Science of Science Research and Tools Tutorial #10 of 12

Dr. Katy Börner

Cyberinfrastructure for Network Science Center, Director Information Visualization Laboratory, Director School of Library and Information Science Indiana University, Bloomington, IN http://info.slis.indiana.edu/~katy

With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Hanning Guo, Mark A. Price, Angela M. Zoss, and Scott Weingart

Invited by Robin M. Wagner, Ph.D., M.S. Chief Reporting Branch, Division of Information Services Office of Research Information Systems, Office of Extramural Research Office of the Director, National Institutes of Health

Suite 4090, 6705 Rockledge Drive, Bethesda, MD 20892 10a-noon, July 27, 2010





12 Tutorials in 12 Days at NIH—Overview

1. Science of Science Research 1st Week

2. Information Visualization

3. CIShell Powered Tools: Network Workbench and Science of Science Tool

4. Temporal Analysis—Burst Detection 2nd Week

5. Geospatial Analysis and Mapping

6. Topical Analysis & Mapping

7. Tree Analysis and Visualization 3rd Week

8. Network Analysis

9. Large Network Analysis

10. Using the Scholarly Database at IU

4th Week

11. VIVO National Researcher Networking

12. Future Developments



12 Tutorials in 12 Days at NIH—Overview

[#10] Using the Scholarly Database at IU

- Motivation
- Functionality / Sample Usage
- Implementation
- Documentation
- Outlook
- Exercise: Identify Promising SDB Collaborations

Recommended Reading

- 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. http://ella.slis.indiana.edu/~katy/paper/07-issi-sdb.pdf
- Scholarly Database home page, http://sdb.slis.indiana.edu.

2



12 Tutorials in 12 Days at NIH—Overview

[#11] VIVO National Researcher Networking

- Motivation
- Users, Their Needs, and Usage Scenarios
- Development
- Implementation
- Usage
- Outlook
- Exercise: Identify Promising VIVO Collaborations

Recommended Reading

VIVO home page, http://vivoweb.org

VIVO Conference in NYC in August 2010, http://conferences.dce.ufl.edu/vivo



12 Tutorials in 12 Days at NIH—Overview

[#12] Future Developments

- Validation Studies
- Needed Data/Documentation
- Needed and New Tool Functionality
- Needed Documentation/Tutorials
- Promising Research Questions
- Exercise: Identify Promising Collaborations

Recommended Reading

Börner, Katy (2010) Atlas of Science. MIT Press. http://scimaps.org/atlas

Börner, Katy, Bettencourt, Luis M. A., Gerstein, Mark & Uzzo, Stephen Miles (Eds.), Knowledge Management and Visualization Tools in Support of Discovery. (2009). NSF CDI Initiative Workshop Report, National Science Foundation, Indiana University. http://ww.slis.indiana.edu/cdi2008/whitepaper.html

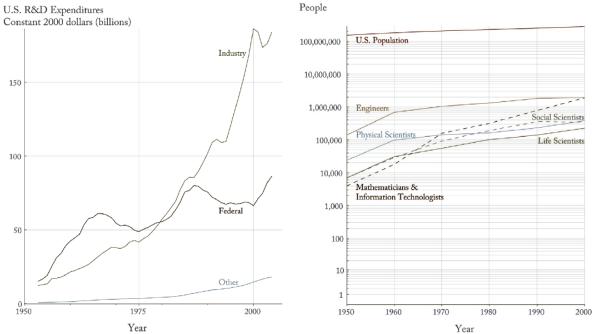
5

[#10] Using the Scholarly Database at IU

- Motivation
- > Functionality / Sample Usage
- > Implementation
- Documentation
- Outlook
- Exercise: Identify Promising SDB Collaborations

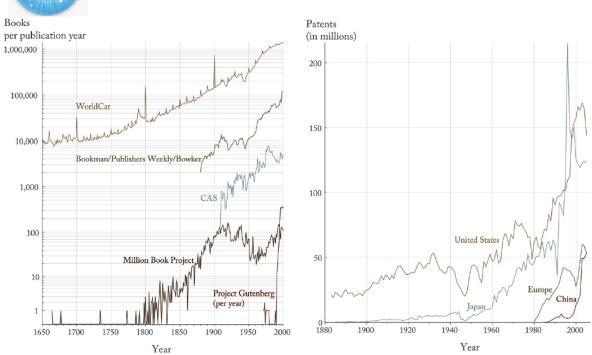


Number of Awards/Funding and Researchers Over Time



Number of Book & Patents Over Time

Börner, Katy (2010) Atlas of Science. MIT Press. http://scimaps.org/atlas

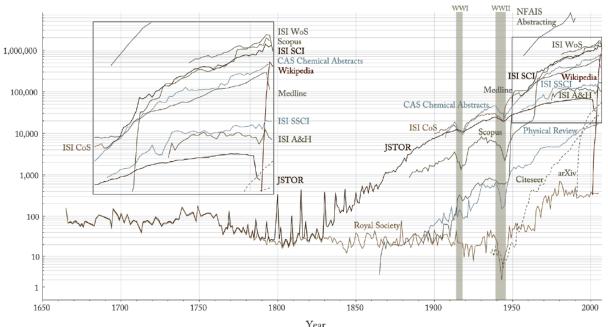


Börner, Katy (2010) Atlas of Science. MIT Press. http://scimaps.org/atlas



Number of Journal Publications (Wikipedia entries) Over Time

Papers & Wikipedia Entries



Börner, Katy (2010) Atlas of Science. MIT Press. http://scimaps.org/atlas

9



Data Needs

Informed science and technology policy (and Science of Science Studies) depend on comprehensive and useful data that has high

- Accuracy
- Integrity (structured & managed)
- Consistency
- Validity (rules, standards are followed)
- Reliability

However, publications, patents, grants are kept in data silos with few interlinkages, incompatible formats, unknown quality and coverage.

Obama Administration is committed to evidence-based policymaking and making data used for policymaking accessible, relevant, and timely. ... Data and analyses should be factual and policy-neutral.

http://www.whitehouse.gov/blog/2010/01/18/science-and-engineering-indicators-2010-a-report-card-us-science-engineering-and-tec

[#10] Using the Scholarly Database at IU

- Motivation
- Functionality / Sample Usage
- Implementation
- Documentation
- Outlook
- Exercise: Identify Promising SDB Collaborations

11



Scholarly Database http://sdb.slis.indiana.edu



Nianli Ma

"From Data Silos to Wind Chimes"

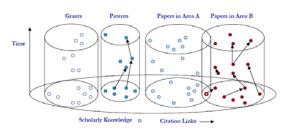
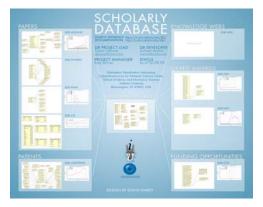
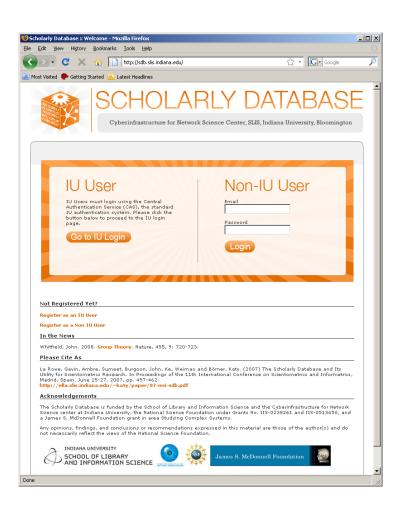


Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type are 'horizontally' interlinked.



- Create public databases that any scholar can use. Share the burden of data cleaning and federation.
- Interlink creators, data, software/tools, publications, patents, funding, etc.

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. http://ella.slis.indiana.edu/~katy/paper/07-issi-sdb.pdf



Online interface:

http://sdb.slis.indiana.edu

Register for free access or test it via tutorial account:

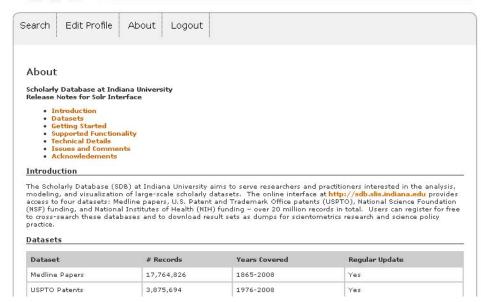
Email: nwb@indiana.edu

Password: nwb

SDB SCHOLARLY DATABASE

Scholarly Database: About





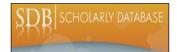


Scholarly Database: # Records, Years Covered

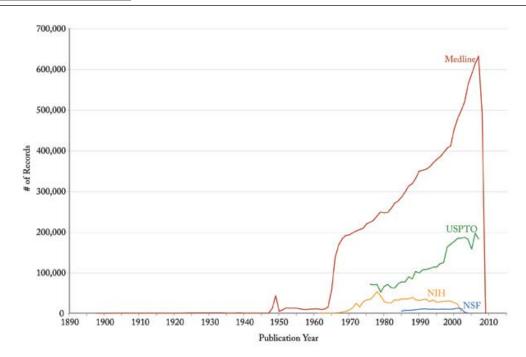
Dataset	# Records	Years Covered	SDB 2.0 Release, Fall 10	Restricted Access
Medline	17,764,826	1898-2008	19,072,547 (1865-2010)	
PhysRev	398,005	1893-2006		Yes
PNAS	16,167	1997-2002		Yes
JCR	59,078	1974, 1979, 1984, 1989 1994-2004		Yes
USPTO	3, 875,694	1976-2008	4,178,196 (1976-2010)	
NSF	174,835	1985-2002	453,687 (1976-2010)	
NIH	1,043,804	1961-2002	1,770,770 (1961-2010)*	
Total	23,167,642	1893-2006		3

NIH awards are not aggregated by base project. Some have up to 3,000 subprojects. http://sdb.slis.indiana.edu/about

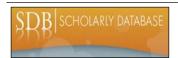
15



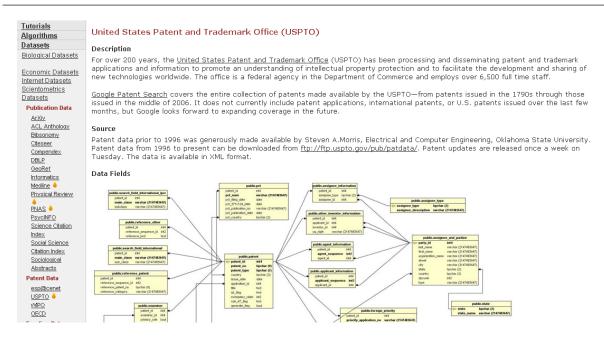
Scholarly Database: Records Per Year



http://sdb.slis.indiana.edu/about



Scholarly Database: Records Per Year



https://nwb.slis.indiana.edu/community/?n=ScientometricsDatasets.USPTO

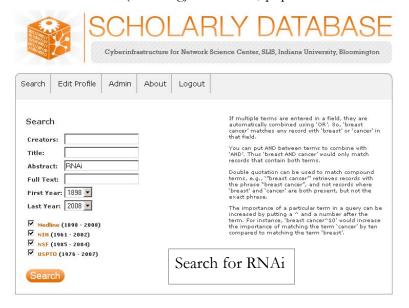
17



Scholarly Database: Web Interface

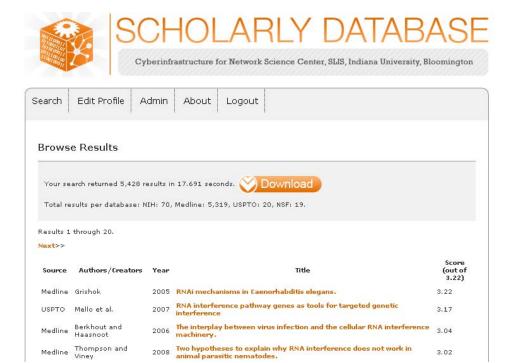
Search across publications, patents, grants.

Download records and/or (evolving) co-author, paper-citation networks.





Scholarly Database: Browse Search Results



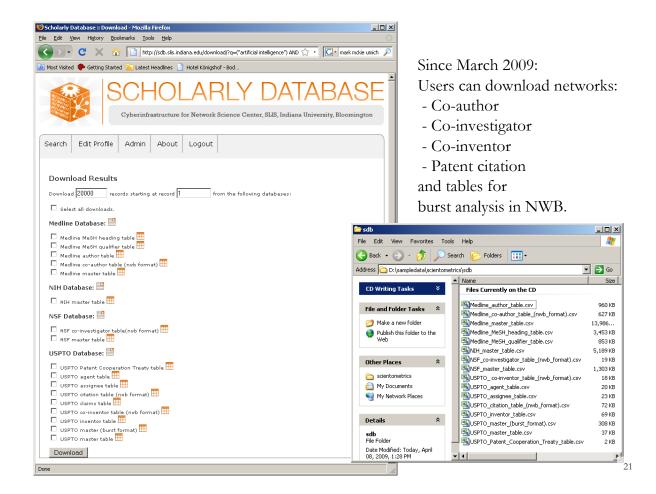
19



Scholarly Database: Download Results



Search Edit Profile Admin About Logout
Download Results Download all Data Dictionary Sample File Download 20000 records starting at record 1 from the following databases: Medline Database: Medline MeSH heading table Medline MeSH qualifier table Medline MeSH qualifier table
Medline author table Medline co-author table (nwb format) Medline master table
NIH Database:
NSF Database:
NSF co-investigator table (nwb format)





Mapping the Field of RNAi Research (SDB Data)

(Sci2 Tutorial, Section 5.2.7)

RNAi	
Time frame:	1865-2008
Region(s):	Miscellaneous
Topical Area(s):	RNAi
Analysis Type(s):	Co-Author Network, Patent-Citation Network, Burst Detection

How many papers, patents, and funding awards exist on a specific topic?

Here we selected research on RNA interference (RNAi) is a system within living cells that helps to control which genes are active and how active they are.

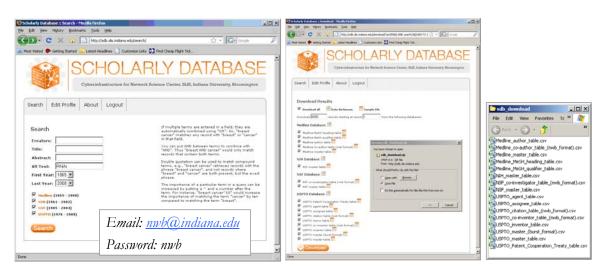
The data for this analysis comes from a search of the Scholarly Database (SDB) (http://sdb.slis.indiana.edu/) for "RNAi" in "All Text" from MEDLINE, NSF, NIH and USPTO. A copy of this data is available in

"yoursci2directory*/sampledata/scientometrics/sdb/RNAi". The default export format is .csv, which can be loaded in the Sci2 Tool directly.



Mapping the Field of RNAi Research (SDB Data)

(Sci2 Tutorial, Section 5.2.7)



The **Scholarly Database** at Indiana University provides free access to 23,000,000 papers, patents, and grants. Since March 2009, users can also download networks, e.g., co-author, co-investigator, co-inventor, patent citation, and tables for burst analysis. For more information and to register, visit http://sdb.slis.indiana.edu.

23



Mapping the Field of RNAi Research (SDB Data)

(Sci2 Tutorial, Section 5.2.7)

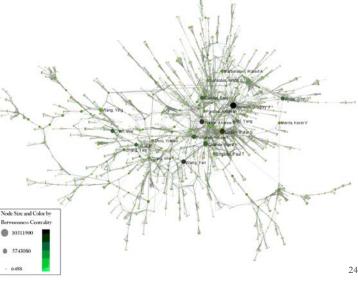
Co-Author Network

Load '*yoursci2directory*/sampledata/scientometrics/sdb/RNAi/Medline_co-author_table_(nwb_format).csv' as a standard csv file. SDB tables are already pre-normalized, so now simply run 'Data Preparation > Text Files > Extract Co-Occurrence Network' using the default parameters.

Network Analysis Toolkit (NAT): 21,578 nodes with 131 isolates, 77,739 edges.

Extract only the largest component by running 'Analysis' Networks > Unweighted and Undirected > Weak Component Clustering.'

Visualize with *GUESS* using '*Layout* > *GEM*'. Use a custom python script to color and size the network.



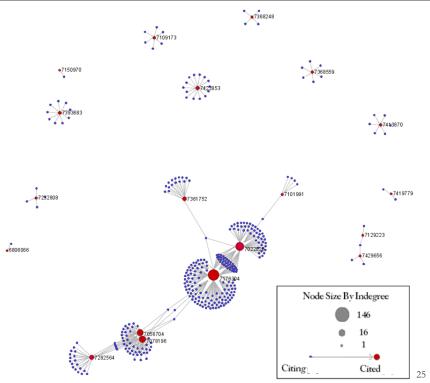


Mapping the Field of RNAi Research (SDB Data)

(Sci2 Tutorial, Section 5.2.7)

Patent Citation Network

To visualize the citation patterns of patents on RNAi, load '*yoursci2directory*/sampledata/scientometrics/sdb/RNAi/USPTO_citation_table_(nwb_format).csv' as a standard csv file and follow the instructions in the tutorial.





Mapping the Field of RNAi Research (SDB Data)

(Sci2 Tutorial, Section 5.2.7)

Topic Bursts

'Load '*yoursci2directory*/sampledat/scientometrics/sdb/RNAi/Medline_master_table.csv'. This table includes full records of MEDLINE papers, and can be used to find bursting terms from MEDLINE abstracts dealing with RNAi.

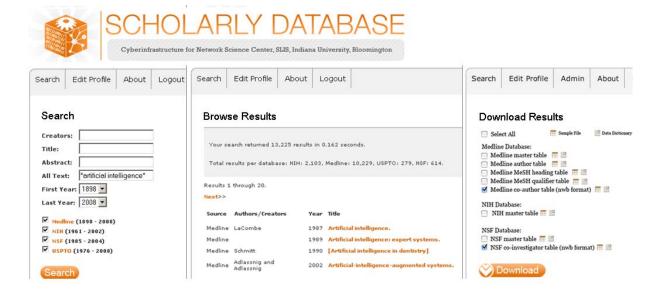
Load the file as a standard csv and run 'Preprocessing > Topical > Normalize Text' with the default separator and the "abstract" box checked. Run 'Analysis > Topical > Burst Detection' with "date_cr_year" in the Date Column and "abstract" in the Text Column, leaving the rest of the values default.

Right click on "Burst detection analysis (date_cr_year, abstract): maximum burst level 1" in the Data Manager and view the file. There are more words than can easily be viewed with the horizontal bar graph, so sort the list by "Strength" and prune all but the strongest 10 words. Save the file as a new .csv and load it into the Sci2 Tool as a standard csv file. Select the new table in the data manager and visualize it using 'Visualize' > Temporal >

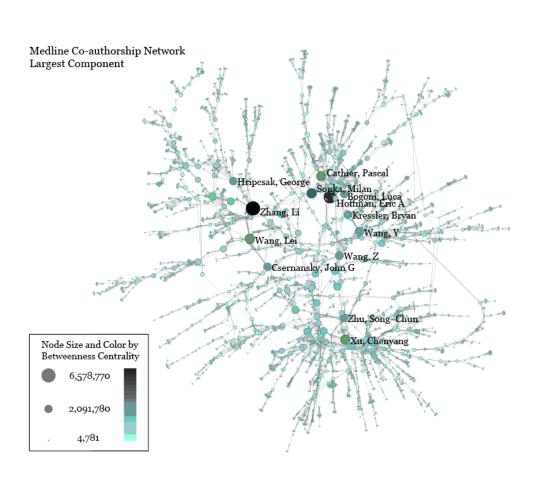
Horizontal Bar Graph.'



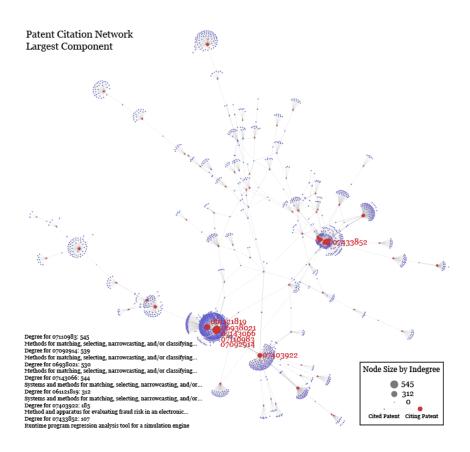
Mapping "Artificial Intelligence Research using SDB Data



Börner, Katy, , Duhon, Russell Jackson &. (2009). Science & Technology Assessment Using Open Data and Open Code. IEEE Intelligent Systems. Vol. 24(4), 78-81, IEEE Computer Systems.



27



Top-10 burst terms from abstracts of the AI search results.

Medline				
Word	Length	Weight	Start	End
medical	17	299.7924	1983	1999
knowledge	5	293.9375	1991	1995
knowledge	6	215.2407	1997	2002
expert	13	171.0443	1985	1997
systems	15	170.3306	1985	1999
intelligence	21	123.9794	1981	2001
patient	21	123.9297	1982	2002
care	12	106.5522	1990	2001
registration	5	104.8139	2005	
knowledge-based	16	98.83778	1987	2002

NSF				
Word	Length	Weight	Start	End
their	6	47.05097	1999	
gray	2	28.19808	2000	2001
learning	2	27.40728	1997	1998
human	5	25.4525	2000	
control	2	24.07877	1992	1993
knowledge	1	21.48756	1998	1998
students	1	21.07674	1997	1997
problems	2	20.77133	1998	1999
more	2	19.96109	2000	2001
use	1	19.38503	2001	2001

NIH				
Word	Length	Weight	Start	End
Phase	8	117.2205	1993	2000
commercial	9	87.57158	1995	
proposed	9	87.57158	1995	
mass	3	83.36952	1978	1980
protein	1	72.15788	1988	1988
networks	4	71.252	1993	1996
patterns	3	66.44826	1977	1979
being	8	66.29254	1971	1978
reasoning	2	65.68178	1984	1985
expert	4	60.49935	1987	1990

USPTO				
Word	Length	Weight	Start	End
human	3	19.03937321	2004	2006
video	3	15.32736425	1998	2000
disclosed	2	14.06694671	1999	2000
neural	3	13.30105906	2004	2006
"correct"	2	12.4336047	1999	2000
unit	2	12.35745838	2002	2003
material	1	12.08487035	2000	2000
feedback	1	12.07730195	2000	2000
rule	1	12.07730195	2000	2000
elevator	4	11.83351857	1991	1994

[#10] Using the Scholarly Database at IU

- Motivation
- > Functionality / Sample Usage
- > Implementation
- Documentation.
- Outlook
- Exercise: Identify Promising SDB Collaborations

SCHOLARLY DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLARLY
DATABASE

SCHOLA



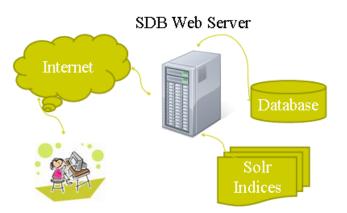
Scholarly Database: Architecture

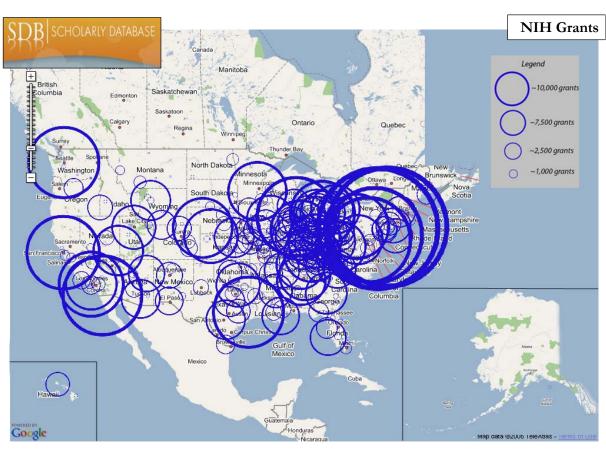
Solr full-text search server

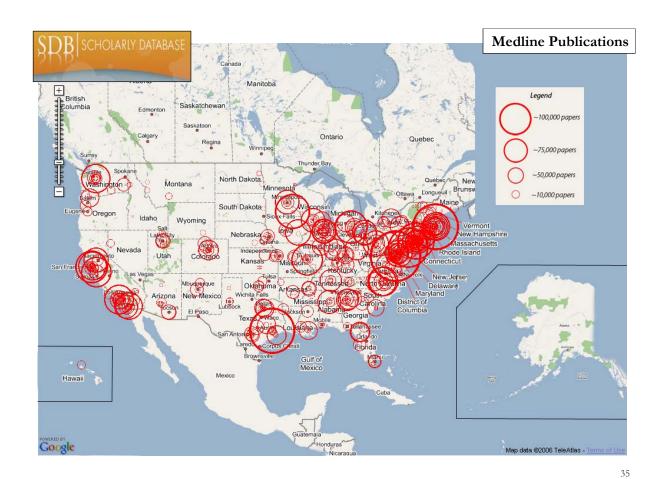
- http://lucene.apache.org/solr/
- Open source
- > Uses the Lucene search library

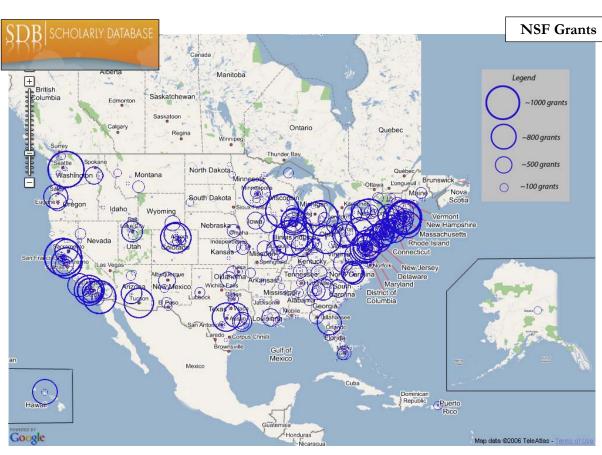
Interface developed in Django

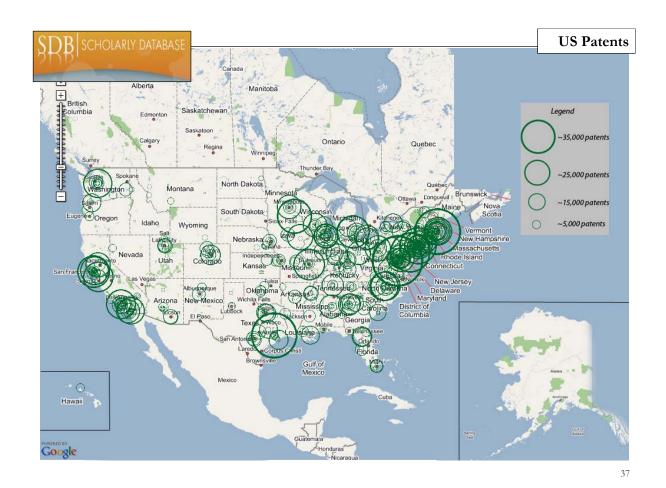
- http://www.djangoproject.com/
- Open source
- Particularly suited for contentfocused web applications











[#10] Using the Scholarly Database at IU Motivation Functionality / Sample Usage Implementation Documentation Outlook Exercise: Identify Promising SDB Collaborations



Scholarly Database: Documentation

Demo

- Wikipedia documentation with table schemas, e.g., https://nwb.slis.indiana.edu/community/?n=ScientometricsDatasets.USPTO
- SDB About page, http://sdb.slis.indiana.edu/about
- Data dictionaries at http://sdb.slis.indiana.edu
- Sample data files at http://sdb.slis.indiana.edu
- Tutorials, e.g., NWB Tool Tutorial, Sci2 Tool Tutorial at http://nwb.slis.indiana.edu/Docs/NWBTool-Manual.pdf
 http://sci.slis.indiana.edu/registration/docs/Sci2 Tutorial.pdf
- Peer reviewed publications, see http://cns.slis.indiana.edu/publications

These types of documentation are needed for scientifically valid studies that are used to inform decision making.

39

[#10] Using the Scholarly Database at IU

- **Motivation**
- > Functionality / Sample Usage
- > Implementation
- Documentation
- Outlook
- **Exercise: Identify Promising SDB Collaborations**



Planned SDB Extensions

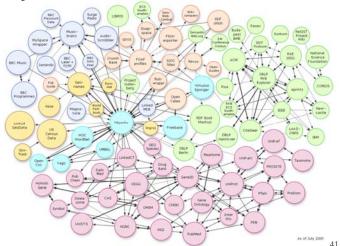
Regular update of SDB Data (add NIH ExPORTER data)

Adding linkage data, e.g., awards-> publications, grants, news.

Adding job market data

Exposing SDB data to the Linked Open Data

Extend SDB-Sci2 Tool synergies.





Adding NIH ExPORTER Data

Source: NIH ExPORTER at

http://projectreporter.nih.gov/exporter/ExPORTER Catalog.aspx?sid=1&index=0

Year coverage: from 2000 till June 2010 Update schedule: monthly for 2010 awards

File formats available: xml/csv Description from Web site

ExPORTER makes downloadable versions of the data accessed through the RePORT Expenditures and Results (RePORTER) interface available to the public. This site is a key component of NIH "open government" initiatives to provide more transparency in NIH activities, improve the quality of the data we collect, and increase its utility.

The NIH ExPORTER now is beta version. Original they only released the data from FY 2005 to FY 2009. On Jun 2010, they increased the historical data from FY 2000 to FY 2004 and refined record formats in response to user feedback. They will post release notes describing these changes until both xml and vsv record formats are finalized on Oct 1, 2010.

Data Fields

Please see the NIH ExPORTER data dictionary.

[#10] Using the Scholarly Database at IU

- Motivation
- > Functionality / Sample Usage
- > Implementation
- **Documentation**
- Outlook
- Exercise: Identify Promising SDB Collaborations

13



Exercise

Please identify promising SDB usages and/or collaborations.

Document it by listing

- Project title
- ➤ User, i.e., who would be most interested in the result?
- Insight need addressed, i.e., what would you/user like to understand?
- Data used, be as specific as possible.
- > Analysis algorithms used.
- Visualization generated. Please make a sketch with legend.



All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.slis.indiana.edu