TLS: Towards a Macroscope for Science Policy Decision Making

NSF SBE-0738111

Katy Börner & Kevin Boyack

Jan. 08 - Dec. 09

Dr. Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu

AAAS-NSF Workshop for SciSIP Grantees, Arlington, VA
Mar 24/25, 2009

Project Goals

(1) Conduct a **detailed analysis of the information needs** of a representative set of science policy makers including existing data, approaches, and tools.

(2) Develop a **theoretic conceptualization of tasks** relevant to science policy-making that map the needs of policy makers to theoretically grounded and practically valuable processing pipelines that transform data into actionable information.

(3) **Design a prototypical tool, a *macroscope***, to see structure, patterns, trends, and outliers in science and technology (S&T) data sets that are too large and complex to be comprehensible to us – just like microscopes and telescopes help us to see things that are too small or too far away. The Macroscope tool development will benefit from the NSF funded *Scholarly Database (SDB)* that provides access to more than 20 million scholarly records, and the *Cyberinfrastructure Shell (CIShell)* which supports the easy plug-and-play of datasets and algorithms and the design of stand-alone tools. Introduce the validated macroscope to a broader audience by means of the *Places & Spaces: Mapping Science* exhibit.

1. Detailed Needs Analysis

A total of 34 science policy makers and researchers at university campus level (8), program officer level (12), and division director level at national, state, and private foundations (10) as well as science policy makers from Europe and Asia (4) were interviewed between Feb. 8th, 2008 and Oct. 2nd, 2008.

Each interview comprised a 40 min, audio-taped, informal discussion on specific information needs, datasets and tools currently used, and information on what a 'dream tool' might look and feel like. There is also a pre-interview questionnaire to acquire demographics and a post-interview questionnaire to get input on priorities.

Data compilation is in progress, should be completed in July 2009, and will be submitted as a journal paper. Some data excerpts are given here.

In the Post-Questionnaire Subjects were asked:

"What are initial thoughts regarding the utility of science of science studies for improving decision making? How would access to datasets and tool speed up and increase the quality of your work?"

Excerpts of answers:

- Two areas have great potential: Understanding S&T as a dynamic system, means to display, visualize and manipulate large interrelated amounts of data in maps that allow better intuitive understanding.
- Look for new areas of research to encourage growth/broader impacts of research--how to assess/ transformative science--what scientific results transformed the field or created a new field/ finding panelists/reviews/ how much to invested until a plateau in knowledge generation is reached/how to define programs in the division.
- Scientometrics as cartography of the evolution of scientific practice that no single actor (even Nobel Laureates) can have. Databases provide a macro-view of the whole of scientific field and its structure. This is needed to make rational decision at the level of countries/states/provinces/regions.
- Understanding where funded scientists are positioned in the global map of science.
- Self-knowledge about effects of funding/ self-knowledge about how to improve funding schemes.
- Ability to see connections between people and ideas, integrate research findings, metadata, clustering career measurement, workforce models, impact (economic/social) on society-interactions between levels of science; lab, institution, agency, Fed Budget, public interests.
- It would be valuable to have tools that would allow one automatically to generate co-citation, co-authorship maps…I am particularly interested in network dynamics.

- It would enable more quantitative decision making in place of an "impression-based" system, and provide a way to track trends, which is not done now.
- When NSF started SciSIP, I was skeptical, but I am more disposed to the idea behind it now although I still don't have a clear idea what scientific metrics will be…..how they will apply across disciplines and whether it's really possible to predict with any accuracy the consequences of any particular decision of a grant award.
- SoS potentially useful to policymakers by providing qualitative and quantitative data on the impacts of science toward government policy goals…ideally these studies would enable policy makers to make better decisions for linking science to progress toward policy goals.
- Tracking faculty's work over time to determine what factors get in the way of productivity and which enhance, e.g. course-releases to allow more time--does this really work or do people who want to achieve do so in spite of barriers.
- I'm not sure that this has relevance to my decision-making. There is a huge need for more reliable data about my organization and similar ones, but that seems distinct from data and tools to study science.
- It would assist me enormously.
- Help to give precedents that would rationalize decisions--help to assess research outside one's major area. Ways of assessing innovation, ways of assessing interactions (among researchers, across areas, outside academia).
- It would allow me to answer questions from members of congress provide visual presentations of data for them.
- Very positive step--could fill important need in understanding innovation systems and organizations.

2. Conceptualizations of Science

See Special Issue of *Journal of Informetrics,* 3(3), Jan 2009.

Science of Science: Conceptualizations and Models of Science

Guest Editors: Katy Börner, Indiana University & Andrea Scharnhorst, Royal Netherlands Academy of Arts and Sciences

This special issue of the journal *Informetrics* aims to improve our understanding of the structure and evolution of science by reviewing and advancing existing conceptualizations and models of scholarly activity.

Existing conceptualizations and models of science have been created by scholars from very different disciplines and backgrounds. They have the form of

- philosophical concepts (Bernal, Kuhn, Popper),
- (utopian) stories (Wells, Lem),
- visual drawings (Otlet),
- empirical measurements (Price, Garfield), or
- mathematical theories (Goffman, Yablonski)

among others.

It is our belief that a theoretically grounded and practically useful shared conceptualization of science can provide the intellectual framework to interlink and puzzle together the hundreds of science models in existence today. This is analogous to how meteorologists or seismologists integrate rather different local weather models or seismic hazard predictions into a global coherent model that has higher predictive value and broader coverage. With this issue we aim to start an interdisciplinary discourse towards a science of science models.

The design of such a conceptualization requires the identification of the

- Boundaries of the system or object.
- Basic building blocks of science, e.g., units of analysis or key actors.
- Interactions of building blocks, e.g., via coupled networks.
- Basic mechanisms of growth and change.

Editorial is available at http://ivl.slis.indiana.edu/km/pub/2009-borner-scharnhorst-joi-sos-intro.pdf

3. Macroscope Tool

Benefits from and extends the *Scholarly Database* at IU

"From Data Silos to Wind Chimes"

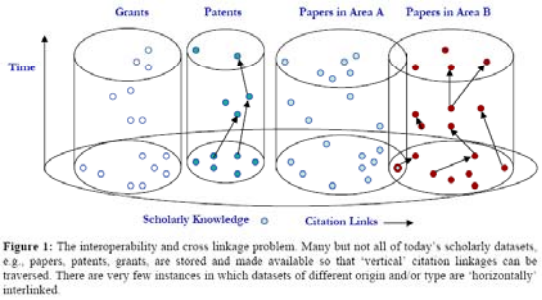
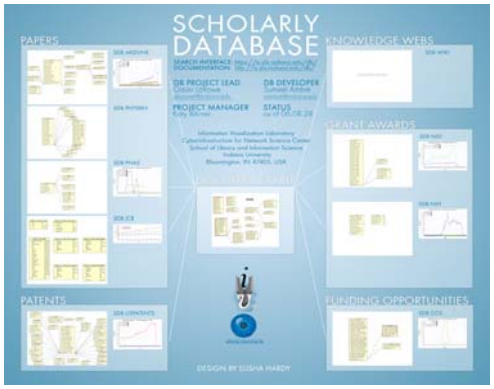


Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type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



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

Scholarly Database: # Records & Years Covered

Datasets available via the Scholarly Database (* internally)

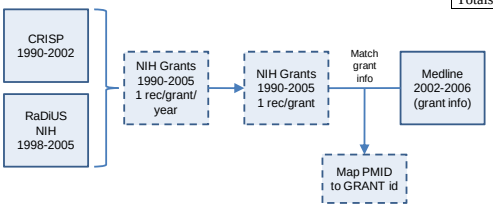
Dataset	# Records	Years Covered	Updated	Restricted Access
Medline	17,764,826	1898-2008	Yes	
PhysRev	398,005	1893-2006		Yes
PNAS	16,167	1997-2002		Yes
JCR	59,078	1974, 1979, 1984, 1989 1994-2004		Yes
USPTO	3, 710,952	1976-2008	Yes*	
NSF	174,835	1985-2002	Yes*	
NIH	1,043,804	1961-2002	Yes*	
Total	**23,167,642**	**1893-2006**	**4**	**3**

Aim for comprehensive time, geospatial, and topic coverage.

Grant-Article Linking

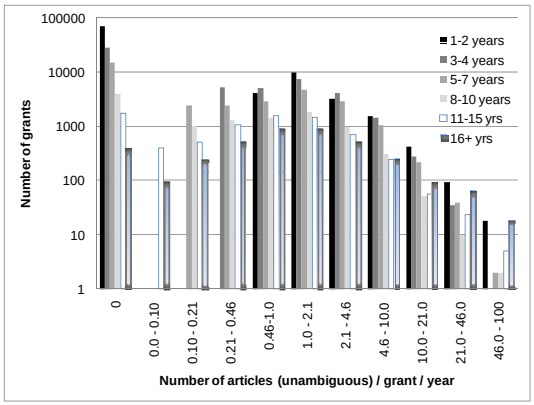
- NIH grant data from CRISP and RaDiUS were linked to Medline papers using the grant information strings in Medline (dirty data using dozens of formats)
- 94% of grant strings were matched with a grant number
- Enables future input-output studies

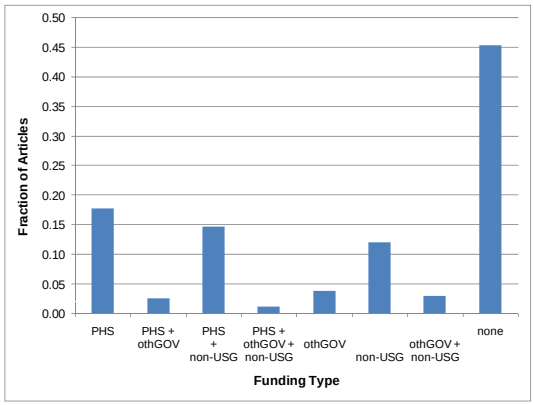
Institute	possible matches	% matched	unambig	ambig	no match	# unique grants	# unique articles	% multi-inst arts
NCI	93,897	92.0%	82,539	3,883	7,475	11,314	51,521	36.1%
NHLBI	82,525	93.5%	72,172	4,952	5,401	9,600	41,901	41.6%
NIGMS	58,749	95.3%	49,886	6,103	2,760	8,421	43,640	35.3%
NIDDK	52,390	95.4%	45,857	4,125	2,408	6,987	31,405	49.5%
NIAID	51,953	92.5%	43,087	4,976	3,890	8,348	30,149	42.8%
NINDS	37,054	94.9%	32,774	2,377	1,903	5,954	24,467	46.7%
NIMH	36,859	93.8%	31,392	3,186	2,281	6,092	21,401	40.0%
NCRR	31,373	95.1%	27,601	2,233	1,539	1,470	24,271	72.7%
NIA	27,424	93.9%	24,104	1,659	1,661	3,369	16,489	50.4%
NICHD	26,691	93.1%	22,596	2,248	1,847	3,975	17,041	49.3%
NIDA	21,145	95.3%	18,234	1,924	987	3,394	11,812	43.1%
NEI	18,835	95.6%	16,183	1,824	828	2,604	10,610	27.8%
NIEHS	16,220	94.3%	14,280	1,008	932	1,540	10,064	52.1%
NIAMS	15,401	93.4%	13,522	856	1,023	2,236	9,931	50.3%
NIAAA	10,643	94.3%	8,885	1,154	604	1,700	5,973	43.3%
NIDCD	9,200	95.0%	7,706	1,033	461	1,916	5,830	29.9%
NIDCR	9,094	94.3%	8,025	554	515	1,536	5,922	38.6%
NIBIB	4,381	95.5%	4,124	60	197	727	3,415	56.5%
FIC	2,813	87.7%	2,404	64	345	547	2,178	54.1%
NINR	2,661	88.2%	2,314	32	315	784	1,996	23.2%
NHGRI	2,559	93.2%	2,098	286	175	492	2,023	50.3%
NCCAM	1,724	93.0%	1,580	23	121	331	1,335	48.5%
NLM	1,609	85.6%	1,362	15	232	232	1,109	35.1%
NCMHHD	559	74.2%	413	2	144	65	373	62.5%
WHI	205	97.1%	199	0	6	41	35	40.0%
Others	598	4.5%	27	0	571	15	26	46.2%
Totals	616,562	93.7%	533,364	44,577	38,621	83,690	374,917	44.0%



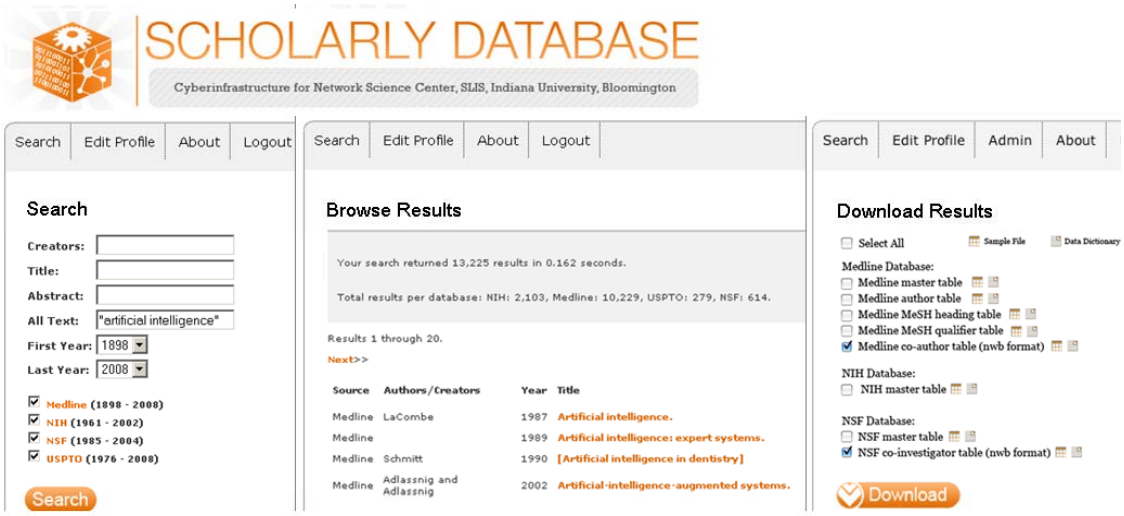
Subsequent Analysis From Matches

- Short grants (1-2 years) produce more papers per year than long grants (3-15 years).
- Data not normalized for grant size.
- Acknowledgement of NIH funding in Medline-indexed articles does seem to be reasonably complete.
- "None" category size consistent with other analyses – these are not "missing NIH" data.





Scholarly Database: Web Interface



Anybody can register for free at https://sdb.slis.indiana.edu to search the about 23 million records and download results as data dumps.

Currently the system has over 100 registered users from academia, industry, and government from over 60 institutions and four continents.

3. Macroscope Tool

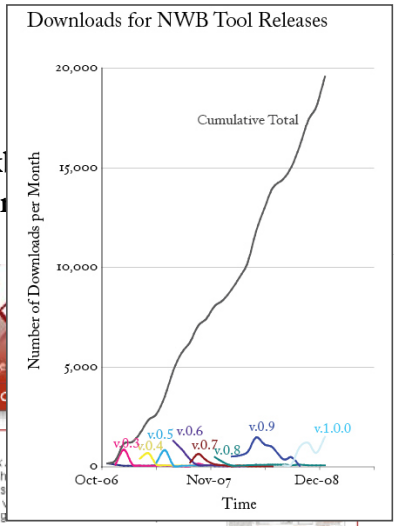
Builds on and extends the Network Workl
ultimately be 'packaged' as a SciPolicy' br

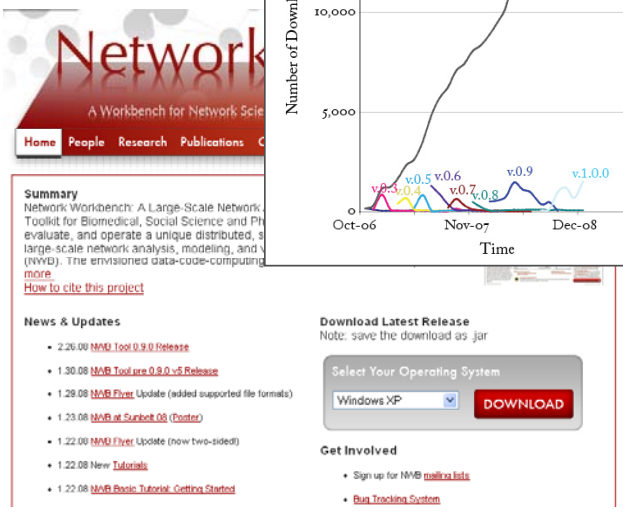
The Network Workbench (NWB) tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

In Feb. 2009, the tool provides more 100 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 40 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 18,000 times since Dec. 2006.





http://nwb.slis.indiana.edu/

Algorithms Currently Available

See https://nwb.slis.indiana.edu/community July 1st, 2008

Preprocessing Edit

- **Remove Nodes**
  - Extract Top Nodes
  - Extract Nodes Above or Below Val
  - Delete High Degree Nodes
  - Delete Random Nodes
  - Delete Isolates
- **Remove Edges**
  - Extract Top Edges
  - Extract Edges Above or Below Val
  - Remove Self Loops
  - Trim By Degree[2]
  - Pathfinder Network Scaling
- **Sampling**
  - Snowball Sampling (n nodes)
  - Node Sampling
  - Edge Sampling
- **Transformations**
  - Symmetrize
  - Dichotomize
  - Multipartite Joining

Modeling Edit

- **General**
  - Random Graph
  - Watts-Strogatz Small World
  - Barabási-Albert Scale-Free
- **Structured**
  - CAN
  - Chord
- **Unstructured**
  - Hypergrid
  - PRU
- **Other**
  - TARL
  - Discrete Network Dynamics

Analysis Edit

- **General Purpose**
  - Network Analysis Toolkit[2]
- **Unweighted & Undirected**
  - **Based on degree/**
    - Node Degree
    - Node Distribution
  - **Based on clustering**
    - k-Nearest Neighbor
    - Watts Strogatz Clustering Coefficie
    - Watts Strogatz Clustering Coefficie
  - **Based on path**
    - Diameter
    - Average Shortest Path
    - Shortest Path Distribution
    - Node Betweenness Centrality
  - **Based on components**
    - Connected Components
    - Weak Component Clustering
  - **K-Core**
    - Extract K-Core[2]
    - Annotate K-Coreness[2]
- **Unweighted & Directed**
  - **Based on degree**
    - Node Indegree
    - Node Outdegree
    - Indegree Distribution
    - Outdegree Distribution
  - **Based on local graph structure**
    - k-Nearest Neighbor
    - Single Node In-Out Degree Correla
  - **Unnamed Category?**
    - Page Rank
  - **Based on local graph structure**
    - Dyad Reciprocity[2]
    - Arc Reciprocity[2]
    - Adjacency Transitivity[2]
  - **Based on components**
    - Weak Component Clustering
    - Extract Attractors[2]

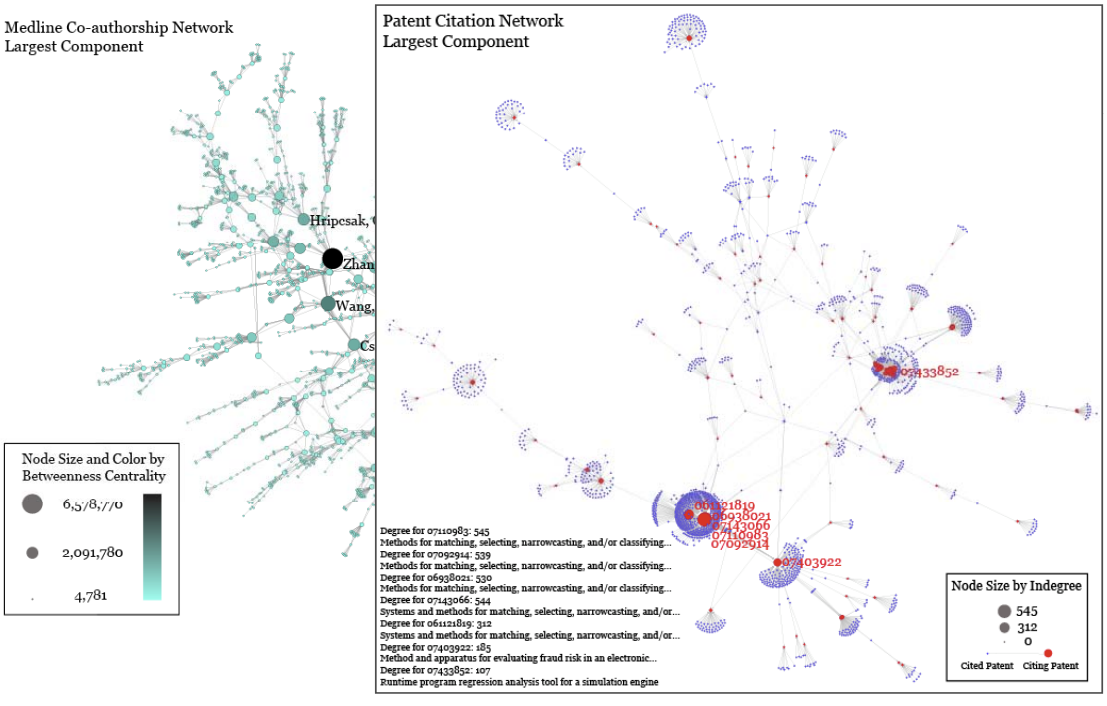
Visualization Edit

- **Tools**
  - GUESS
  - GnuPlot[2]
- **Predefined Positions Layout**
  - DrL (VxOrd)
  - Pre-defined Positions (prefuse beta)[2]
- **Move**
  - Circular
- **Tree Layouts**
  - Radial Tree (prefuse alpha)
  - Radial Tree with Annotations (prefuse beta)[2]
  - Tree Map
  - Tree View
  - Balloon Graph (prefuse alpha)[2]
- **Network Layouts**
  - Force Directed with Annotation (prefuse beta)
  - Kamada-Kawai (JUNG)
  - Fruchterman-Reingold (JUNG)
  - Fruchterman-Reingold with Annotation (prefuse beta)
  - Spring (JUNG)
  - Small World (prefuse alpha)
- **Other Layouts**
  - Parallel Coordinates (demo)[2]
  - LaNet (k-Core Decomposition)

Scientometrics Edit

- **Extract Network From Table**
  - Extract Co-Authorship Network
  - Extract Co-Occurrence Network From Table[2]
  - Extract Directed Network From Table[2]
- **Extract Network From Another Network**
  - Extract Bibliographic Coupling Similarity Network
  - Extract Co-Citation Similarity Network[2]
- **Cleaning**
  - Remove ISI Duplicate Records
  - Detect Duplicate Nodes
  - Remove Rows With Multitudinous Fields[2]

SciPolicy Studies - Using Open Data and Open Code

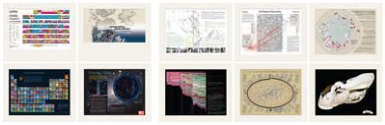


Mapping Science Exhibit – 10 Iterations in 10 years

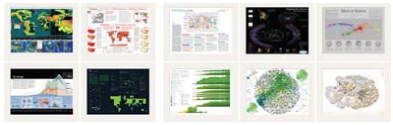
http://scimaps.org/

The Power of Maps (2005)

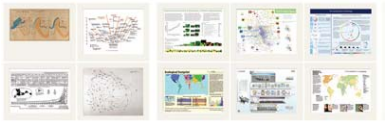
The Power of Reference Systems (2006)



The Power of Forecasts (2007)



Science Maps for Economic Decision Makers (2008)



Science Maps for Science Policy Makers (2009)

Science Maps for Scholars (2010)

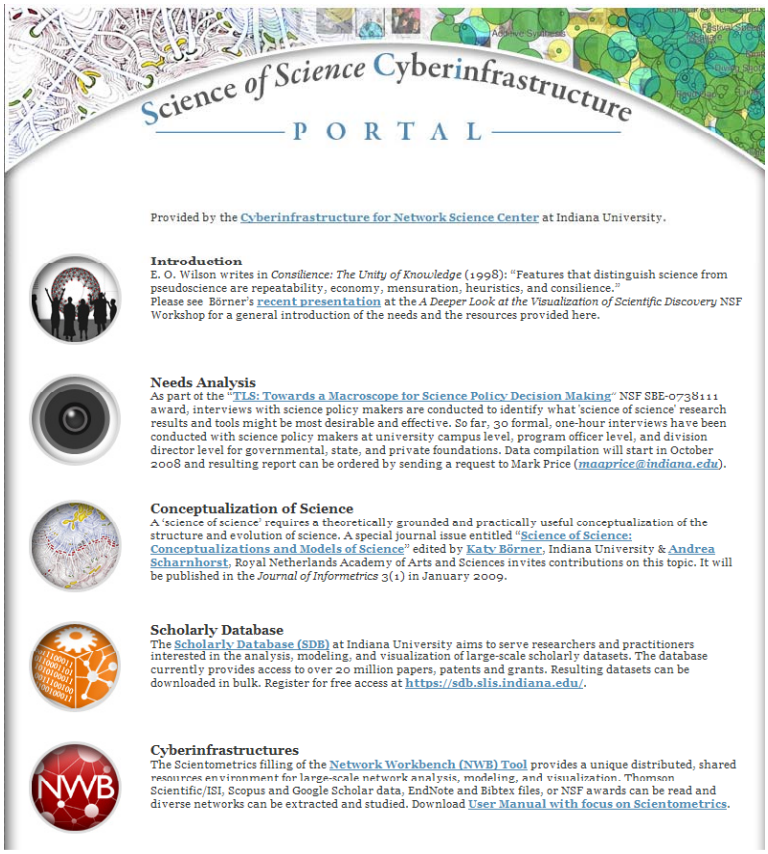
Science Maps as Visual Interfaces to Digital Libraries (2011)

Science Maps for Kids (2012)

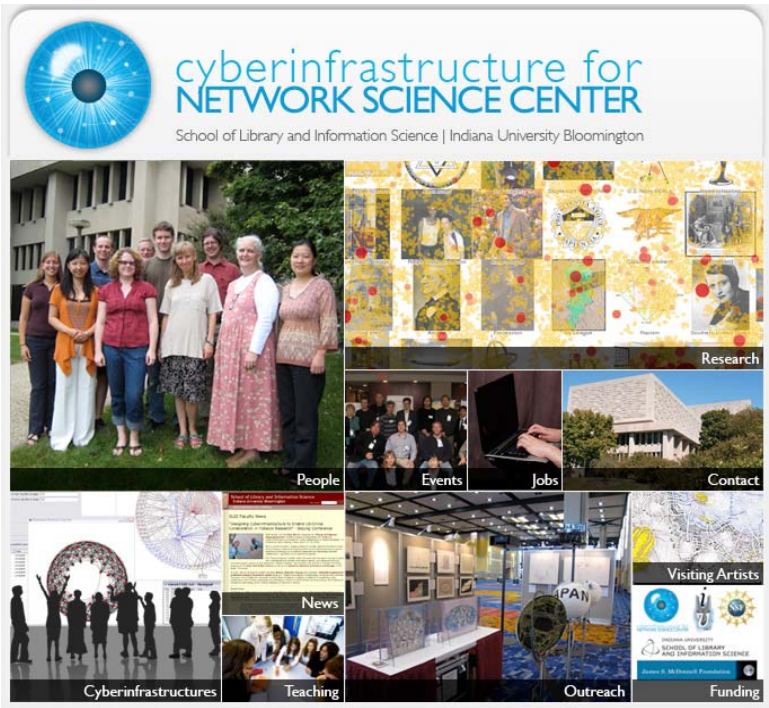
Science Forecasts (2013)

How to Lie with Science Maps (2014)

Exhibit has been shown in 49 venues on four continents. Also at
- NSF, 10th Floor, 4201 Wilson Boulevard, Arlington, VA.
- Chinese Academy of Sciences, China, May 17-Nov. 15, 2008.
- University of Alberta, Edmonton, Canada, Nov 10-Jan 31, 2009
- Center of Advanced European Studies and Research, Bonn, Germany, Dec. 11-19, 2008.



http://sci.slis.indiana.edu



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