This is a preprint of: Börner, Katy, Nianli Ma, Russell J. Duhon, Angela M. Zoss. 2009.
Science & Technology Assessment Using Open Data and Open Code.
Hsinchun Chen, Ed. *IEEE IS Trends & Controversies*, *IEEE Intelligent Systems*, July/August.

Science & Technology Assessment Using Open Data and Open Code

Katy Börner, Nianli Ma, Russell J. Duhon, Angela M. Zoss
Cyberinfrastructure for Network Science Center, School of Library and Information Science
Indiana University, Wells Library, 1320 E. Tenth Street, Bloomington, IN 47405, USA
katy | nianma | rduhon | amzoss @indiana.edu

Facing the Data Deluge

The number of active researchers exceeds the number of all previous researchers. Researchers either publish or perish. Some areas of science produce more than 40,000 papers a month. Not only library buildings and storage facilities, but also databases are filling up more quickly than they can be built. In addition, there are scientific datasets, algorithms, and tools that need to be mastered in order to advance science. No single man or machine can process and make sense of this enormous stream of data, information, knowledge, and expertise.

The tools we use to access, manage, and utilize our collective knowledge are primitive. Our main means of accessing everything we collectively know is search engines. While this seems to work well for fact-finding, it keeps us scrounging on the floor among confirmed and unconfirmed records. There is no "zoom out" button that provides us with a global view of what we collectively know – how everything is interlinked, what patterns, trends or outliers exist, or in what context a specific piece of knowledge was created or can be used. Without context, intelligent data selection, prioritization, and quality judgments become extremely difficult to make.

This reality leads to increasing specialization of researchers, practitioners, and other knowledge workers, a disconcerting fragmentation of science, a world of missed opportunities for collaboration, and a nightmarish feeling that we are doomed to reinvent the wheel forever. This is a major concern. Scientific results are needed to enable all human beings to live healthy, productive, and fulfilling lives.

Embracing Science Maps

Recent advances in the digitization, federation, mining, and mapping of data make it possible to chart the structure and dynamics of science (Börner et al 2003, Chen 2003, Shiffrin & Börner 2004). The resulting maps of science serve today's explorers navigating the world of scientific research. The maps are generated through analysis of large-scale scholarly datasets in an effort to connect and make sense of the bits and pieces of knowledge they contain. They objectively identify major research areas, experts, institutions, collections, grants, papers, journals, and ideas in domains of interest. They provide overviews of specific fields of science: homogeneity, import-export factors, and relative speed of innovation. They allow one to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research.

Currently, many of the datasets and tools used to generate maps of science are proprietary and particular to each analyst. There are few, if any, standardized tools that can access appropriate data, link them, and present the results in such a way as to enable decision making by non-experts. This paper introduces open data and code that can be used by anyone to map scientific activity and technologically-relevant data in a user-friendly yet professional manner.

Chasing Free Data

The Scholarly Database (SDB) (http://sdb.slis.indiana.edu) at Indiana University evolved from seven years of development towards a free data source for science and technology (S&T) studies (La Rowe 2007). It offers three distinct advantages that are critical for S&T studies:

- Search queries, e.g., for an author/investigator/inventor name or topic term, can be run against multiple databases offering simultaneous retrieval – e.g., all funding, publications, and patents relevant for a query.
- Search results can be downloaded as complete record data dumps in a format easily processed.
- As query results are processed, derivative datasets such as co-author/investigator/inventor tables or patent-citation tables can be downloaded as well.

Currently, SDB provides access to four datasets: 17,764,826 Medline papers provided by the National Library of Medicine (NLM), 1,043,804 funding awards by the National Institutes of Health (NIH) and 174,835 from the National Science Foundation (NSF), and 3,875,694 U.S. Patent and Trademark Office patents (USPTO). Information regarding data provenance, system architecture, table schemas, and search functionality is available on the 'About' page of SDB. Any researcher or layperson can register to search the approximately 23 million records. Currently, the system has over 120 registered users from four continents and over 60 institutions in academia, industry, and government.

Sharing Free Code

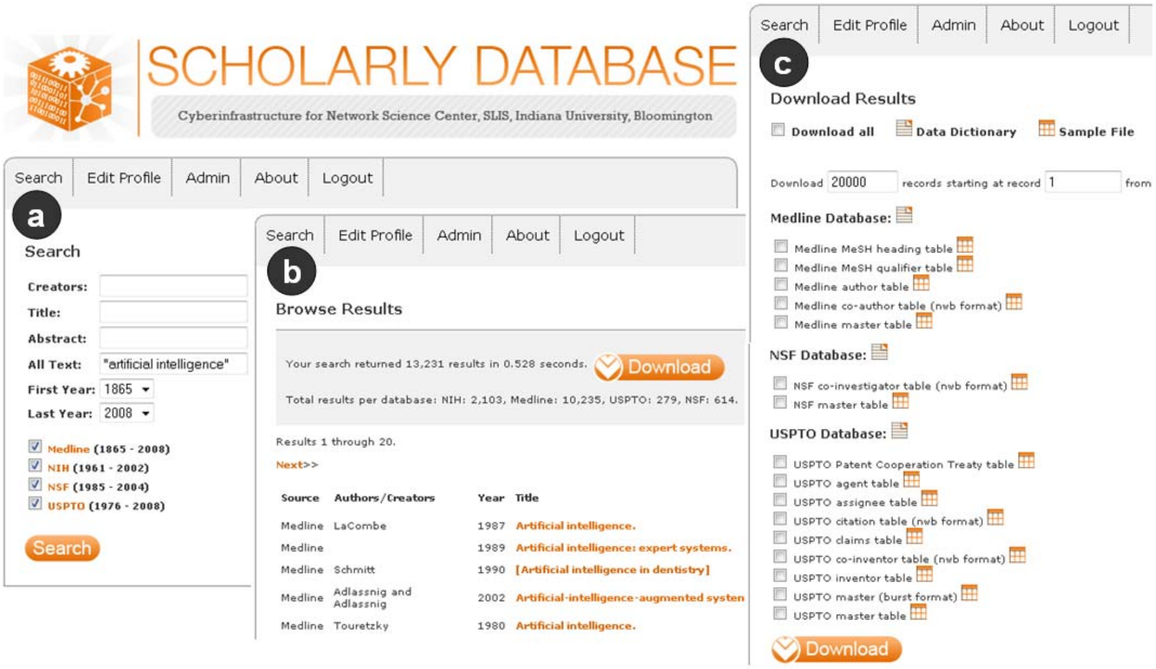
The Network Workbench (NWB) tool (http://nwb.slis.indiana.edu) supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks (NWB Team, 2006). As of February 2009, the tool contains more than 100 plug-ins for the preprocessing, analysis, modeling, and visualization of networks. About 40 of the plug-ins can be applied to or were specifically designed for S&T studies. The NWB tool comes with an associated community wiki (https://nwb.slis.indiana.edu/community) with extensive documentation of algorithms and sample datasets. The tool has been downloaded more than 22,000 times since Dec. 2006.

S&T Studies That Can Be Replicated Anyone

The Scholarly Database, in combination with the NWB Tool, can be used to study S&T professionally in a manner easily replicated by anyone. The three steps are: dataset retrieval and download using SDB, data analysis and visualization using the NWB Tool, and interpretation of results. The sections below show how to begin applying this process.

Data Acquisition

A query for "artificial intelligence" in the "All Text" field over all datasets available in SDB was run; see Fig. 1a. The *Browse Result* page comprises 13,231 records – 10,235 Medline papers, 2,103 NIH awards, 614 NSF awards, and 279 USPTO patents. The top-5 highest scoring records are five Medline papers; see Fig. 1b. Clicking on the record title opens a page with abstract and other information associated with the record.


SCHOLARLY DATABASE
Cyberinfrastructure for Network Science Center, SLIS, Indiana University, Bloomington
Search
Edit Profile
Admin
About
Logout
a
b
c
Search
Download
Download

Search

Creators:
Title:
Abstract:
All Text: "artificial intelligence"
First Year: 1865
Last Year: 2008

☑ Medline (1865 - 2008)
☑ NIH (1961 - 2002)
☑ NSF (1985 - 2004)
☑ USPTO (1976 - 2008)

Browse Results

Your search returned 13,231 results in 0.528 seconds.

Total results per database: NIH: 2,103, Medline: 10,235, USPTO: 279, NSF: 614.

Results 1 through 20.

Next>>

Source	Authors/Creators	Year	Title
Medline	LaCombe	1987	Artificial intelligence.
Medline		1989	Artificial intelligence: expert systems.
Medline	Schmitt	1990	[Artificial intelligence in dentistry]
Medline	Adlassnig and Adlassnig	2002	Artificial-intelligence-augmented systen
Medline	Touretzky	1980	Artificial intelligence.

Download Results

□ Download all   Data Dictionary   Sample File

Download 20000 records starting at record 1 from

Medline Database:

□ Medline MeSH heading table
□ Medline MeSH qualifier table
□ Medline author table
□ Medline co-author table (nwb format)
□ Medline master table

NSF Database:

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

USPTO Database:

□ USPTO Patent Cooperation Treaty table
□ USPTO agent table
□ USPTO assignee table
□ USPTO citation table (nwb format)
□ USPTO claims table
□ USPTO co-inventor table (nwb format)
□ USPTO inventor table
□ USPTO master (burst format)
□ USPTO master table

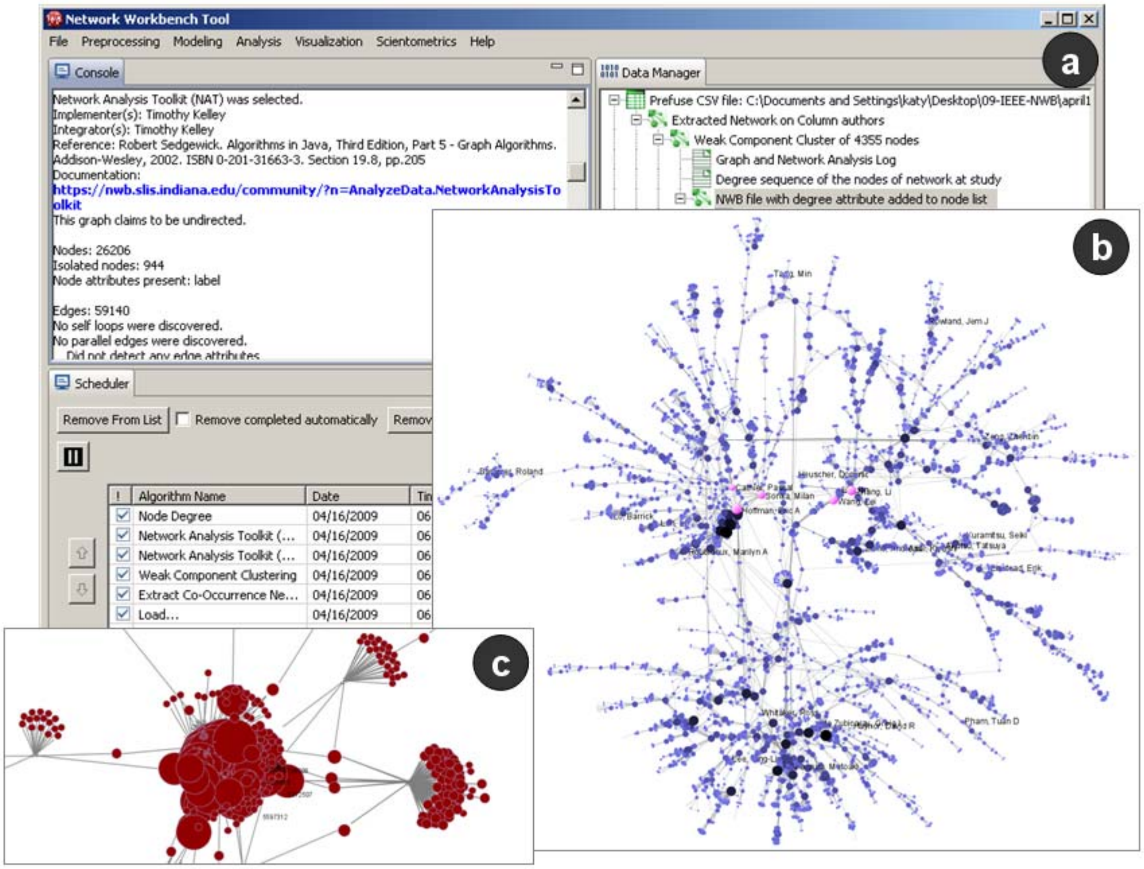
Figure 1: SDB interfaces for search (a), browse results (b), and download results (c).

The *Download Results* page in Fig. 1c allows users to select different types of data. For example, the Medline database offers a master table with general information, an author table that provides paper-author associations, a co-author table that stores the co-author network in a format compatible with the NWB Tool, as well as several other tables. Data dictionaries are provided for each database, and sample datasets are given.

Medline Co-Authorship Network

The Medline master table lists all paper records. The top-five most frequently occurring journals are: *IEEE Transactions on Pattern Analysis and Machine Intelligence* with 761 papers, *IEEE Transactions on Image Processing (*526), *Bioinformatics* (469), *IEEE Transactions on Systems, Man, and Cybernetics – Part B, Cybernetics* (456) and *International Conference on Medical Image Computing and Computer-Assisted Intervention* (443).

The Medline co-author table provides information on paper-author linkages. It can be loaded into the NWB Tool – see NWB Tool (Cyberinfrastructure for Network Science Center, 2009) for details. The table then appears in the *Data Manager* window on the right; see Fig. 2a. Using NWB plug-ins specific to Scientometrics research, the co-authorship network can be extracted. The *Network Analysis Toolkit* computes basic properties: The network has 26,206 author nodes and 59,140 co-author edges. Exactly 944 authors are unconnected (also called isolates). There are almost 5,000 clusters. The largest component with 4,355 nodes and 13,804 edges was extracted using *Weak Component Clustering*. Subsequently, the degree of each node was computed via *Node Degree* analysis. The betweenness centrality (BC) of each node, i.e., the fraction of shortest paths between node pairs that pass through the node of interest, was determined by running the *Node Betweenness Centrality* algorithm. The resulting network was visualized using the *GUESS* graph exploration tool available under the *Visualization* menu.



```
Network Analysis Toolkit (NAT) was selected.
Implementer(s): Timothy Kelley
Integrator(s): Timothy Kelley
Reference: Robert Sedgewick. Algorithms in Java, Third Edition, Part 5 - Graph Algorithms.
Addison-Wesley, 2002. ISBN 0-201-31663-3. Section 19.8, pp.205
Documentation:
https://nwb.slis.indiana.edu/community/?n=AnalyzeData.NetworkAnalysisTo
olkit
This graph claims to be undirected.

Nodes: 26206
Isolated nodes: 944
Node attributes present: label

Edges: 59140
No self loops were discovered.
No parallel edges were discovered.
   Did not detect any edge attributes
```

```
⊟ Prefuse CSV file: C:\Documents and Settings\katy\Desktop\09-IEEE-NWB\april1
   ⊟ Extracted Network on Column authors
      ⊟ Weak Component Cluster of 4355 nodes
            Graph and Network Analysis Log
            Degree sequence of the nodes of network at study
         ⊟ NWB file with degree attribute added to node list
```

Remove From List   □ Remove completed automatically   Remov

!	Algorithm Name	Date	Tin
☑	Node Degree	04/16/2009	06
☑	Network Analysis Toolkit (...	04/16/2009	06
☑	Network Analysis Toolkit (...	04/16/2009	06
☑	Weak Component Clustering	04/16/2009	06
☑	Extract Co-Occurrence Ne...	04/16/2009	06
☑	Load...	04/16/2009	06

Figure 2: NWB Tool interface (a) and GUESS layouts of the largest component of the co-authorship network (b) and zoom into patent citation network (c).

Fig. 2b shows the co-author network with author nodes size and color coded according to their degree, which is the same as the number of distinct co-authors. The top-20 highest degree nodes are labeled. The five nodes with the highest BC value are shown in pink. The highest BC node, “Zhang, Li”, is the author of ten papers from the Medline AI search results. His papers have been published in journals with ISI Subject Categories varying from “Computer Science, Hardware & Architecture” to “Endocrinology & Metabolism”, a diversity that is mirrored by his co-authorship connections to researchers from many different clusters in the network. Medline contains little Computer Science research but covers works in the biomedical sciences. Consequently, the network features major experts that apply AI techniques to biomedical research and practice.

USPTO Patent Citation Network

Loading the *USPTO citation table* and applying the Scientometrics-specific *Extract Directed Network* algorithm, the patent-citation network can be extracted (see the NWB Tool User Manual for details). The USPTO citation network has 3,614 nodes, 8,393 edges, and 107 components. The network shows many network components connected by weak linkages. The 20 nodes with the highest outdegree, i.e., the highest number of citations within the set, are labeled by patent number. Fig. 2c shows a zoom into the set of most highly cited patents. Among them are patent no. 5597312 entitled “Intelligent tutoring method and system”, no. 5372507 describing a “Machine-aided tutorial method”, and no. 5696885 “Expert system and method employing hierarchical knowledge base, and interactive multimedia/hypermedia applications”.

Acknowledgments

We would like to thank the NWB Team and Kevin W. Boyack for stimulating discussions that helped shape this work. This material is based upon work supported by the National Science Foundation under grants SBE-0738111 and IIS-0750993, and the James S. McDonnell Foundation. 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

Börner, Katy, Chaomei Chen, Kevin W. Boyack. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology* (Vol. 37, pp. 179-255). Medford, N.J.: American Society for Information Science and Technology.

Chen, Chaomei (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization.* London: Springer

Cyberinfrastructure for Network Science Center. (2009). *Network Workbench Tool: User Manual*, 1.0.0 beta, http://nwb.slis.indiana.edu/Docs/NWB-manual-1.0.0beta.pdf. (accessed on 04/13/2009)

La Rowe, Gavin, Ambre, Sumeet Adinath, Burgoon, John W., Ke, Weimao & Börner, Katy. (2007). The Scholarly Database and Its Utility for Scientometrics Research. Torres-Salinas, D. & Moed, H. F. (Eds.), *Proceedings of the 11th International Conference on Scientometrics and Informetrics*, Madrid, Spain, June 25-27, pp. 457-462.

NWB Team. (2006). *Network Workbench Tool*. Indiana University, Northeastern University, and University of Michigan, http://nwb.slis.indiana.edu. (accessed on 04/13/2009)

Shiffrin, Richard, Katy Börner. (2004). Mapping Knowledge Domains. *PNAS*, 101(Suppl. 1), 5183–5185.