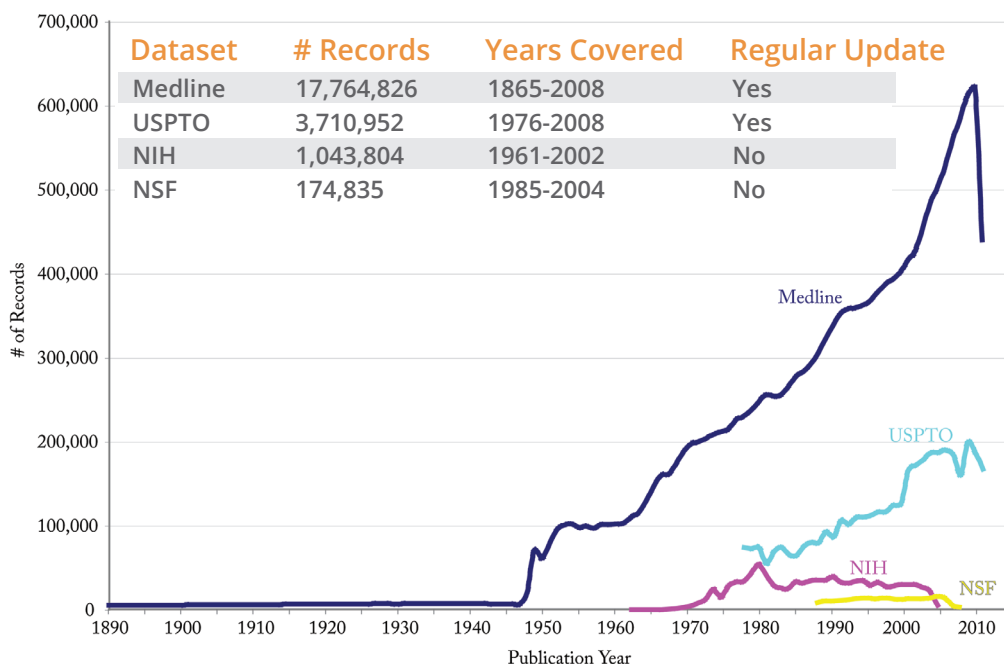
Acknowledgements

The SDB team comprises Nianli Ma, Russell J. Duhon, Elisha F. Hardy, and Katy Börner at the Cyberinfrastructure for Network Science Center, Indiana University. The initial database was setup by Gavin La Rowe, Sumeet Ambre, John Burgoon, and Weimao Ke.

The Scholarly Database is funded by the School of Library and Information Science and the Cyberinfrastructure for Network Science center at Indiana University, the National Science Foundation under Grants No. IIS-0238261 and IIS-0513650, and a James S. McDonnell Foundation grant in area Studying Complex Systems.

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 National Science Foundation.

sdb.cns.iu.edu

Data

Dataset	# Records	Years Covered	Regular Update
Medline	17,764,826	1865-2008	Yes
USPTO	3,710,952	1976-2008	Yes
NIH	1,043,804	1961-2002	No
NSF	174,835	1985-2004	No

Medline Papers

Medline is the National Library of Medicine (NLM) premier bibliographic database that contains references to journal articles in the life sciences with a concentration on biomedicine. A distinctive feature of Medline is that the records are indexed with NLM's Medical Subject Headings (MeSH).

The database contains citations from 1950 to the present, with some older material. New citations that have been indexed with MeSH terms, publication types, GenBank accession numbers, and other indexing data are available daily and display with the tag. Medline licensing information can be found **www.nlm.nih.gov/databases/journal.html**.

U.S. Patent and Trademark Office Patents

For over 200 years, the United States Patent and Trademark Office (USPTO) has been processing and disseminating patent and trademark applications and information to promote an understanding of intellectual property protection and to facilitate the development and sharing of new technologies worldwide. The USPTO home page is at **www.uspto.gov.**

National Institutes of Health Awards

CRISP (Computer Retrieval of Information on Scientific Projects) is a searchable database of federally funded biomedical research projects conducted at universities, hospitals, and other research institutions. The database is maintained by the Office of Extramural Research at the National Institutes of Health. The main CRISP site is at **www.nsf.gov/awardsearch**.

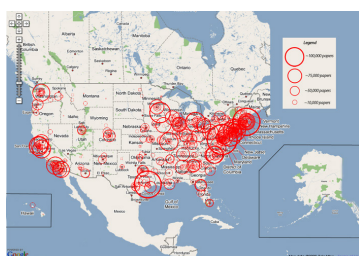## National Science Foundation Awards

The National Science Foundation (NSF) funds research and education in science and engineering. It does this through grants, contracts, and cooperative agreements to and with more than 2,000 colleges, universities, and other research and/or education institutions in all parts of the United States. The NSF Awards search page is at **www.nsf.gov/awardsearch**.
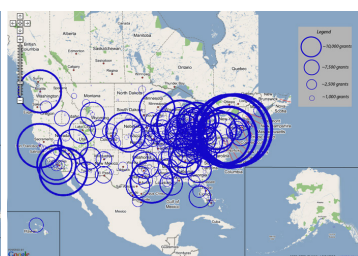
Data Documentation

Each dataset has a Wiki page linked from the SDB site which gives more detailed information about the dataset, including the table schema and data coverage.
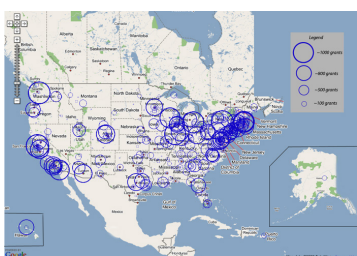
Geography Overlays

Medline Publications   ### NIH Grants   ### US Patents