Scientometrics
DOI 10.1007/s11192-014-1238-2

Open data and open code for big science of science studies

Robert P. Light · David E. Polley · Katy Börner

Received: 18 November 2013

Abstract Historically, science of science (Sci2) studies have been performed by single investigators or small teams. As the size and complexity of data sets and analyses scales up, a "Big Science" approach (Price, Little science, big science, 1963) is required that exploits the expertise and resources of interdisciplinary teams spanning academic, government, and industry boundaries. Big Sci2 studies utilize "big data", i.e., large, complex, diverse, longitudinal, and/or distributed datasets that might be owned by different stakeholders. They apply a systems science approach to uncover hidden patterns, bursts of activity, correlations, and laws. They make available open data and open code in support of replication of results, iterative refinement of approaches and tools, and education. This paper introduces a database-tool infrastructure that was designed to support big Sci2 studies. The open access Scholarly Database (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records. The open source Sci2 tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies. The scalability of the infrastructure is examined. Results show that temporal analyses scale linearly with the number of records and file size, while the geospatial algorithm showed quadratic growth. The number of edges rather than nodes determined performance for network based algorithms.

Keywords Open data · Visualization software · Big data · Scalability · Workflows

R. P. Light (✉) · D. E. Polley · K. Börner
Cyberinfrastructure for Network Science Center, School of Informatics and Computing,
Indiana University, Bloomington, IN, USA
e-mail: lightr@indiana.edu

D. E. Polley
e-mail: dapolley@indiana.edu

K. Börner
e-mail: katy@indiana.edu

Published online: 19 February 2014

Introduction and related work

Many science of science (Sci2) studies use heterogeneous datasets and advanced data mining and visualization algorithms to advance our understanding of the structure and dynamics of science. The quality of results depends on the quality and coverage of the data used. Data cleaning and preprocessing can easily consume 80 % or more of the overall project effort and budget. As the number of data records grows, different types of tools and expertise are required to handle the data. MS Excel can load a maximum of 1,048,576 rows of data by 16,384 columns per sheet. MS Access file sizes cap at 2 GB, including indices, forms, and macros along with the data. Larger datasets need to be stored in a database designed with scalability in mind. As the diversity of datasets increases, the structures of different datasets need to be aligned. As data covers more and more years, dealing with format changes becomes necessary. Many studies require extensive preprocessing and augmentation of the data, such as identification of unique records or record values, geocoding of records in preparation for geospatial analysis, or the extraction of networks for network studies. For many researchers, the effort to compile ready-to-analyze-and-visualize data is extremely time consuming and challenging and sometimes simply insurmountable.

Many datasets relevant for Sci2 studies, e.g., papers, patents, grants, and clinical trials, are freely available by different providers. However, they are stored in separate silos with diverse interfaces of varying usability that deliver data in many different formats. Research projects seeking to use one or many of these data sources face major data access, integration, and unification challenges. Indiana University's Scholarly Database (SDB), originally launched in 2005, makes over 26 million scholarly records freely available via a unified interface and in data formats that are easy to use and well documented. In the last 4 years, SDB has answered thousands of queries and delivered millions of records to users around the globe. The 2012 update to the SDB improves the quality of data offered and integrates new humanities and clinical trial datasets.

Equipped with high quality, high coverage data in standard data formats, tools that scale in terms of the number of records that can be read and processed are needed to truly make sense of big data (Robertson et al. 2009). While most tools work well for micro and meso level studies (up to 100,000 records), few scale to macro level big-data studies with millions or even billions of records. Another type of scalability relates to the ease of usage and ease of interpretation of big data visualizations. How to best communicate temporal trends or burst of activity over a 100 year time span? How to depict the geospatial location of millions of records in a scalable fashion? Can the topical evolution of massive document datasets be communicated to a general audience? Most visualizations of million node networks resemble illegible spaghetti balls—do advanced network analysis algorithms scale and help to derive insights?

Frequently, different types of analysis have to be applied to truly understand a natural, social, or technological system. Examples are temporal studies that answer WHEN questions, geospatial studies that answer WHERE questions and draw heavily on research in cartography, topical studies that use linguistic analysis to answer WHAT questions, and network studies that employ algorithms and techniques developed in social sciences, physics, information science and other domains to answer WITH WHOM questions. However, most existing systems support only one general type of analysis and visualization and many require programming skills. For example, four of the top 20 data visualization tools listed by.*net* in September of 2012 support charts and graphs while six support geospatial maps exclusively (Suda 2012). Only the D3 (data-driven documents) and

Raphaël JavaScript libraries, the Google Chart API, and R support a larger array of charts, graphs, and maps yet all three require programming or scripting skills that most users do not possess. Excel might be the only tool on the list that can be used by a large number of non-programmers. A listing of tools commonly used in Sci2 studies can be found at http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/8.2+Network+Analysis+and+Other+Tools but most support a very limited range of workflows (Cobo et al. 2011).

This paper presents a database-tool infrastructure that applies a divide-and-conquer approach to support big Sci2 studies. It combines an online database supporting bulk download of data in easy to process formats with a plug-and-play tool to read, clean, interlink, mine, and visualize data using easy to manipulate graphical user interfaces. An earlier version of this paper was published in the proceedings of the International Society of Scientometrics and Infometrics Conference in Vienna (Light et al. 2013).

The remaining paper is organized as follows: the next two sections present the database and tool functionalities. We then present a sample workflow, complete from initial question to data acquisition to visualization and interpretation. Subsequently, we test and discuss the scalability of data readers, preprocessing, analysis and visualization algorithms. We conclude the paper with a discussion of the presented work and an outlook to future work.

The Scholarly Database (SDB)

SDB was created in 2005 to provide researchers and practitioners easy access to various datasets offered by different publishers and agencies (LaRowe et al. 2009). The SDB is implemented using PostgreSQL 8.4, a free and open source relational database management system. Since the introduction of version 8.1, PostgreSQL developers have been focused on improving the scalable performance of the system and this software is now employed by many companies to provide large-scale data solutions, including Yahoo!, Sony Online and Skype. Today, the SDB provides easy access to paper, patent, grant, and clinical trials records authored by 13.8 million people in 208 countries (some, such as Yugoslavia, no longer in existence), interlinked by 58 million patent citation links, and over 2.5 million links connecting grant awards to publications and patents. As of November 2012, the SDB features over 26 million records from MEDLINE (19,039,860 records spanning from 1865 to 2010), United States Patent and Trademark Office (USPTO) patents (4,178,196, 1976–2010), National Institutes of Health (NIH) awards (2,490,837, 1972–2012), National Science Foundation (NSF) awards (453,687, 1952–2010), National Endowment for the Humanities (NEH) awards (47,197, 1970–2012), and clinical trials (119,144, 1900–2012).

Unique features of SDB comprise:

- *Open access* the SDB is composed entirely of open data so there are no copyright or proprietary issues for the researcher to contend with in its use. Data is provided to researchers free of charge.
- *Ease of use* simple user interfaces provide a one-stop data access experience making it possible for researchers to focus on answering their questions, rather than spending much time on parsing, searching, and formatting data.
- *Federated search* by aggregating the data into a single environment, SDB offers a federated search environment powered by a Solr core. Users can search one, some, or

all of the available datasets over some or all years using the same set of terms and get a combined set of results that are ranked by relevance.

- *Bulk download* most databases do not support downloads and those that do only permit access to a limited number of records. SDB supports bulk download of data records; data linkages—co-author, patent citations, grant-paper, grant-patent; burst analysis files. Users are granted a base number of downloads by default to prevent abuse of the system, but this number can be extended by request without charge.
- *Unified file formats* SDB source data comes in different file formats. NIH funding data is stored in flat files; clinical trials are offered in XML, while patents come in a variety of formats, depending on the year. Old patents come in a fixed width data format while newer patents are provided in XML. Much time and effort was spent to normalize this data into easy-to-use file formats, e.g., comma-delimited tables for use in spreadsheet programs and common graph formats for network analysis and visualization.
- *Well-documented* SDB publishes data dictionaries for every dataset offered. Information on data provenance, table structure, data types, and individual field comments are available. In addition, the SDB offers a set of small sample files, giving researchers an easily usable test-bed for working out their algorithms before committing to analysis of a larger set.

The SDB Wiki (http://sdb.wiki.cns.iu.edu) provides more information including a user guide, information on each dataset, and release notes.

The Sci2 tool

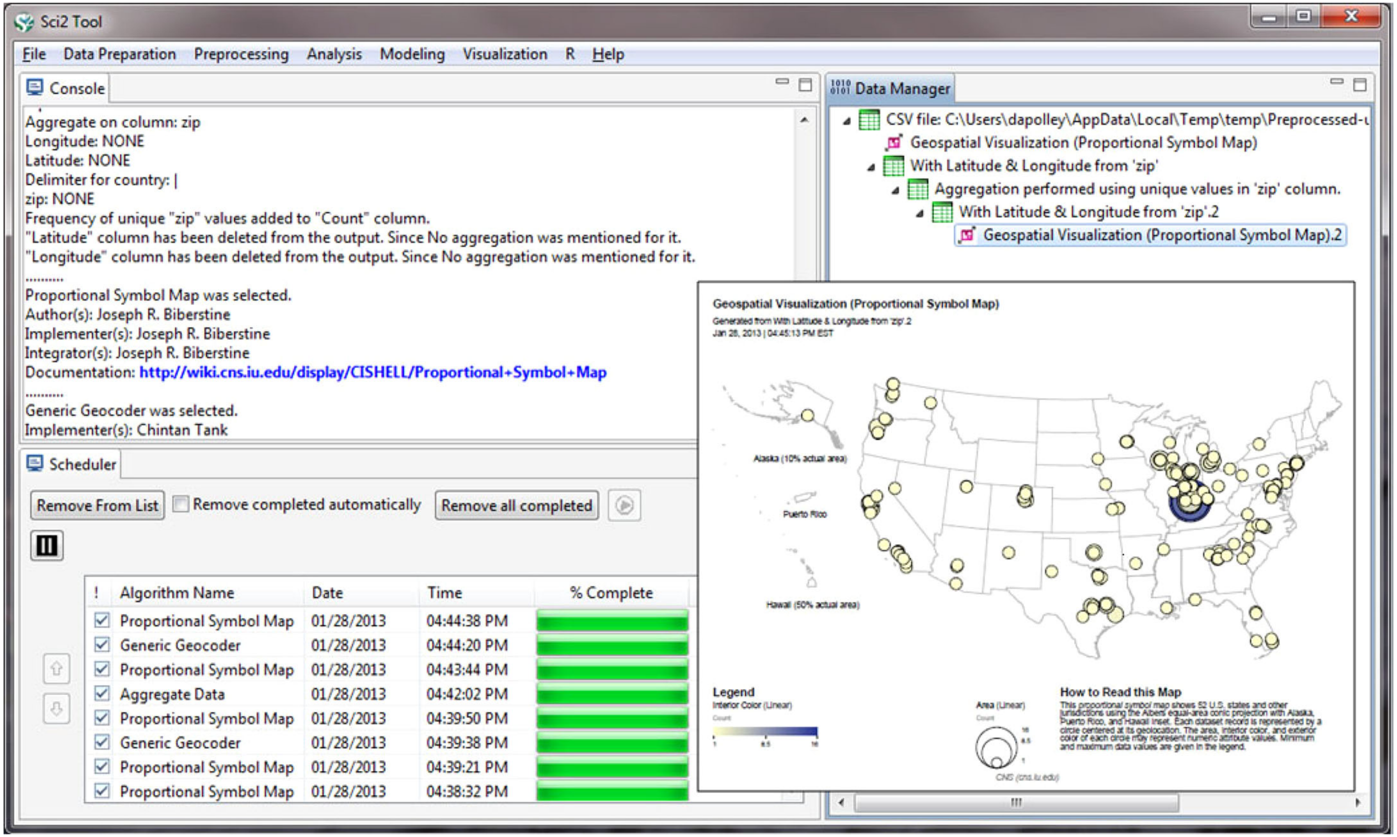
The Sci2 tool is a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels, see screenshot in Fig. 1, general workflow in Fig. 2 and specific workflows discussed in the "Scalability tests" section.

The tool's OSGi/CIShell core architecture makes it possible for domain scientists to contribute new algorithms written in a variety of programming languages using a plug-and-play macroscope approach (Börner 2011).

As of November 2012, the Sci2 tool has 171 algorithms, 112 of which are visible to the user (see Table 1) written in Java, C, C++, and Fortran. In addition, a number of tools (Gnuplot, GUESS, and Cytoscape) were implemented as plugins and bridges to R and to Gephi were created, allowing the seamless use of different tools. The Sci2 user interface and sample map is shown in Fig. 1.

Unique features of Sci2 comprise:

- *Open source* anybody can examine the source code and advance it.
- *Extensive use of well-defined reference systems* to improve readability and to support interpretation, Sci2 uses a number of carefully designed reference systems, see Fig. 3. Each comes with a title, legend, and a brief "How to read this visualization" section that provides further details, e.g., on used geospatial projections.
- *Interactivity* while visualizations of small datasets can be explored interactively, visualizations of big data are rendered into Postscript files that can be converted to PDF files and examined using pan and zoom as well as filtered, e.g., by searching for specific text in the display.



Aggregate on column: zip
Longitude: NONE
Latitude: NONE
Delimiter for country: |
zip: NONE
Frequency of unique "zip" values added to "Count" column.
"Latitude" column has been deleted from the output. Since No aggregation was mentioned for it.
"Longitude" column has been deleted from the output. Since No aggregation was mentioned for it.
....
Proportional Symbol Map was selected.
Author(s): Joseph R. Biberstine
Implementer(s): Joseph R. Biberstine
Integrator(s): Joseph R. Biberstine
Documentation: http://wiki.cns.iu.edu/display/CISHELL/Proportional+Symbol+Map
....
Generic Geocoder was selected.
Implementer(s): Chintan Tank

CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed-u
- Geospatial Visualization (Proportional Symbol Map)
- With Latitude & Longitude from 'zip'
  - Aggregation performed using unique values in 'zip' column.
    - With Latitude & Longitude from 'zip'.2
      - Geospatial Visualization (Proportional Symbol Map).2

!	Algorithm Name	Date	Time	% Complete
☑	Proportional Symbol Map	01/28/2013	04:44:38 PM	🟩
☑	Generic Geocoder	01/28/2013	04:44:20 PM	🟩
☑	Proportional Symbol Map	01/28/2013	04:43:44 PM	🟩
☑	Aggregate Data	01/28/2013	04:42:02 PM	🟩
☑	Proportional Symbol Map	01/28/2013	04:39:50 PM	🟩
☑	Generic Geocoder	01/28/2013	04:39:38 PM	🟩
☑	Proportional Symbol Map	01/28/2013	04:39:21 PM	🟩
☑	Proportional Symbol Map	01/28/2013	04:38:32 PM	🟩

Geospatial Visualization (Proportional Symbol Map)
Generated from With Latitude & Longitude from 'zip'.2
Jan 28, 2013 | 04:45:13 PM EST

How to Read this Map

This proportional symbol map shows 52 U.S. states and other jurisdictions using the Albers equal-area conic projection with Alaska, Puerto Rico, and Hawaii inset. Each dataset record is represented by a circle centered at its geolocation. The area, interior color, and exterior color of each circle may represent numeric attribute values. Minimum and maximum data values are given in the legend.

Fig. 1 Sci2 tool user interface with *Proportional Symbol Map* visualization

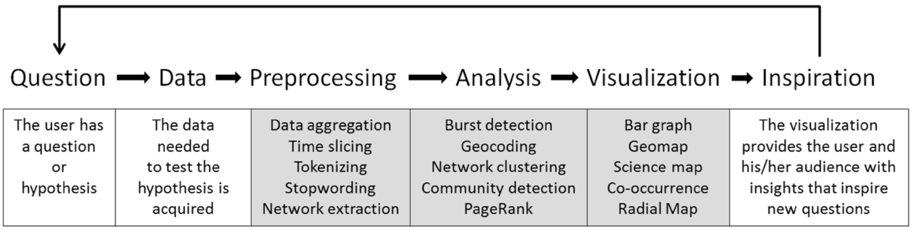


Fig. 2 General Sci2-based visualization creation workflow (tool-specific tasks in *gray*)

Table 1 Sci2 algorithm summary tables

Categories	Algorithms	Examples
Acquisition	5	Google citation user ID search algorithm
Data preparation	13	Extract co-occurrence network
Preprocessing	22	Slice table by time, extract ZIP code
Analysis	47	*K*-nearest neighbor, *Burst Detection*
Modeling	4	Watts–Strogatz small world, TARL
R	4	Create an R instance, send a table to R
Visualization	17	Choropleth map, bipartite network graph
Total	112	

- *Workflows* all user actions are recorded in a log file to ensure proper documentation and easy replicability of workflows that might comprise 15–20 analysis and visualization algorithms with a range of parameter settings.
- *Online documentation* all Sci2 plugins as well as major workflows are documented in the Sci2 Wiki (http://sci2.wiki.cns.iu.edu) together with release notes.

Sample workflow

This sample workflow aims to answer the question: "What were the emergent topics in hurricane research during the 2000s?" Questions of emergent topics lend themselves well to burst analysis, which highlights words that appear or increase in frequency in a dataset over a portion of the time examined. In order to perform burst analysis, records need to include text as well as a time-stamp.

Data query

On the SDB website, we search for the terms "hurricane" or "typhoon" in the All Text field, accessing the NSF dataset from 2000 to 2010. The dataset includes 509 grants, the strongest matches of which are shown in Table 2.

Data download

The SDB offers a variety of tables for download depending on the dataset(s) queried. The contents of all tables are described via the data dictionary (downloadable at http://wiki.cns.

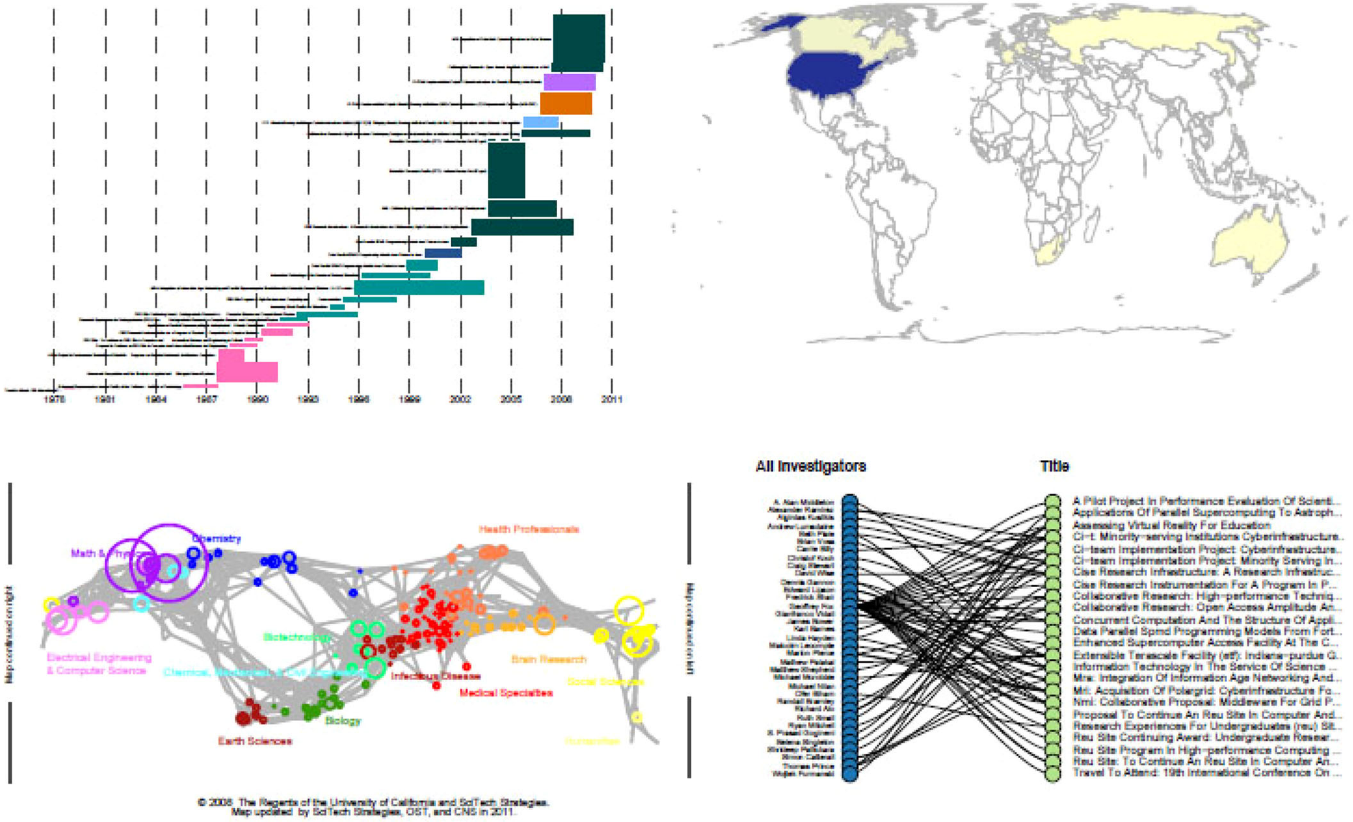
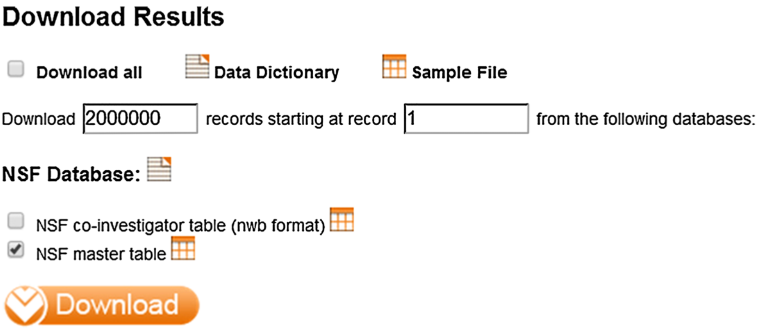


Fig. 3 Exemplary reference systems supported by Sci2 including *Temporal Bar Graph* (*top*, *left*), Choropleth *map* (*top*, *right*), UCSD science *map* (*bottom*, *left*), bimodal network visualization (*bottom*, *right*). Full versions available at http://wiki.cns.iu.edu/display/SCI2TUTORIAL/1+Introduction

Table 2 Highest rated results from a search of the terms "hurricane" or "typhoon" in the NSF database from 2000 to 2010, as run on 12 November 2013

Source	Authors/ creators	Years	Titles	Score (out of 1.84)
NSF	Eisner	2000	Extratropical linkages to tropical cyclone activity	1.84
MEDLINE	Chen et al.	2003	Strategies of disaster response in the health care system for tropical cyclones: experience following Typhoon Nari in Taipei City	1.67
MEDLINE	Bengtsson	2001	Weather. Hurricane threats	1.46
NSF	Ritchie and Tyo	2007	Enhancing forecasts of tropical cyclone extratropical transition by statistical pattern recognition	1.34
MEDLINE	Schiermeier	2005	Hurricane link to climate change is hazy	1.26
MEDLINE	Bohannon and Enserink	2005	Hurricane Katrina. Questioning the 'Dutch solution'	1.26
MEDLINE	Odom-Forren	2005	Hurricane Katrina	1.26
MEDLINE	Witze	2006	Tempers flare at hurricane meeting	1.26
MEDLINE	Kerr	2006	Climatology. A tempestuous birth for hurricane climatology	1.26
MEDLINE	Baum and Fendell	2006	Operational hurricane intensity forecasting	1.26



Download Results

□ Download all

Download 2000000 records starting at record 1 from the following databases:

NSF Database:

□ NSF co-investigator table (nwb format)
☑ NSF master table

Fig. 4 SDB download screen, showing datasets, data dictionary and sample files available for download

iu.edu/display/SDBDOC/NSF+Awards) that is available for all datasets, as seen in Fig. 4. The required data for this analysis is provided in the master table. The file contains columns for titles, abstracts and years of publication and can be loaded into the Sci2 tool for preparation and visualization.

Data preparation

Once the data has been loaded into the Sci2 tool, the first step is to pre-process the text. The *Lowercase*, *Tokenize*, *Stem and Stopword* routine is designed to normalize text in preparation for analysis. Initially, the default stopword list included with Sci2 is used. The

Table 3 Top bursts in hurricane research (2000–2010) by weight (raw data)

Time	Terms	Weights	Start	End	Ranks	Terms	Weights	Start	End
1	Sger	25.15	2005	2005	11	Law	7.91	2009	2010
2	Katrina	9.92	2005	2005	12	Applic	7.80	2007	2010
3	Sampl	9.83	2005	2005	13	Intens	7.60	2003	2004
4	Orlean	9.16	2005	2005	14	Mississippi	7.57	2005	2006
5	2009	8.86	2009	2010	15	Expect	7.49	2009	2010
6	Wind	8.67	2002	2004	16	Abstract	7.40	2005	2005
7	Aftermath	8.45	2005	2006	17	Complet	7.21	2002	2004
8	111-5	8.24	2009	2010	18	Gulf	7.00	2005	2006
9	Health	8.05	2004	2005	19	Act	6.93	2009	2010
10	Reinvest	8.02	2009	2010	20	Scienc	6.72	2007	2007

normalized dataset is then analysed with the *Burst Detection* algorithm using default parameters. This generates a list of bursting topics, with weights, start, and end dates. The top-20 most highly weighted topics are listed in Table 3.

Some terms, like “katrina” (referring to Hurricane Katrina the deadliest and most destructive Atlantic tropical cyclone in 2005) that bursts in 2005 are easy to understand, while others, like “sger” and “111-5” are difficult to interpret. More detailed examination reveals that “111-5” is the Public Law number of the American Recovery and Reinvestment Act that provided funding for a great many NSF proposals in the late 2000s. Likewise, “sger” refers to the “Small Grants for Exploratory Research” program run by the NSF. Both burst during 2005 as much funding is devoted to hurricane research in that year. If desired, the two terms can be excluded from further analysis by modifying the stopword list. A copy of the original stopword list is made and terms that are undesirable are added, including “sger”, “111-5” “2009” and several other years that appeared later in the data, as well as terms such as “grant” and “award” that are clearly more mechanical to the NSF process than relevant to the science conducted. The process is repeated, starting from textual normalization and the modified results are shown in Table 4.

Table 4 Top bursts in hurricane research (2000–2010) by weight (refined)

Ranks	Terms	Weights	Start	End	Ranks	Terms	Weights	Start	End
1	Katrina	9.92	2005	2005	11	Complet	7.21	2002	2004
2	Sampl	9.83	2005	2005	12	Gulf	7.00	2005	2006
3	Orlean	9.16	2005	2005	13	Exploratori	6.56	2005	2005
4	Wind	8.67	2002	2004	14	Public	6.46	2009	2010
5	Aftermath	8.45	2005	2006	15	Louisiana	6.33	2005	2005
6	Health	8.05	2004	2005	16	Bring	6.20	2007	2008
7	Applic	7.80	2007	2010	17	Advanc	6.11	2009	2010
8	Intens	7.60	2003	2004	18	Graduat	6.07	2009	2010
9	Mississippi	7.57	2005	2006	19	Approach	5.97	2008	2010
10	Expect	7.49	2009	2010	20	Open	5.75	2008	2010

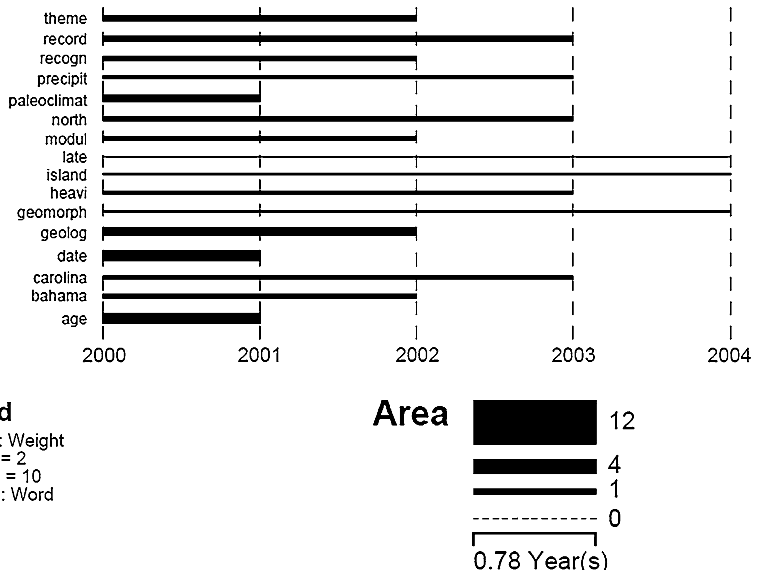


Fig. 5 Excerpt of *Temporal Bar Graph* of hurricane research bursts (2000–2010)

Visualization

While further refinement is possible via tweaking the parameters of the algorithm or making more additions to the stopword file, this dataset reflects the expected impact of Hurricane Katrina on research in 2005 and is the one that will be visualized here. The visualization selected is the *Temporal Bar Graph*, creating bars for each burst over time. The full visualization, a six page PDF document covering 290 bursts, is too large to reproduce here, but a portion is shown in Fig. 5. The length of bars indicates the length of the burst, while area represents the weight of the burst.

This sample has used only a few hundred records from a 1.2 MB file, far from what would be considered truly big data, but the principles remain consistent regardless of data size. Subsequently, we present the results of extensive tests that were run to determine the scalability of key Sci2 tool algorithms and visualizations.

Scalability tests

To demonstrate the scalability of the database and tool, tests were performed using synthetic datasets with pre-defined properties generated in Python and datasets retrieved from the SDB. All four types of analysis supported by Sci2 were tested: temporal analysis, geospatial analysis, topical analysis, and network analysis. Initially, we identified workflows indicative of these four main types of analysis. From there, we broke down each workflow into the specific steps (algorithms) involved in the workflow, starting with loading the data and ending in visualization. For each algorithm, e.g., data reader, analysis, visualization, we measured (in seconds) the length of time it took for an algorithm to finish processing. We considered the start of the algorithm to be the point at which the user inputs his or her parameters (where applicable) and then executes the algorithm. We considered all algorithms to be finished when the associated data files appeared in the Data Manager and were displayed as complete in the Scheduler. For each test, we calculated the average for 10 trials. Between trials, we closed down Sci2 in order to minimize any adverse effects of residual memory. Tests were performed on a common system: an Intel(R) Core(TM) Duo CPU E8400 3.00 GHz processor and 4.0 GB of memory running a 64bit version of Windows 7 and a 32bit version of Java 7. Memory allotted to Sci2 was extended to 1,500 MB.

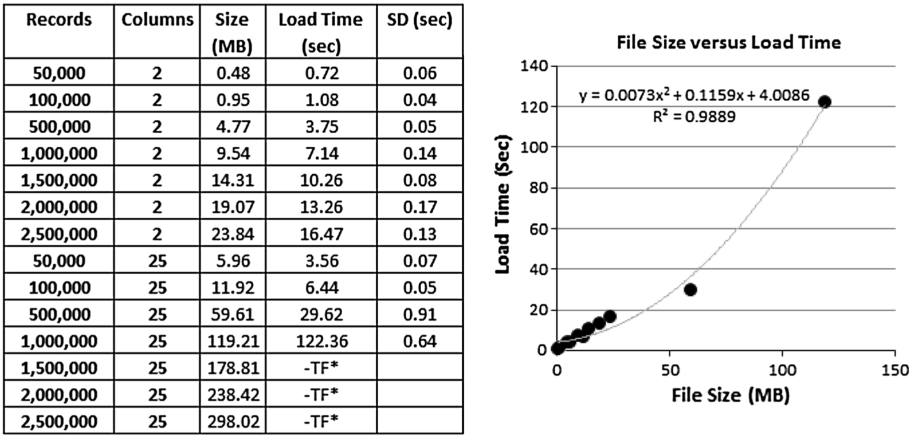
File loading

Synthetic data was used to measure how file loading times vary in terms of number of records and length of individual record in bytes. Two series of datasets were generated, one with only 2 rows, a small integer, and a short string and one with 25 rows, a small integer and 24 short strings, each with increasing numbers of rows. Average loading times over 10 trials are given in Fig. 6. The three largest datasets did not load but returned a Java heap space error (-TF*). At first glance, there seems to exist a direct relationship between file size and loading time ($R^2 = 0.9384$), a closer look at the plot of size versus time reveals that a quadratic regression line has a noticeably better fit ($R^2 = 0.9889$). This is likely a result of the tool having to devote resources to file management that would otherwise be available for completing functions more efficiently.

Next, SDB data prepared for usage in Sci2 workflows was read comprising.

- NIH data at 3.4 GB, NSF data at 489 MB, NIH data at 139 MB, and NEH data at 12.1 MB data prepared for temporal analysis.
- Data from NIH, NSF, MEDLINE, USPTO, and clinical trials at 11.5 MB and MEDLINE data at 1 GB to be used in geospatial analysis.
- MEDLINE data at 514 kB for topical analysis.
- NSF data at 11.9 MB and USPTO data at 1.04 GB network analysis.

Average load times measured across 10 trials are shown in Table 5. The three largest datasets, would not load but returned a Java heap space error (-TF*, Table 5).



Records	Columns	Size (MB)	Load Time (sec)	SD (sec)
50,000	2	0.48	0.72	0.06
100,000	2	0.95	1.08	0.04
500,000	2	4.77	3.75	0.05
1,000,000	2	9.54	7.14	0.14
1,500,000	2	14.31	10.26	0.08
2,000,000	2	19.07	13.26	0.17
2,500,000	2	23.84	16.47	0.13
50,000	25	5.96	3.56	0.07
100,000	25	11.92	6.44	0.05
500,000	25	59.61	29.62	0.91
1,000,000	25	119.21	122.36	0.64
1,500,000	25	178.81	-TF*	
2,000,000	25	238.42	-TF*	
2,500,000	25	298.02	-TF*	

Fig. 6 Comparison of load times, measured in seconds, across standardized datasets, tabulated (*left*) and plotted with quadratic regression *line* (*right*)

Table 5 Comparison of load times, measured in seconds, across nine different datasets

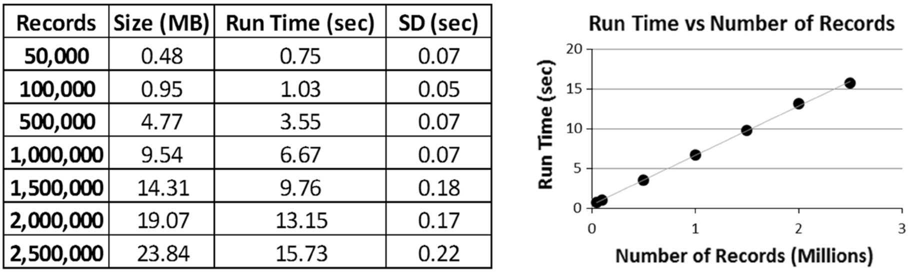
Datasets	Sizes	Number of records	Mean	Standard deviation	Minimum	Maximum
NIH (year, title, abstract)	3.4 GB	2,490,837	-TF*			
USPTO (patent, citations)	1.04 GB	57,902,504	-TF*			
MEDLINE (geospatial)	1.0 GB	9,646,117	-TF*			
NSF (year, title, abstract)	489 MB	453,740	64.54	0.991	63.2	65.9
NIH (title, year)	139 MB	2,490,837	83.86	1.32	82.3	85.6
NEH (year, title, abstract)	12.1 MB	47,197	2.05	0.070	1.9	2.1
NSF (co-author network)	11.9 MB	341,110	4.52	0.063	4.4	4.6
Combined geo-spatial	11.5 MB	11,549	1.91	0.056	1.8	2.0
MEDLINE journals	0.5 MB	20,775	0.44	0.096	0.3	0.6

Temporal studies ("When")

To test the scalability of temporal analysis within Sci2 we selected the *Burst Detection* algorithm as described by Kleinberg (2003). To test this in a standardized fashion, we generated a randomized set of years from 1980 to 2000, assigning each year a distribution of short strings to test the accuracy of the algorithm. We then calculated the average time, minimum time, and the maximum time it took the *Burst Detection* algorithm to complete across 10 trials. In all cases, the algorithm was able to detect a pre-programmed burst of a word over a short time frame.

A look at the table and graph in Fig. 7 shows linear growth with number of records that holds equally true with file size. It is possible that with larger files, this may begin to show the same quadratic tendency as the file loading, but 2.5 million records was the largest file loaded. The data does illustrate that, barring resource exhaustion issues, Sci2 runs this algorithm in a linear timescale.

We then conducted a burst analysis of the title fields for NIH, NSF, and NEH grant data. The NSF and NEH datasets contain three columns: title, abstract, and year. The NIH data contains only two columns: title and year. The NIH grant data set is the largest at 139 MB and 2,490,837 records, followed by the NSF grant data at 489 MB and 453,740 records, and finally the NEH grant data at 12.1 MB with 47,197 records (Table 6). In order to obtain accurate results with the *Burst Detection* algorithm we had to normalize the title text with the *Lowercase*, *Tokenize*, *Stem*, *and Stopword Text* algorithm prior to running the *Burst*



Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	0.48	0.75	0.07
100,000	0.95	1.03	0.05
500,000	4.77	3.55	0.07
1,000,000	9.54	6.67	0.07
1,500,000	14.31	9.76	0.18
2,000,000	19.07	13.15	0.17
2,500,000	23.84	15.73	0.22

Fig. 7 Comparison of *Burst Detection* run times, measured in seconds, across standardized datasets, tabulated (*left*) and plotted (*right*)

Table 6 Temporal analysis algorithm run time in seconds

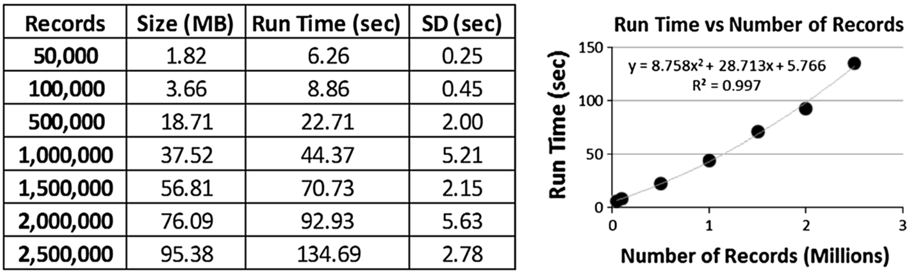
Datasets	Sizes (MB)	Rows	Mean	SD	Min	Max
Burst Detection						
NSF	489	453,740	13.64	0.648	12.9	14.8
NIH	139	2,490,837	-NT*			
NEH	12.1	47,197	1.57	0.094	1.4	1.7

Detection algorithm, a step not necessary with the synthetic data since it was optimized for burst analysis. Due to the number of records in the NIH dataset, the *Lowercase*, *Tokenize*, *Stem*, *and Stopword Text* algorithm failed to terminate and as a result the *Burst Detection* algorithm was not tested with this dataset (-NT*).

Geospatial studies (“Where”)

In order to test Sci2 performance for geomapping, randomized datasets with lists of US cities and associated longitude and latitude, were generated. There was only one distinct step (algorithm) involved in this geospatial workflow: visualizing the geolocated data with the *Proportional Symbol Map* (Biberstine 2012), see US geomap in Fig. 2. We projected this on a map of the United States, as this data set only included locations within the US average run times are shown in Fig. 8. Like with file loading, the *Proportional Symbol Map* data is better fit by a quadratic model (R^2 of 0.997 as opposed to 0.9834 for a linear fit).

Next, 11,848 SDB records related to gene therapy funding (NIH, NSF), publications (MEDLINE), patents (USPTO), and clinical trials were loaded and the *Proportional Symbol Map* was used to display the geocoded data (Table 7). Exactly 299 records had no or incomplete geolocation data and were removed resulting in 11,549 rows at 11.5 MB. The run time, at 4.37 s is lower than predicted by the model (6.11 s), implying that the quadratic model may not perfectly describe the run time, particularly with smaller sets.

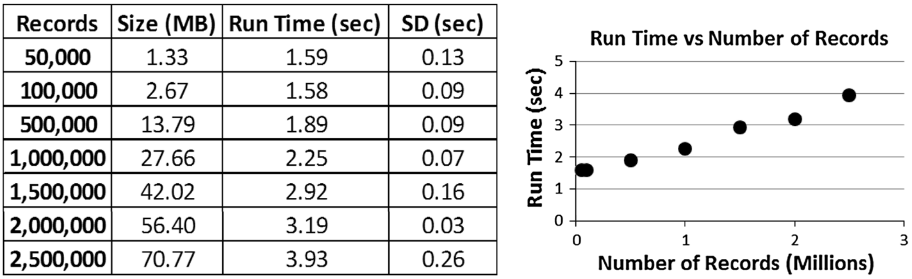


Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	1.82	6.26	0.25
100,000	3.66	8.86	0.45
500,000	18.71	22.71	2.00
1,000,000	37.52	44.37	5.21
1,500,000	56.81	70.73	2.15
2,000,000	76.09	92.93	5.63
2,500,000	95.38	134.69	2.78

Fig. 8 Comparison of *Proportional Symbol Map* run times, measured in seconds, across standardized datasets

Table 7 Geospatial analysis algorithm run time in seconds

Dataset	Size (MB)	Rows	Mean	SD	Min	Max
Algorithm 1: *Proportional Symbol Map*						
Pre-located	11.5	11,549	4.37	0.125	4.2	4.6



Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	1.33	1.59	0.13
100,000	2.67	1.58	0.09
500,000	13.79	1.89	0.09
1,000,000	27.66	2.25	0.07
1,500,000	42.02	2.92	0.16
2,000,000	56.40	3.19	0.03
2,500,000	70.77	3.93	0.26

Fig. 9 Comparison of UCSD *map* of science generation run times, measured in seconds, across standardized datasets

Topical studies ("What")

The Sci2 tool supports the generation of science map overlays. Specifically, it uses the UCSD map of science and classification system (Börner et al. 2012), a visual representation of 554 sub-disciplines within 13 disciplines of science and their relationships to one another, see lower left map in Fig. 2. This basemap is then used to show the result of mapping a data set's journals to the underlying subdiscipline(s) those journals represent (Biberstine 2011). Mapped subdisciplines are shown with node sizes relative to the number of articles matching journals and color is based on the discipline as defined in the basemap. To create a standardized dataset, random lists of valid journal names were generated. The number of records and run time results are tabulated and plotted in Fig. 9. Linear and quadratic models fit about equally well, but both show that the intercept is about 1.5 s, more than half of the run time for all but the largest sets. This stands to reason as the lookup tables must be loaded and accessed regardless of the size of the dataset being used.

Next, MEDLINE data was obtained from SDB including all 20,773 journals indexed in MEDLINE and the number of articles published in those journals. Average *Map of Science* via *Journals* run times are given in Table 8.

Network studies ("With Whom")

Sci2 supports the extraction of diverse network types. The *Extract Directed Network* algorithm (Alencar 2010) accepts tabular data and constructs a directed network from entities in the specified source column to entities in the specified target column. Run times across 10 trials for networks with different numbers of nodes and edges are shown in Fig. 10. As to be expected, there is a direct linear relationship between the number of edges and the run time (Fig. 10).

Next we retrieved from the SDB all 6,206 USPTO patents that cite patents with numbers 591 and 592 in the patent number field. We ran the *Extract Directed Network*

Table 8 Topical visualization algorithm run time in seconds

Dataset	Size (kB)	Rows	Mean	SD	Min	Max
Algorithm 1: map of science via journals						
MEDLINE journals	514	20,773	7.84	0.096	7.7	8.0

Records	% Conn	Edges	Size (MB)	Run (sec)	SD (sec)	Records	% Conn	Edges	Size (MB)	Run (sec)	SD (sec)
500	**2**	**5,000**	0.017	1.13	0.05	**250**	**50**	**31,250**	0.124	1.86	0.05
500	**5**	**12,500**	0.045	1.44	0.07	**500**	**50**	**125,000**	0.546	5.89	0.1
500	**10**	**25,000**	0.093	1.92	0.04	**1,000**	**50**	**500,000**	2.28	20.74	0.12
500	**25**	**62,500**	0.247	3.46	0.08	**1,500**	**50**	**1,125,000**	5.21	45.28	0.44
500	**50**	**125,000**	0.546	5.89	0.1	**2,000**	**50**	**2,000,000**	9.33	79.41	0.62

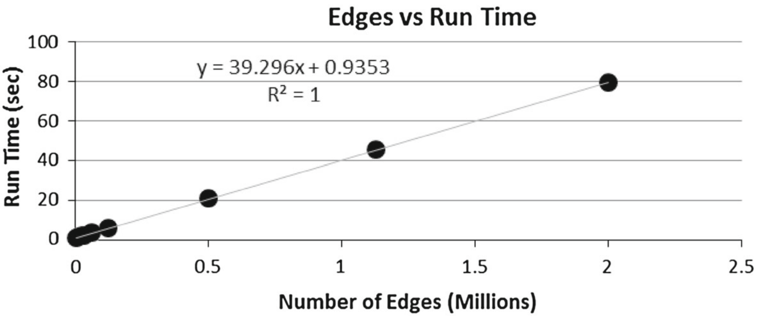


Fig. 10 Average directed network extraction run times, measured in seconds versus the number of edges in the dataset, across standardized datasets, tabulated with varying connectivity (*left*) and number of nodes (*right*, *top*) and plotted (*below*)

Table 9 Network analysis algorithm run time in seconds

Dataset	Size (MB)	Nodes	Edges	Mean	SD	Min	Max
Algorithm 1: extract co-occurrence network							
U.S. patent references	0.147	12,672	7,940	7.88	0.103	7.7	8.1

algorithm, creating a network pointing from the patent numbers to the numbers those patents reference in the dataset and results are given in Table 9. While the scalability of Sci2 third-party visualization tools such as GUESS, Cytoscape, and Gephi do not pertain to Sci2 in a direct way, we were interested to understand their scalability. Neither Cytoscape nor GUESS were capable of rendering the network in a Fruchterman–Reingold layout, while Gephi loaded the network in 2.1 s and rendered it in about 40 s (the actual process in Gephi is non-terminating, but this was the time to a reasonably defined network). Gephi is able to achieve higher performance due to its ability to leverage GPUs in computing intensive tasks.

Discussion and future work

This paper introduced, exemplified, and examined the scalability of a database-tool infrastructure for big Sci2 studies. SDB relational database functionality was exploited to store, retrieve, and preprocess datasets. Subsequently, the data were processed using the Sci2 tool. The scalability of this approach was tested for exemplary analysis workflows using synthetic and SDB data. Techniques used were similar to those employed in testing the performance of web-native information visualizations (Johnson and Jankun-Kelly

2008). Most run-times scale linearly or exponentially with file size. The number of records impacts run-time more than file size. Files larger than 1.5 million records (synthetic data) and 500 MB (SDB) cannot be loaded and hence cannot be analyzed. Run times for rather large datasets are commonly less than 10 s. Only large datasets combined with complex analysis require more than one minute to execute.

A forthcoming paper will compare the runtime of Sci2 with other tools that have similar functionality, e.g., TEXTrend or VOSViewer for topical analysis and visualization; CiteSpace, Leydesdorff's Software, DynaNets, SISOB, Cytoscape, and Gephi for network analysis and visualization, see below and (Cobo et al. 2011) for links and references.

Recent work has added web services to the Sci2 tool and selected workflows can now be run online. Other efforts aim to expand the adoption of OSGi/CIShell in support of algorithm and tool plugin implementation and sharing across scientific boundaries. Tools that are OSGi/CIShell compatible comprise TEXTrend (http://textrend.org) led by George Kampis at Eötvös Loránd University, Budapest, Hungary supports natural language processing, classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component and DynaNets (http://www.dynanets.org) coordinated by Peter Sloot at the University of Amsterdam for the study of evolving networks, or SISOB (http://sisob.lcc.uma.es) an observatory for science in society based in social models.

Much of the development time for the SDB for the last year has been focused on adding data to the system and refactoring code to make it easier to manage and update. Going forward, we plan to implement an API to further ease access and usage of the SDB and we are exploring an RDF conversion to add SDB to the Web of Linked Open Data (Heath and Bizer 2011). In addition, we are considering a visual interface to SDB that uses Sci2 Web services to empower users to interactively explore, analyze, and visualize search results.

Documentation and teaching of tool functionality and workflows are important for research and practice. SDB and Sci2 are used in the Information Visualization MOOC (http://ivmooc.cns.iu.edu) which debuted in Spring 2013 to over 1,700 users, making existing and new workflows available via video tutorials to a much broader audience.

Acknowledgments The Scholarly Database is funded by the National Science Foundation under Grant No. IIS-0238261. SDB and Sci2 are supported by the National Science Foundation under Grants No. IIS-0513650, SBE-0738111, a James S. McDonnell Foundation Grant, and the Cyberinfrastructure for Network Science Center at the School of Informatics and Computing, Indiana University. 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.

References

Alencar, A. (2010). CIShell: Extract directed network. Retrieved January 24, 2013, from http://wiki.cns.iu.edu/display/CISHELL/Extract+Directed+Network.

Biberstine, J. R. (2011). CIShell: Proportional symbol map. Retrieved January 24, 2013, from http://wiki.cns.iu.edu/display/CISHELL/Map+of+Science+via+Journals.

Biberstine, J. R. (2012). CIShell: Proportional symbol map. Retrieved January 24, 2013, from http://wiki.cns.iu.edu/display/CISHELL/Proportional+Symbol+Map.

Börner, K. (2011). Plug-and-play macroscopes. *Communications of the ACM, 54*(3), 60–69.

Börner, K., Klavans, R., Patek, M., Zoss, A., Biberstine, J. R., Light, R., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464. doi:10.1371/journal.pone.0039464.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. San Rafael, CA: Morgan and Claypool Publishers.

Johnson, D. W., & Jankun-Kelly, T. J. (2008). *A scalability study of web-native information visualization*. Paper presented at the Graphics Interface Conference 2008, Windsor, ON, Canada.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373–397.

LaRowe, G., Ambre, S., Burgoon, J., Ke, W., & Börner, K. (2009). The scholarly database and its utility for Scientometrics research. *Scientometrics, 79*(2), 219–234.

Light, R. P., Polley, D. A., & Börner, K. (2013). *Open data and open code for big science of science studies.* Paper presented at International Society of Scientometrics and Infometrics Conference, Vienna, Austria.

Price, D. J. de Solla (1963). *Little science, big science*. New York: Columbia University Press.

Robertson, G., Ebert, D., Eick, S., Keim, D., & Joy, K. (2009). Scale and complexity in visual analytics. *Information Visualization, 8*(4), 247–253. doi:10.1057/ivs.2009.23.

Suda, B. (2012). The top 20 data visualization tools. *.net*. http://www.netmagazine.com/features/top-20-data-visualisation-tools.