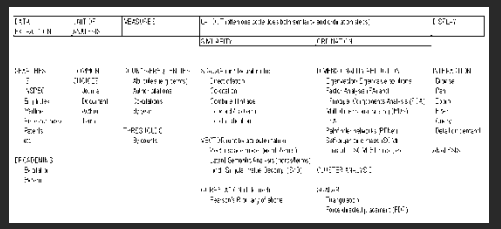
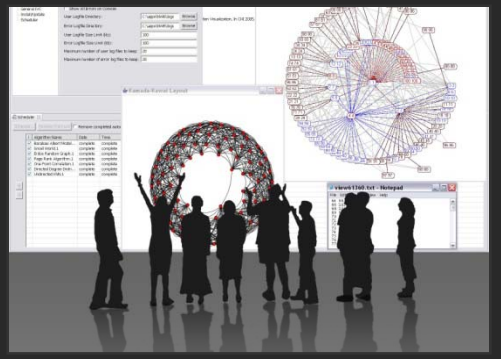
Towards a Science of Science (Policy)

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

NSF Workshop on "A Deeper Look at the Visualization of Scientific Discovery in the Federal Context", Arlington, VA
Sept 12th, 2008





Overview

1. Features of Science

2. Detailed Needs Analysis
3. Conceptualization of Science / Terms & Definitions
4. Data Quality, Coverage and Interlinkage
5. Algorithms and Cyberinfrastructures
6. Case Studies and Evaluation

↓

7. Standards and Practices

8. Dissemination of Results

1. Features of Science

E. O. Wilson in *Consilience: The Unity of Knowledge* (1998) writes:

"Features that distinguish science from pseudoscience are repeatability, economy, menuration, heuristics, and consilience."

For a true science of science (policy) this can be detailed as:

- *Repeatability* refers to the fact that any science analysis, model, or map can be rerun or regenerated by a scholar other than the author. This requires that datasets are accessible and documented in a way that they can be recompiled, software is made available or is documented in sufficient detail so that it can be reimplemented and run with the exact same parameter settings.
- *Economy* entails that results are presented in a form that is both simplest and most pleasing not only for the expert but also for a general audience.
- *Mensuration* means properly measured, using universally accepted scales in an unambiguous fashion.
- *Heuristics* – the best science stimulates further discovery.
- *Consilience* – those explanations and results that are consistent with each other are most likely to survive.

2. Detailed Needs Analysis

As part of

- TLS: Towards a Macroscope for Science Policy Decision Making. NSF SBE-0738111 award (Katy Börner, Weixia Huang, Kevin W. Boyack).

that aims to design qualitatively new tools for science policy makers, we are conducting interviews with science policy makers at different levels of the science enterprise to identify what 'science of science' research results and tools might be most beneficial.

Each interview comprises 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 your input on priorities.

Results of 30+ interviews should become available at the end of 2008.

But see also Paul Gemperline's slides.

3. Conceptualization of Science / Terms & Definitions

- Identify 'basic units' of science, their 'interlinkages', and major static and dynamic properties of interest.
- Define (ideally operationalize) major terms such as impact, interdisciplinarity, etc.
- What conceptual description might best represent science? See also Special Issue of *Journal of Informetrics,* 3(1), 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

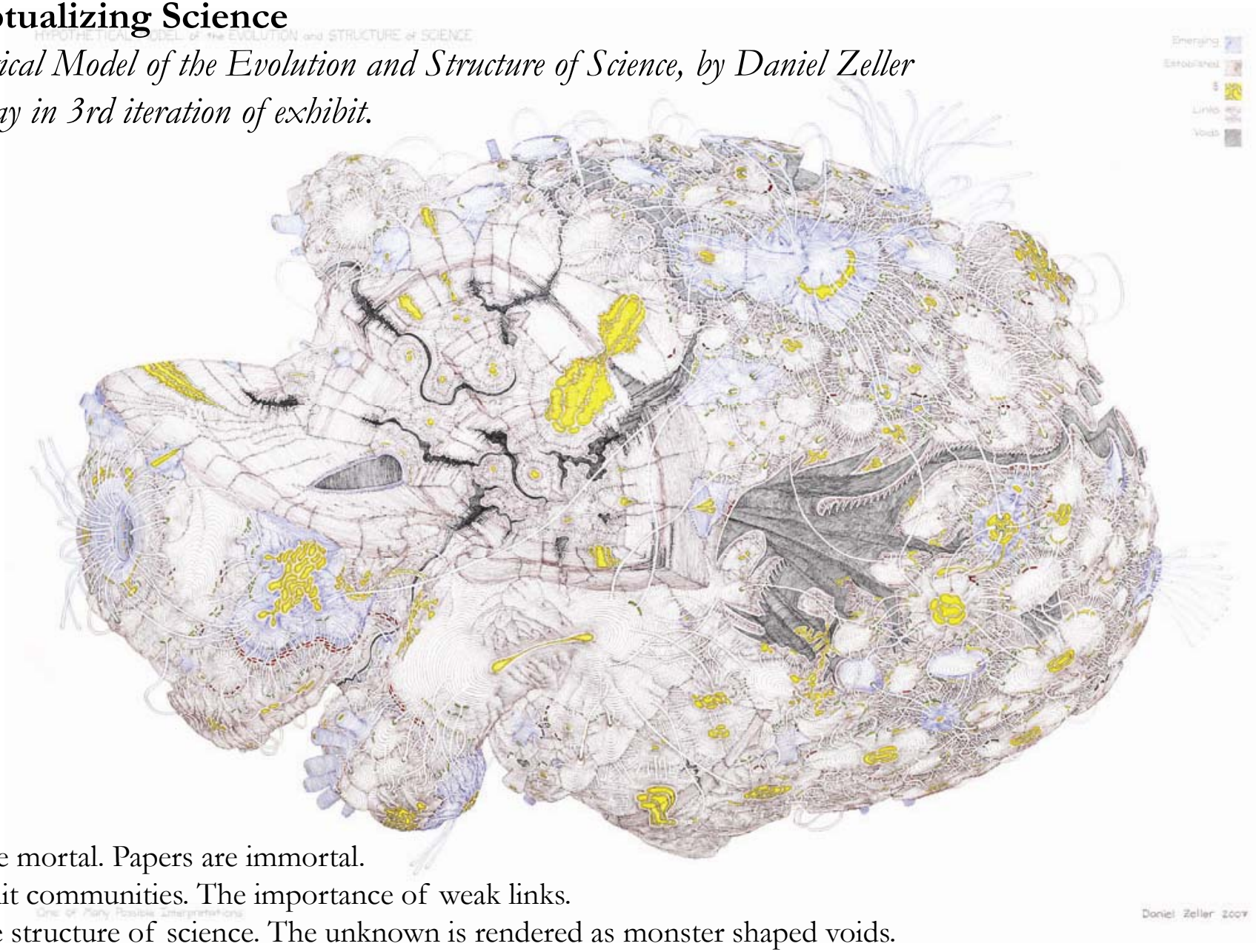
- Boundaries of the system or object.
- Basic building blocks of science, e.g., units of analysis or key actors.
- Interactions of building blocks, e.g., via coupled networks.
- Basic mechanisms of growth and change.
- Existing laws (static and dynamic).

Ideally, the conceptualizations can be also presented in a visual form so that disciplinary and cultural boundaries can be bridged more easily.

Conceptualizing Science

Hypothetical Model of the Evolution and Structure of Science, by Daniel Zeller

On display in 3rd iteration of exhibit.



Authors are mortal. Papers are immortal.
Densely knit communities. The importance of weak links.
Cumulative structure of science. The unknown is rendered as monster shaped voids.
Impact of funding on science (yellow).

3. Conceptualization of Science / Terms & Definitions

Process of Analyzing and Mapping Knowledge Domains

DATA EXTRACTION	UNIT OF ANALYSIS	MEASURES	LAYOUT (often one code does both similarity and ordination steps)		DISPLAY
			SIMILARITY	ORDINATION	
SEARCHES ISI INSPEC Eng Index Medline ResearchIndex Patents etc. BROADENING By citation By terms	COMMON CHOICES Journal Document Author Term	COUNTS/FREQUENCIES Attributes (e.g. terms) Author citations Co-citations By year THRESHOLDS By counts	SCALAR (unit by unit matrix) Direct citation Co-citation Combined linkage Co-word / co-term Co-classification VECTOR (unit by attribute matrix) Vector space model (words/terms) Latent Semantic Analysis (words/terms) incl. Singular Value Decomp (SVD) CORRELATION (if desired) Pearson's R on any of above	DIMENSIONALITY REDUCTION Eigenvector/ Eigenvalue solutions Factor Analysis (FA) and Principal Components Analysis (PCA) Multi-dimensional scaling (MDS) LSA Pathfinder networks (PFNet) Self-organizing maps (SOM) includes SOM, ET-maps, etc. CLUSTER ANALYSIS SCALAR Triangulation Force-directed placement (FDP)	INTERACTION Browse Pan Zoom Filter Query Detail on demand ANALYSIS

Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003) Visualizing Knowledge Domains. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Volume 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.

4. Data Quality, Coverage and Interlinkage

"From Data Silos to Wind Chimes"

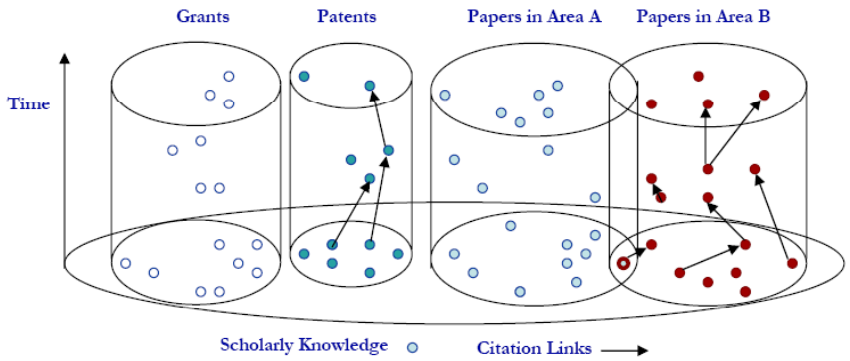
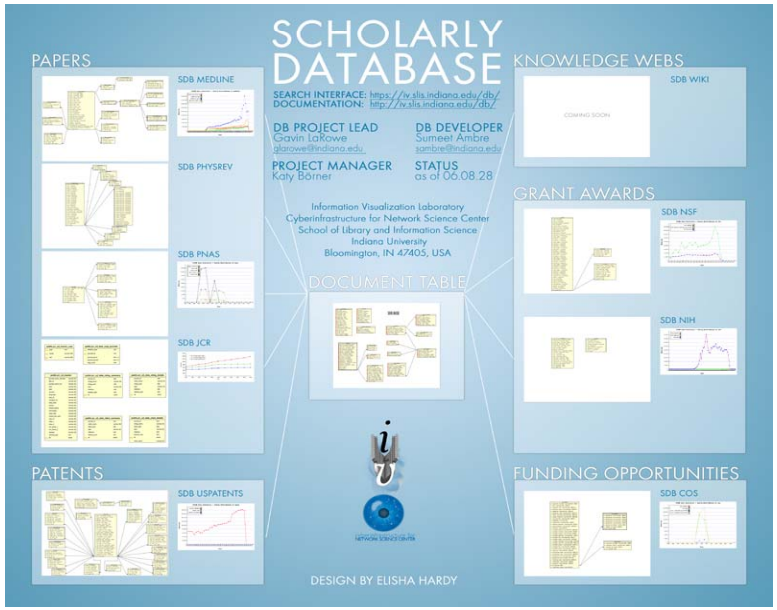


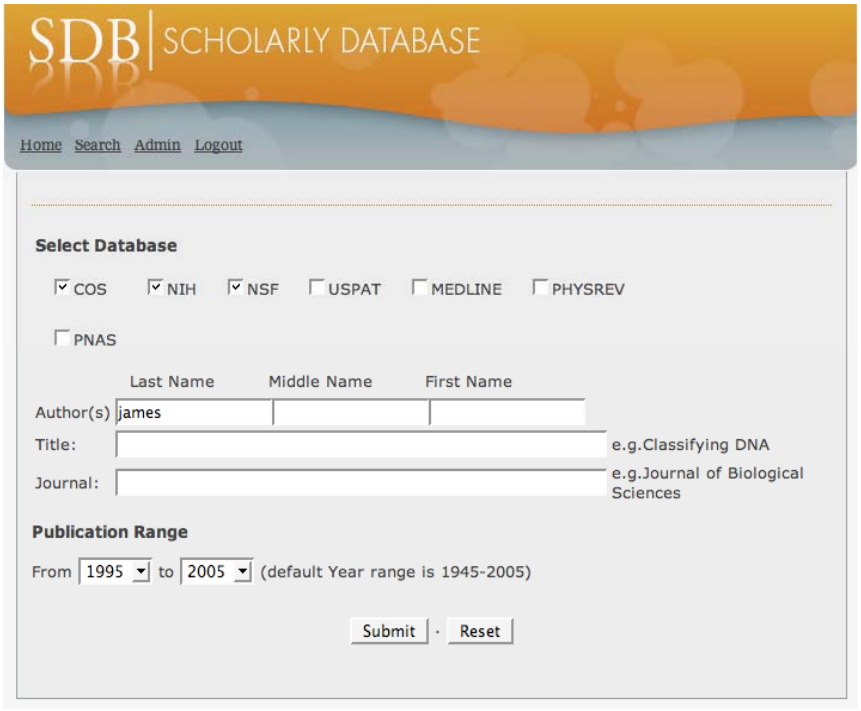
Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type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

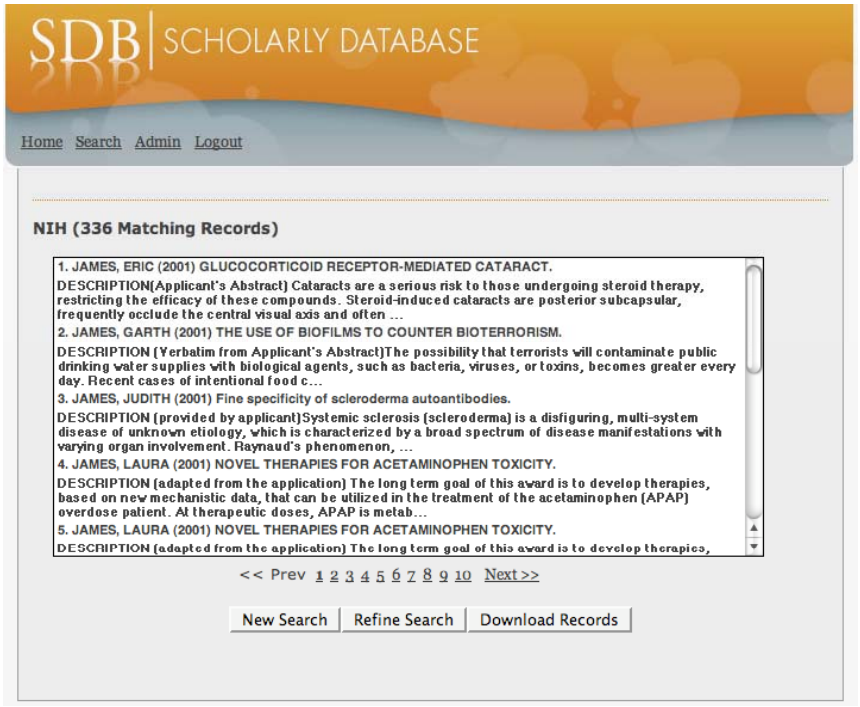


- Interlink creators, data, software/tools, publications, patents, IP, funding, etc.
- Create public databases that any scholar can use. Share the burden of data cleaning and federation.

Scholarly Database: Web Interface

Search across publications, patents, grants.
Download records and/or (evolving) co-author, paper-citation networks.





Register for free access at https://sdb.slis.indiana.edu.

Scholarly Database: # Records & Years Covered

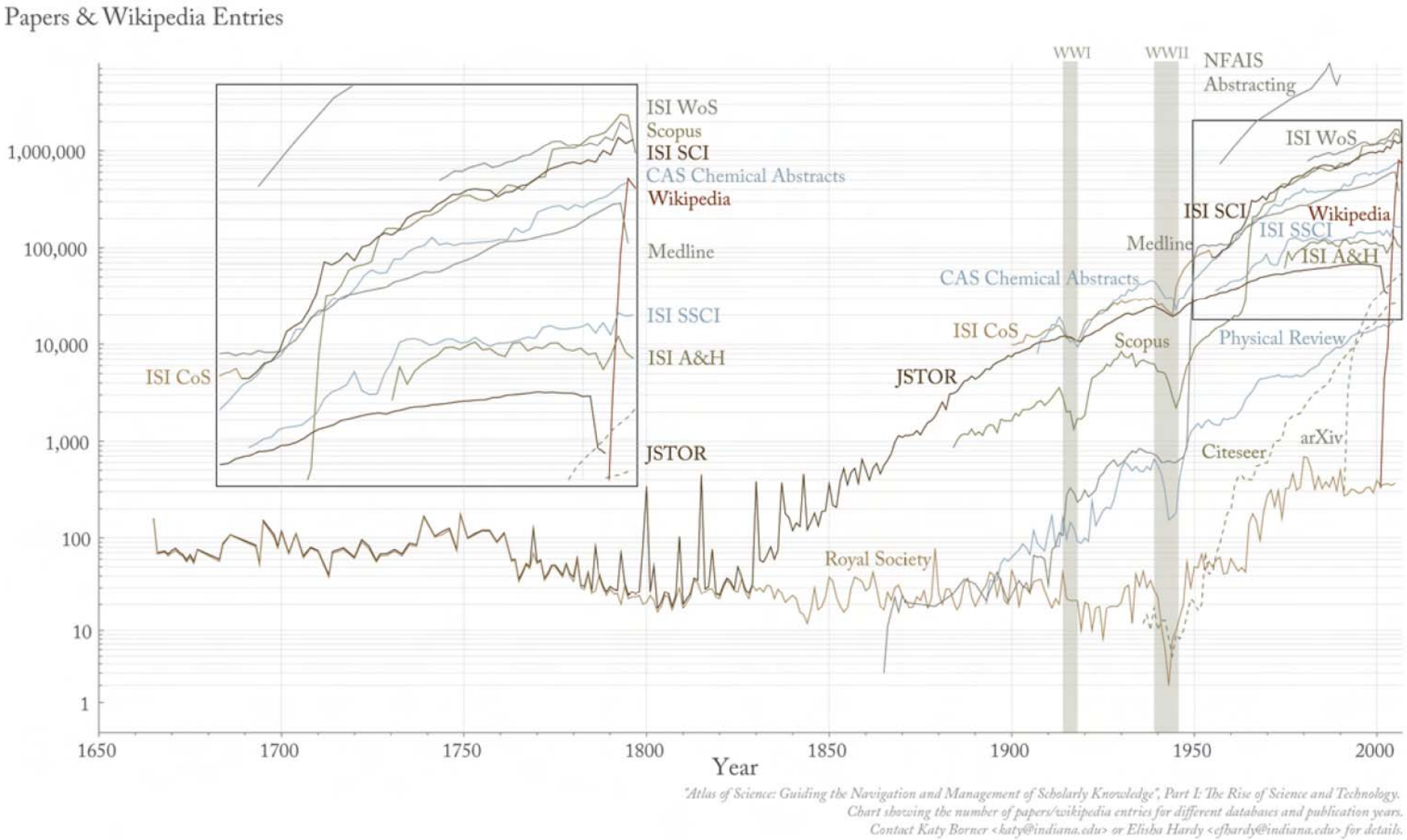
Datasets available via the Scholarly Database (* future feature)

Dataset	# Records	Years Covered	Updated	Restricted Access
Medline	13,149,741	1965-2005	Yes	
PhysRev	398,005	1893-2006		Yes
PNAS	16,167	1997-2002		Yes
JCR	59,078	1974, 1979, 1984, 1989 1994-2004		Yes
USPTO	3,179,930	1976-2004	Yes*	
NSF	174,835	1985-2003	Yes*	
NIH	1,043,804	1972-2002	Yes*	
Total	**18,021,560**	**1893-2006**	**4**	**3**

Aim for comprehensive time, geospatial, and topic coverage.

Holdings of Scholarly Databases/Wikipedia, 1665 to 2006

#papers per publication year



"Atlas of Science: Guiding the Navigation and Management of Scholarly Knowledge", Part I: The Rise of Science and Technology.
Chart showing the number of papers/wikipedia entries for different databases and publication years.
Contact Katy Borner <katy@indiana.edu> or Elisha Hardy <eehardy@indiana.edu> for details.

5. Algorithms and Cyberinfrastructures

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

http://epic.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[?]
  - 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[?]
- **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[?]
    - Annotate K-Coreness[?]
- **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[?]
    - Arc Reciprocity[?]
    - Adjacency Transitivity[?]
  - **Based on components**
    - Weak Component Clustering
    - Extract Attractors[?]

Visualization Edit

- **Tools**
  - GUESS
  - GnuPlot[?]
- **Predefined Positions Layout**
  - DrL (VxOrd)
  - Pre-defined Positions (prefuse beta)[?]
- **Move**
  - Circular
- **Tree Layouts**
  - Radial Tree (prefuse alpha)
  - Radial Tree with Annotations (prefuse beta)[?]
  - Tree Map
  - Tree View
  - Balloon Graph (prefuse alpha)[?]
- **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)[?]
  - LaNet (k-Core Decomposition)

Scientometrics Edit

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

6. Case Studies and Evaluation

Desirable Features

- Theoretically grounded
- Practically relevant
- Testable hypotheses
- Well documented results
- Repeatable

Types of Studies

- Individual, local, global level
- Temporal, geospatial, semantic, network analysis

Result Presentation

- Textual, tabular, charts, geo maps, science maps, other reference systems

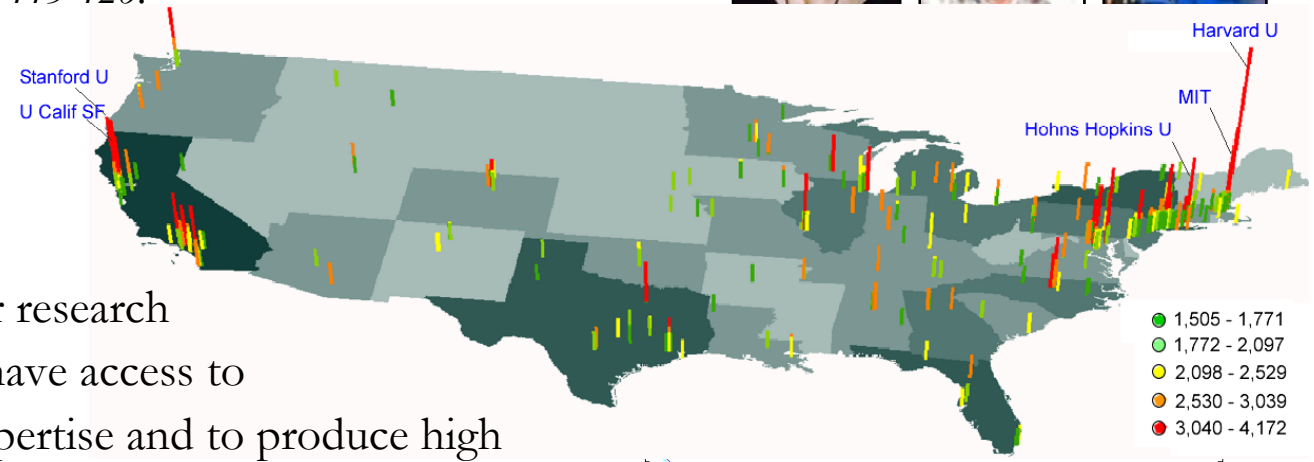
Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions

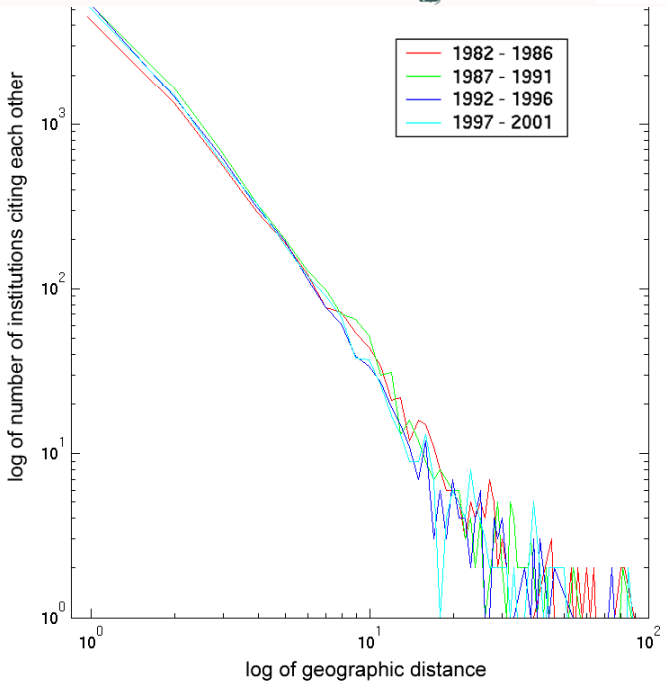


Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao. (2006) Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. Scientometrics. 68(3), pp. 415-426.

Research questions:

1. Does space still matter in the Internet age?
2. Does one still have to study and work at major research institutions in order to have access to high quality data and expertise and to produce high quality research?
3. Does the Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?



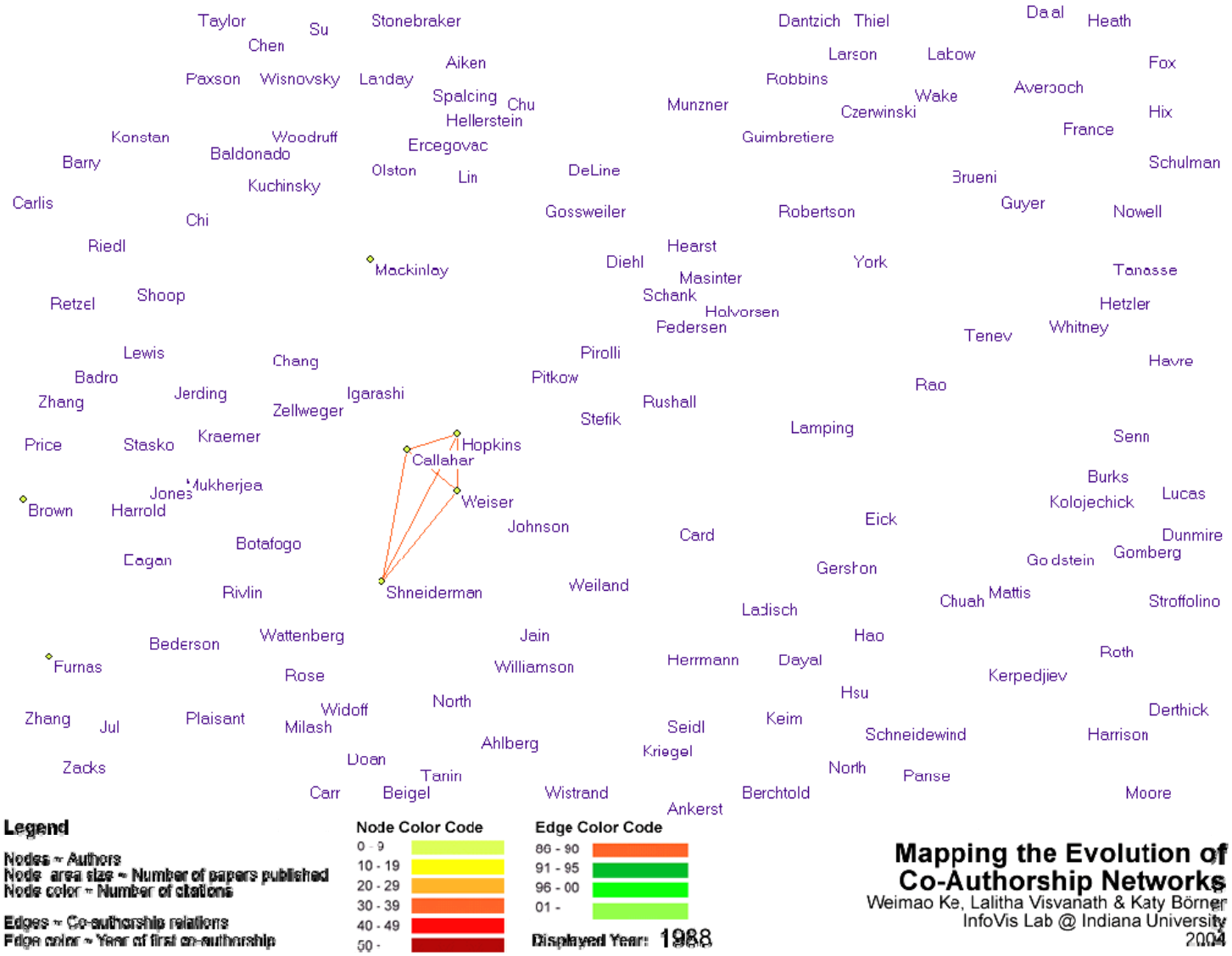


Contributions:

- Answer to Qs 1 + 2 is YES.
- Answer to Qs 3 is NO.
- Novel approach to analyzing the dual role of institutions as information producers and consumers and to study and visualize the diffusion of information among them.

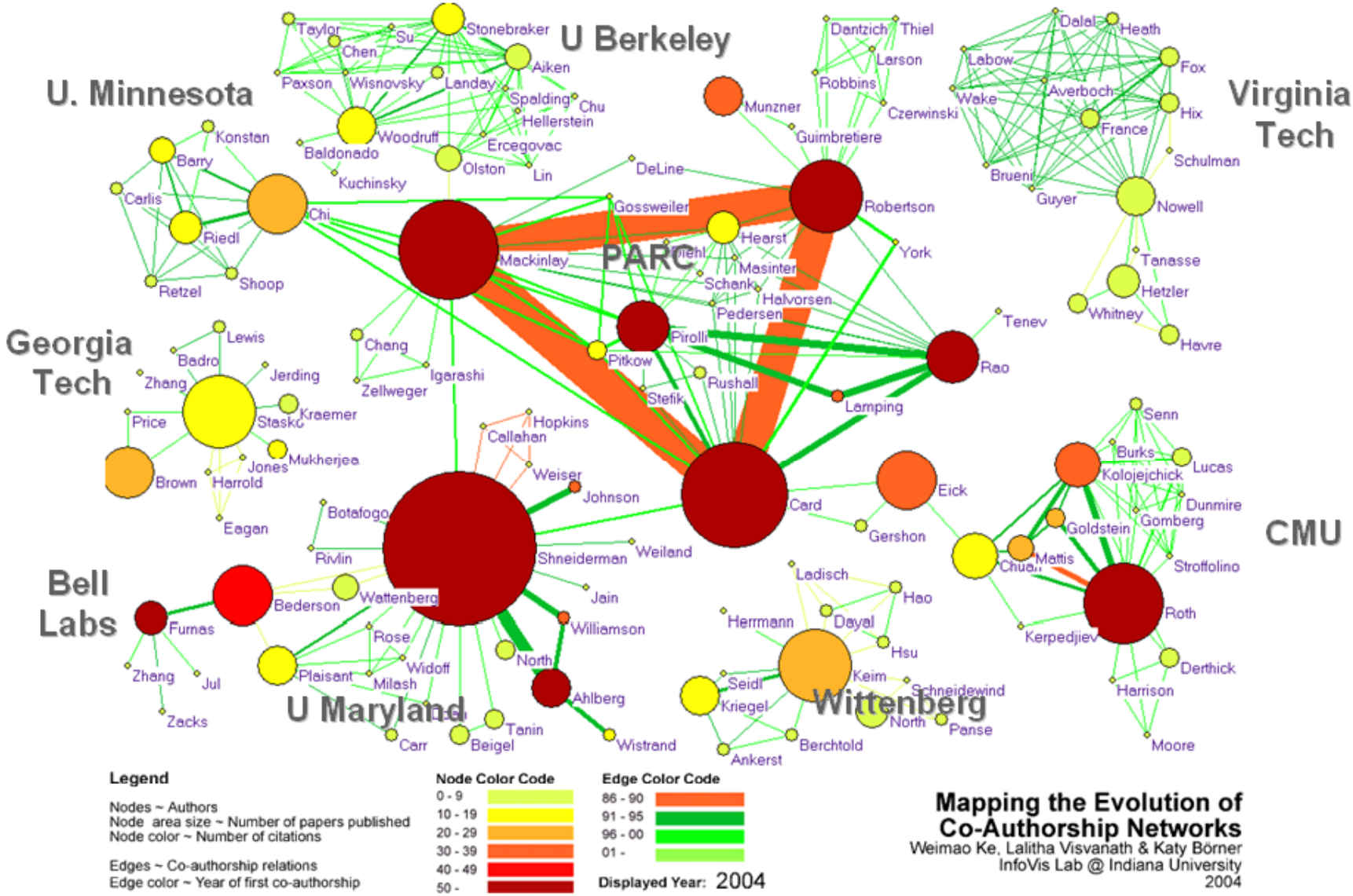
Mapping the Evolution of Co-Authorship Networks

Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest.



Mapping the Evolution of Co-Authorship Networks

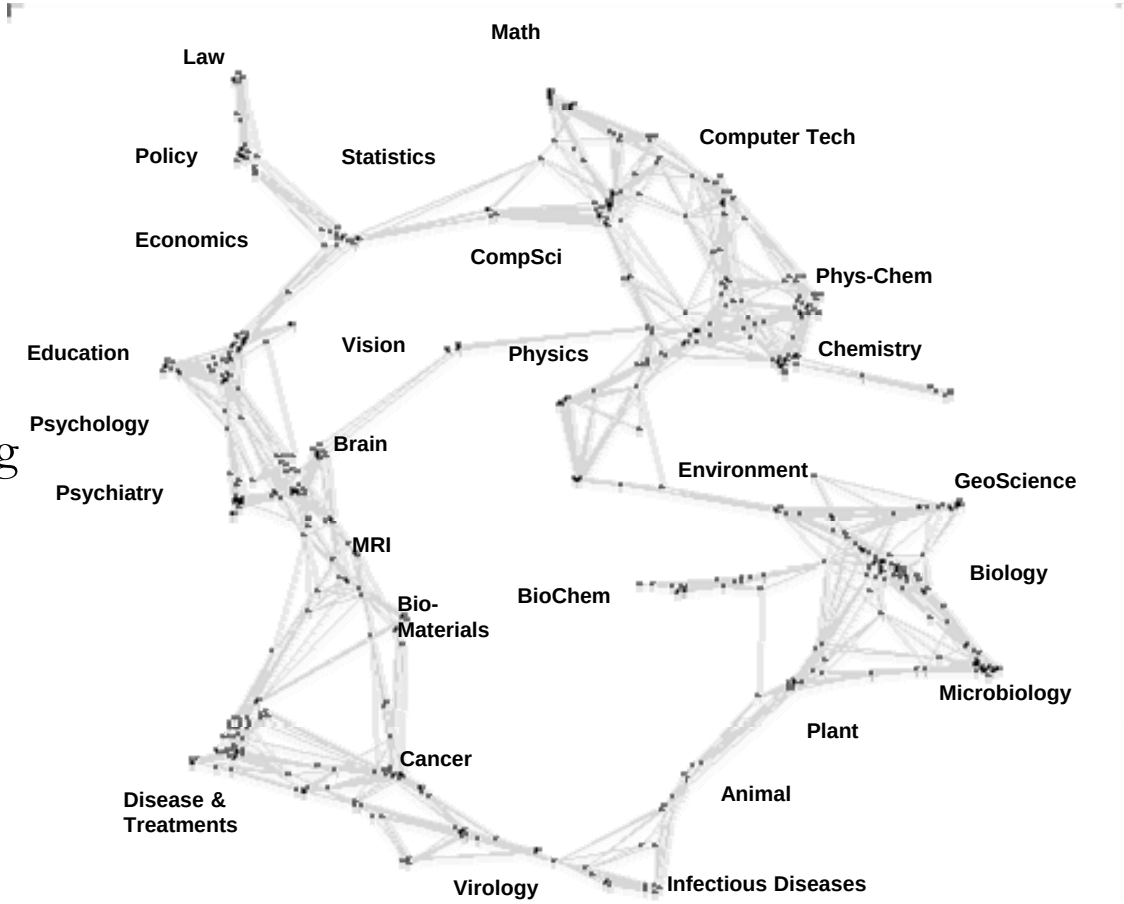
Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest



Latest 'Base Map' of Science

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007). Mapping the Structure and Evolution of Chemistry Research. 11th International Conference on Scientometrics and Informetrics. pp. 112-123.

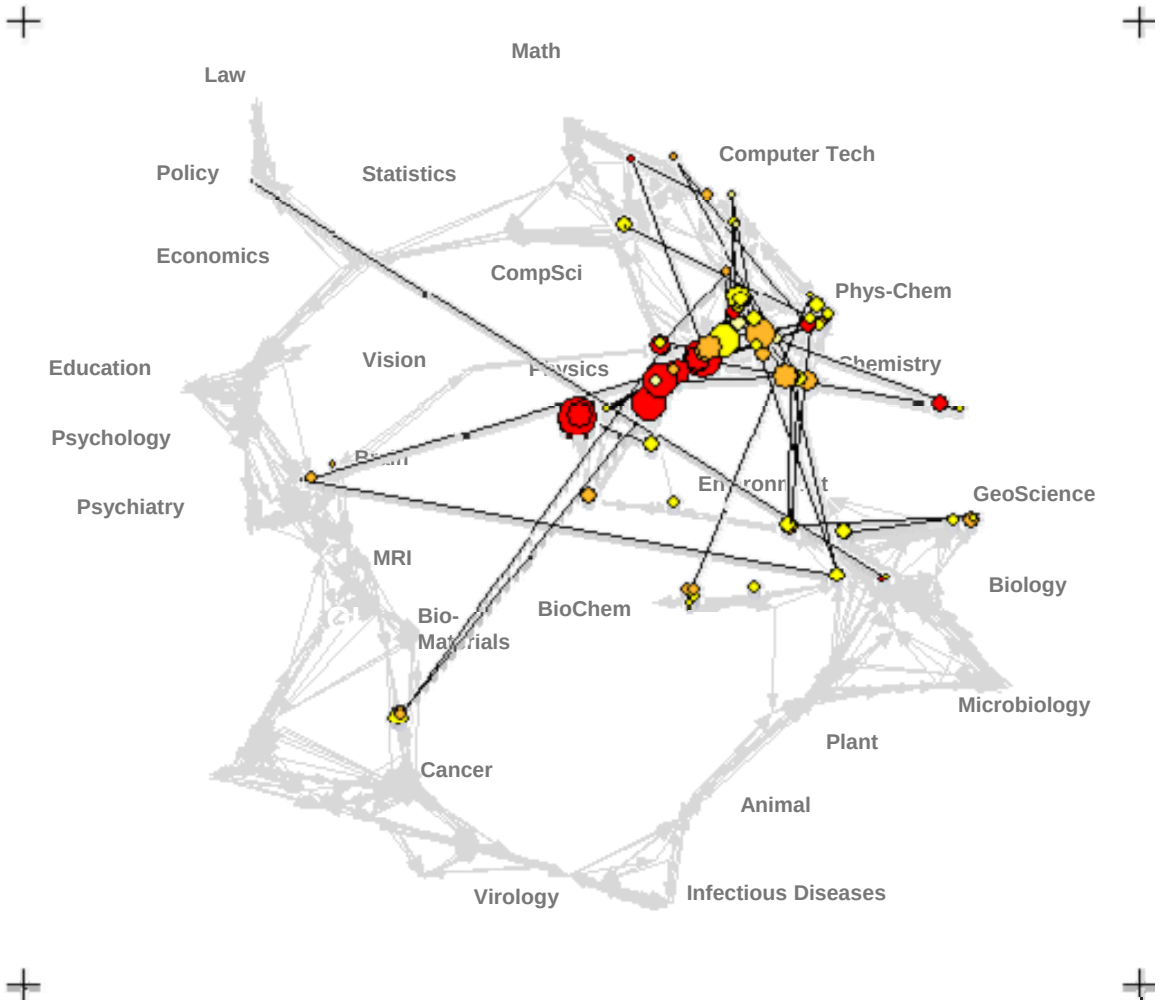
- Uses combined SCI/SSCI from 2002
  - 1.07M papers, 24.5M references, 7,300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts were reaggregated at the journal cluster level to calculate the
  - (x,y) positions for each journal cluster
  - by association, (x,y) positions for each journal



Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

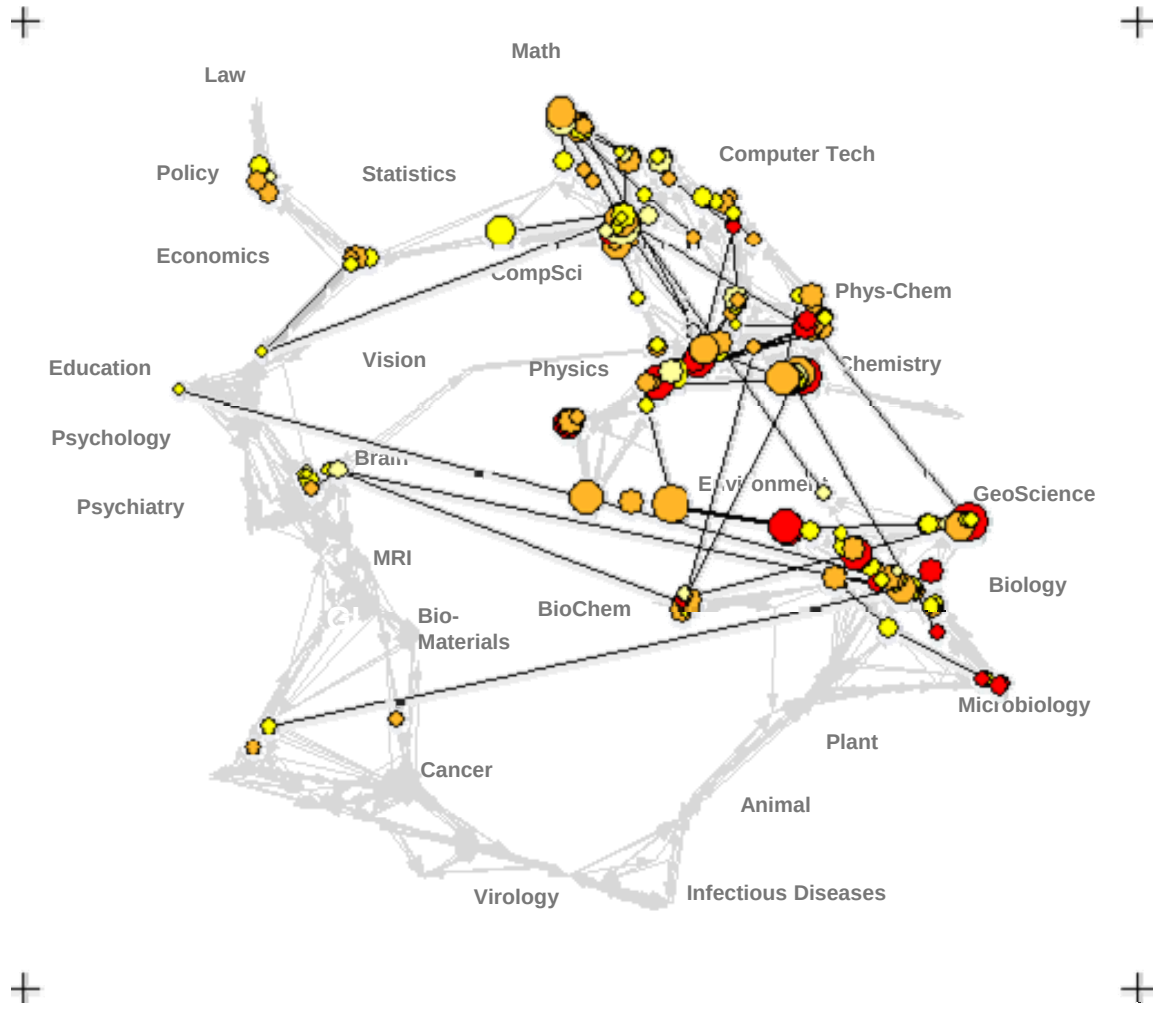
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

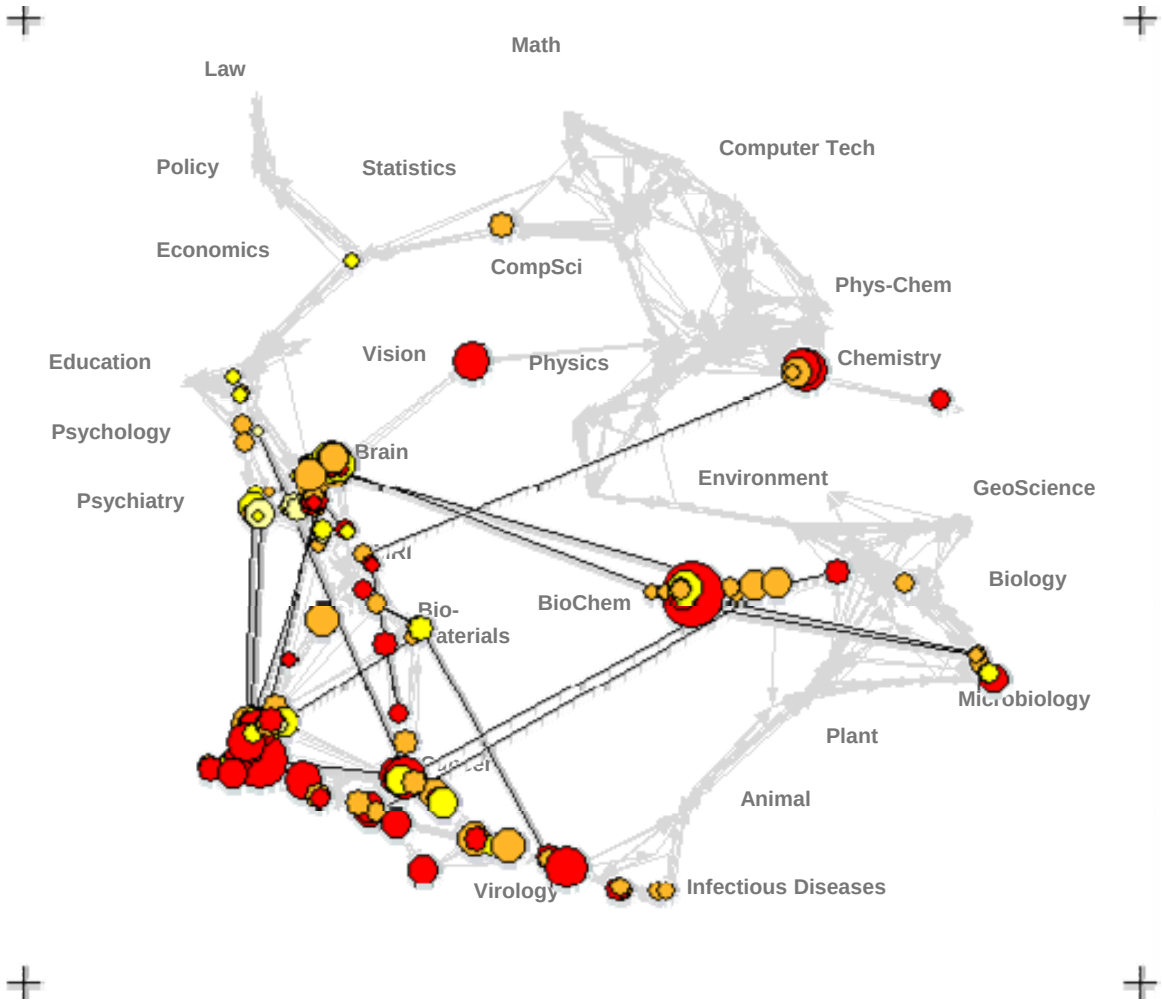
Funding Patterns of the National Science Foundation (NSF)



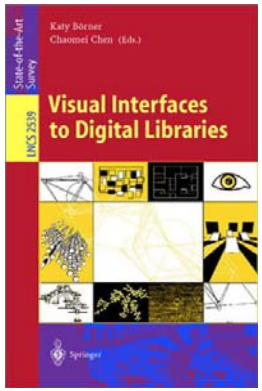
Science map applications: Identifying core competency

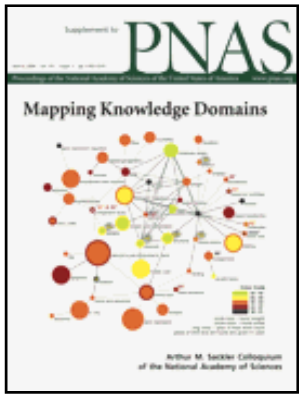
Kevin W. Boyack, Katy Börner, & Richard Klavans (2007).

Funding Patterns of the National Institutes of Health (NIH)



7. Standards and Practices





- Which ones exist?

Relevant Scholarly Reviews/Issues

- Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, Volume 37, Chapter 5, pp. 179-255. http://ivl.slis.indiana.edu/km/pub/2003-borner-arist.pdf
- Shiffrin, Richard M. and Börner, Katy (Eds.) (2004). Mapping Knowledge Domains. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl_1). http://www.pnas.org/content/vol101/suppl_1/
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). Network Science. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, Volume 41, Chapter 12, pp. 537-607. http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf

8. Dissemination of Results

Advantages for Funding Agencies

- Supports monitoring of (long-term) money flow and research developments, evaluation of funding strategies for different programs, decisions on project durations, funding patterns.
- Staff resources can be used for scientific program development, to identify areas for future development, and the stimulation of new research areas.

Advantages for Researchers

- Easy access to research results, relevant funding programs and their success rates, potential collaborators, competitors, related projects/publications **(research push).**
- More time for research and teaching.

Advantages for Industry

- Fast and easy access to major results, experts, etc.
- Can influence the direction of research by entering information on needed technologies **(industry-pull)**.

Advantages for Publishers

- Unique interface to their data.
- Publicly funded development of databases and their interlinkage.

For Society

- Dramatically improved access to scientific knowledge and expertise.

Places & Spaces: Mapping Science

a science exhibit that introduces people to maps of sciences, their makers and users.

http://scimaps.org.

Exhibit Curators: Dr. Katy Börner & Elisha F. Hardy

Mapping Science Exhibit – 10 Iterations in 10 years

The Power of Maps (2005)

The Power of Reference Systems (2006)

The Power of Forecasts (2007)

Science Maps for Economic Decision Makers (2008)

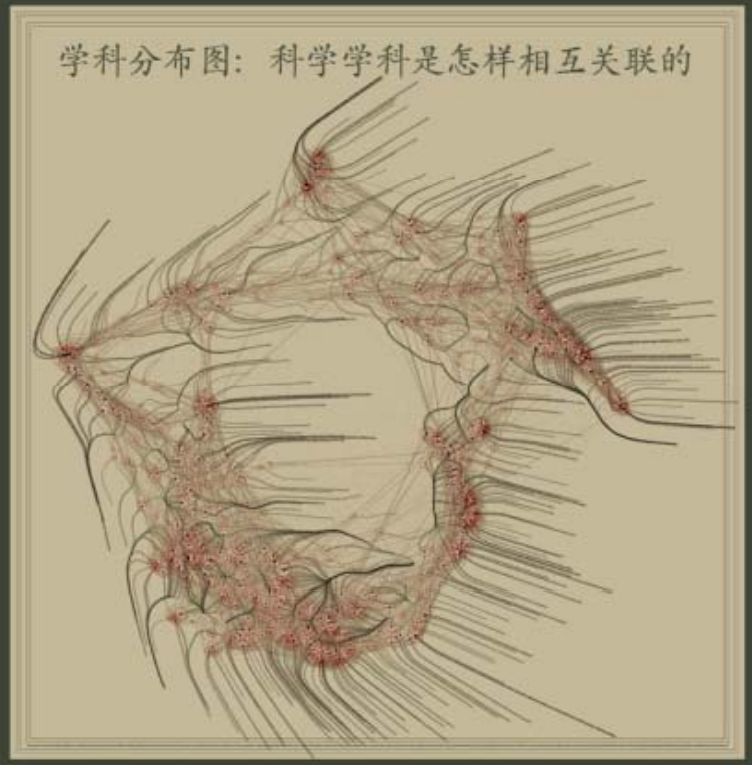
Science Maps for Science Policy Makers (2009)
Science Maps for Scholars (2010)
Science Maps as Visual Interfaces to Digital Libraries (2011)
Science Maps for Kids (2012)
Science Forecasts (2013)

How to Lie with Science Maps (2014)

So far, the exhibit has been shown in 49 venues on four continents. It is currently on display at
- National Science Foundation, 10th Floor, 4201 Wilson Boulevard, Arlington, VA, permanent display.
- National Research Council in Ottawa, Canada, April 3-Aug. 29, 2008.
- National Science Library of the Chinese Academy of Sciences, Beijing, China, May 17-Nov. 15, 2008.

Illuminated Diagram Display

W. Bradford Paley, Kevin W. Boyack, Richard Kalvans, and Katy Börner (2007) Mapping, Illuminating, and Interacting with Science. SIGGRAPH 2007, San Diego, CA.



你可以通过触摸屏在地图上随意指点来改变所到之处的光亮强度。当你触摸世界地图的某一点时，在那个地理位置上的所有研究机构会被点亮。同时在这些研究机构工作的学者的论文所属的学科会在学科分布图上被点亮。而当你触摸学科分布图的某一点时，在那个位置上的科学学科会被点亮。同时从事这些学科研究的研究机构在世界地图上的分布会被点亮。

纳米技术

这里显示所有和纳米技术相关的科学学科。纳米技术和科学研究人类在无形的空间里改造世界的能力。这些空间存在于极其微小以至单个原子的结构中。目前大部分有关纳米的研究主要集中在物理、化学和材料科学领域。它们主要位于学科分布图上半部分的右面。不过，纳米技术在生物学和医药学研究里的应用也越来越多。生物学和医药学位于学科分布图下半部分的右面。

探索科学学科的相互关联性

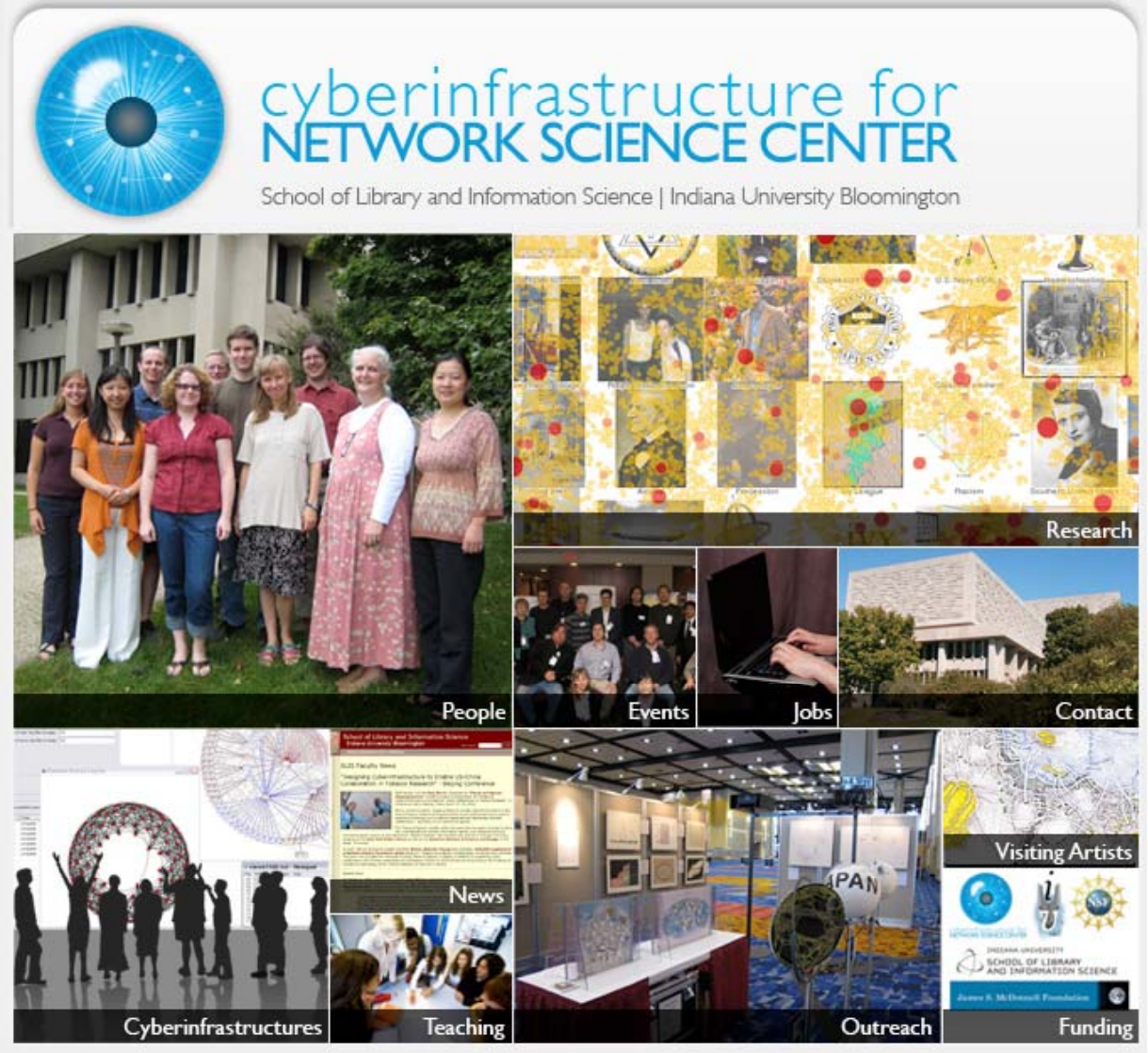
所有科学学科 显示所有776种科学学科	纳米技术 有关微观粒子的科学
可持续性 一些与人类寄予长期希望相关的科学	化学和生物 化学和生物科学的交叉部分

光柱缓慢的扫过所有相互关联的科学学科，每一个学科以及从事这方面科学研究的研究机构在世界地图上的位置会被逐一点亮。首先，显示屏会点亮那些产出论文最多，最活跃的科学学科。然后那些小学科或冷门学科会被逐一点亮。

探索某个学者的科学著作的影响力的传播

弗郎西.科里克 DNA双螺旋状的发现者之一	阿尔伯特.爱因斯坦 用相对论重新激活了物理学	迈克尔.费舍尔 发现了物质转变模式的关键步骤	苏珊.费斯克 研究人的认知是如何产生偏见的
约舒亚.雷德伯格 细菌遗传机制研究的先驱	德里克.德索拉.普里斯 著名的“科学计量学之父”	理查德.扎尔 采用激光化学技术研究分子动态分布	关于本次展览 与此展览相关人员和机构

显示屏通过四步来展示某个学者对科学的贡献以及影响力的传播。首先，显示屏点亮该学者所发表的论文所属的学科在学科分布图上的位置以及该学者从事这项研究时所在的研究机构在世界地图上的位置。到目前为止，所有这些论文的引用率仍然很高。第二步，显示屏点亮所有引用在第一步中被点亮的原始论文的论文在学科分布图上的位置以及它们在世界地图上的位置。第三步，显示屏点亮所有引用了在第二步中被点亮的论文的学科在学科分布图上的位置以及它们在世界地图上的位置。第四步，显示屏点亮所有引用了在第三步中被点亮的论文的学科在学科分布图上的位置以及它们在世界地图上的位置。



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