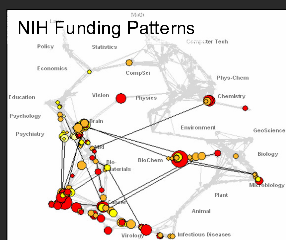
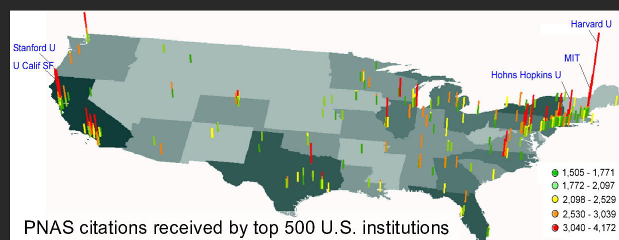
Towards Effective KM Tools

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

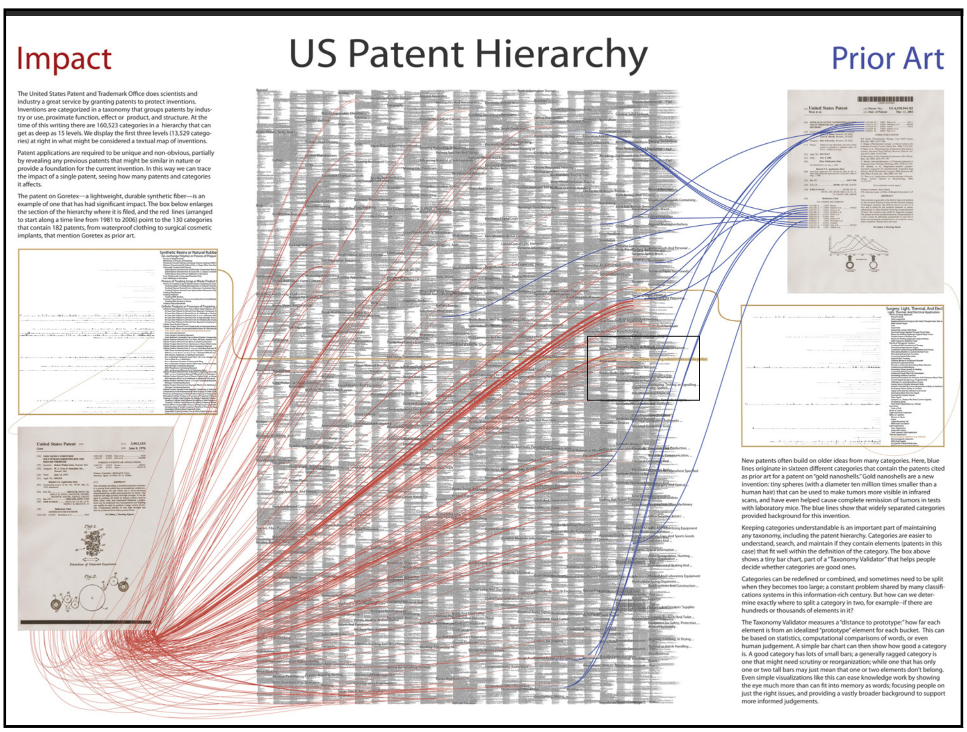
KEBR, December 11-12, 2006

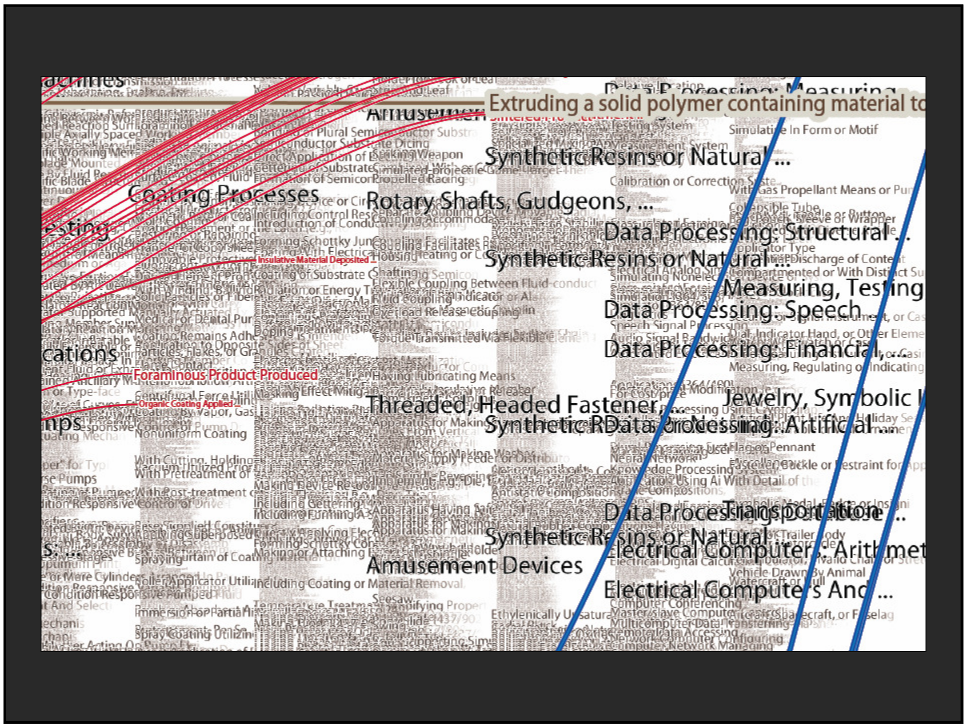


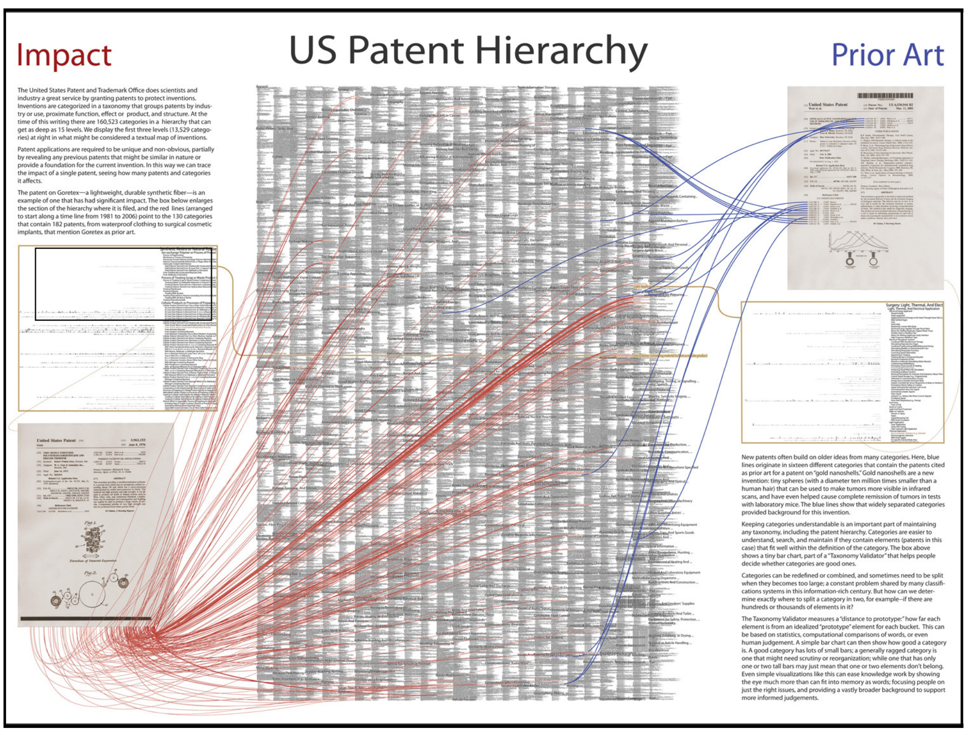


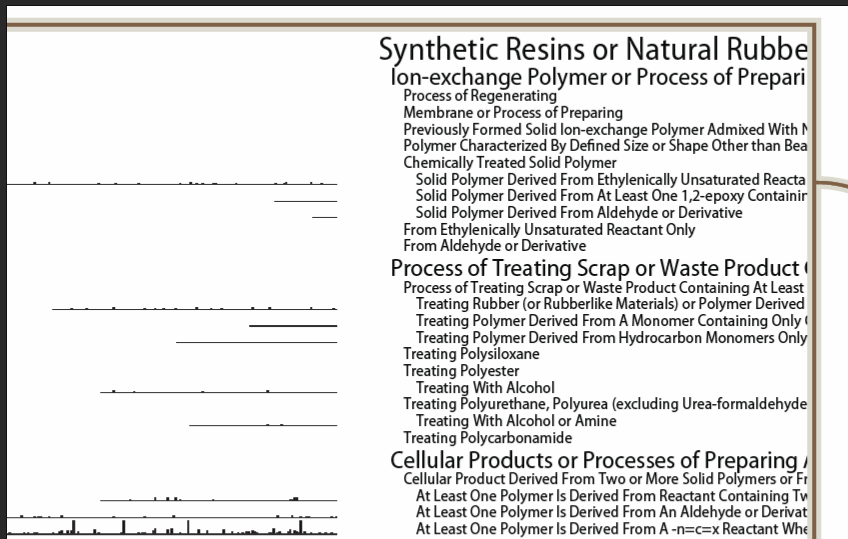
- **Taxonomies / Ontologies / Classification Hierarchies**
- **Mapping Knowledge Spaces such as Science**
- **Building Data/Software/Resources Bazaars not Cathedrals**

Taxonomies / Ontologies / Classification Hierarchies

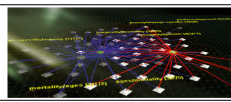



Extruding a solid polymer containing material to
Coating Processes
Rotary Shafts, Gudgeons, ...
Synthetic Resins or Natural ...
Data Processing: Structural ...
Measuring, Testing
Data Processing: Speech
Data Processing: Financial, ...
Foraminous Product Produced
Threaded, Headed Fastener
Jewelry, Symbolic
Amusement Devices
Electrical Computers: Arithmet
Electrical Computers And ...


Impact
US Patent Hierarchy
Prior Art
The United States Patent and Trademark Office does scientists and industry a great service by granting patents to protect inventions. Inventions are categorized in a taxonomy that groups patents by industry or use, proximate function, effect or product, and structure. At the time of this writing there are 160,523 categories in a hierarchy that can get as deep as 15 levels. We display the first three levels (13,529 categories) at right in what might be considered a textual map of inventions.
Patent applications are required to be unique and non-obvious, partially by revealing any previous patents that might be similar in nature or provide a foundation for the current invention. In this way we can trace the impact of a single patent, seeing how many patents and categories it affects.
The patent on Goretex—a lightweight, durable synthetic fiber—is an example of one that has had significant impact. The box below enlarges the section of the hierarchy where it is filed, and the red lines (arranged to start along a time line from 1981 to 2006) point to the 130 categories that contain 182 patents, from waterproof clothing to surgical cosmetic implants, that mention Goretex as prior art.
New patents often build on older ideas from many categories. Here, blue lines originate in sixteen different categories that contain the patents cited as prior art for a patent on "gold nanoshells." Gold nanoshells are a new invention: tiny spheres (with a diameter ten million times smaller than a human hair) that can be used to make tumors more visible in infrared scans, and have even helped cause complete remission of tumors in tests with laboratory mice. The blue lines show that widely separated categories provided background for this invention.
Keeping categories understandable is an important part of maintaining any taxonomy, including the patent hierarchy. Categories are easier to understand, search, and maintain if they contain elements (patents in this case) that fit well within the definition of the category. The box above shows a tiny bar chart, part of a "Taxonomy Validator" that helps people decide whether categories are good ones.
Categories can be redefined or combined, and sometimes need to be split when they becomes too large; a constant problem shared by many classifications systems in this information-rich century. But how can we determine exactly where to split a category in two, for example--if there are hundreds or thousands of elements in it?
The Taxonomy Validator measures a "distance to prototype:" how far each element is from an idealized "prototype" element for each bucket. This can be based on statistics, computational comparisons of words, or even human judgement. A simple bar chart can then show how good a category is. A good category has lots of small bars; a generally ragged category is one that might need scrutiny or reorganization; while one that has only one or two tall bars may just mean that one or two elements don't belong. Even simple visualizations like this can ease knowledge work by showing the eye much more than can fit into memory as words; focusing people on just the right issues, and providing a vastly broader background to support more informed judgements.

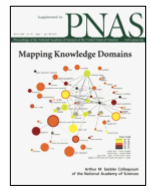


Mapping Knowledge Spaces such as Science



Process of Mapping Knowledge Domains

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

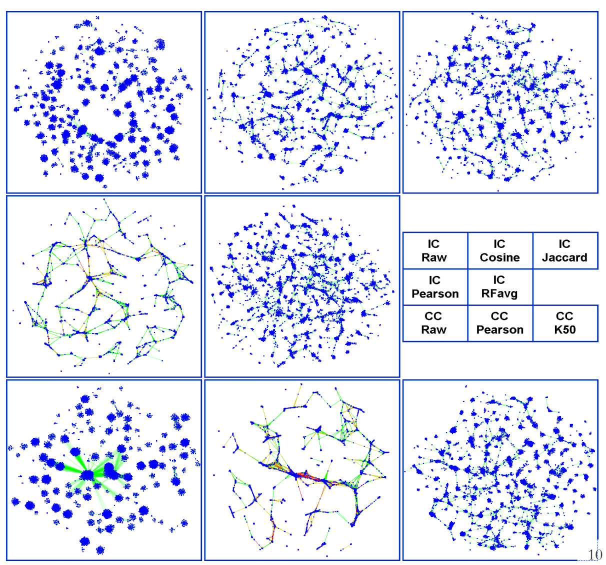


Börner, Chen & Boyack.. (2003) Visualizing Knowledge Domains. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Volume 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.

9

Comparison of Similarity Metrics

- ISI file year 2000, SCI and SSCI: 7,121 journals.
- Different similarity metrics
  - Inter-citation (raw counts, cosine, modified cosine, Jaccard, RF, Pearson)
  - Co-citation (raw counts, cosine, modified cosine, Pearson)
- Maps were compared based on
  - regional accuracy,
  - the scalability of the similarity algorithm, and
  - the readability of the layouts.

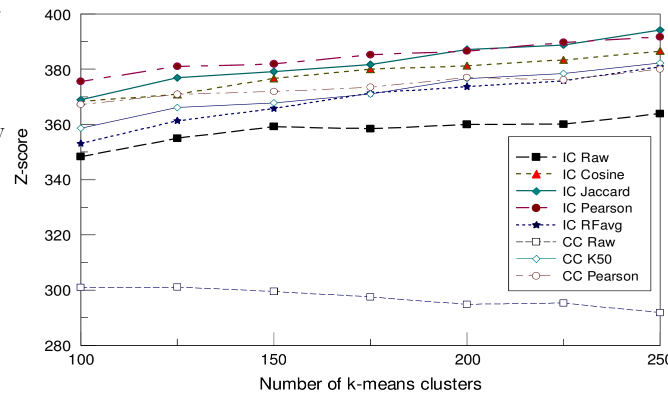


Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. Scientometrics. 64(3), 351-374.

10

Selecting the similarity measure with the best regional accuracy

- For each similarity measure, the VxOrd layout was subjected to k-means clustering using different numbers of clusters.
- Resulting cluster/category memberships were compared to actual category memberships using entropy/mutual information method by Gibbons & Roth, 2002.
- Increasing Z-score indicates increasing distance from a random solution.
- Most similarity measures are within several percent of each other.

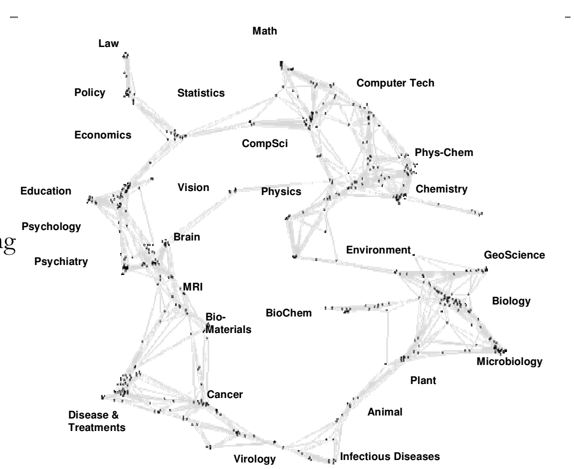


Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. Scientometrics. 64(3), 351-374.

Latest 'Base Map' of Science

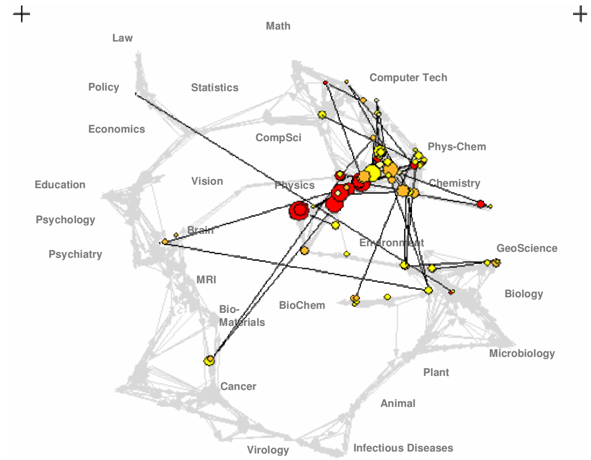
Kevin W. Boyack & Richard Klavans, unpublished work.

- 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



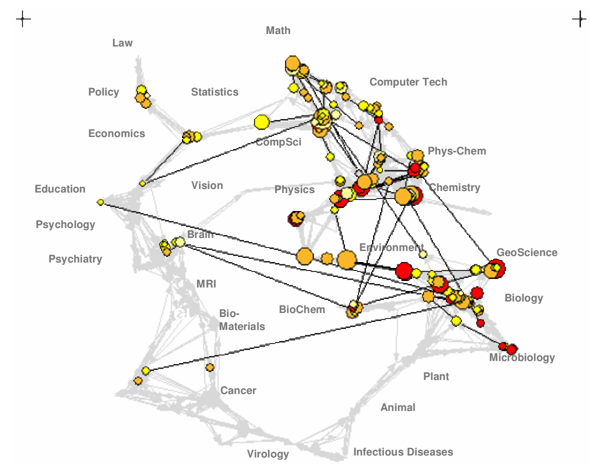
Science map applications: Identifying core competency
Kevin W. Boyack & Richard Klavans, unpublished work.

Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency
Kevin W. Boyack & Richard Klavans, unpublished work.

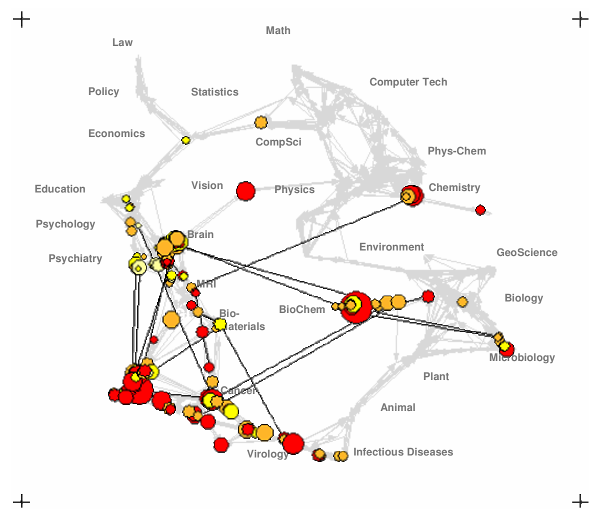
Funding Patterns of the National Science Foundation (NSF)

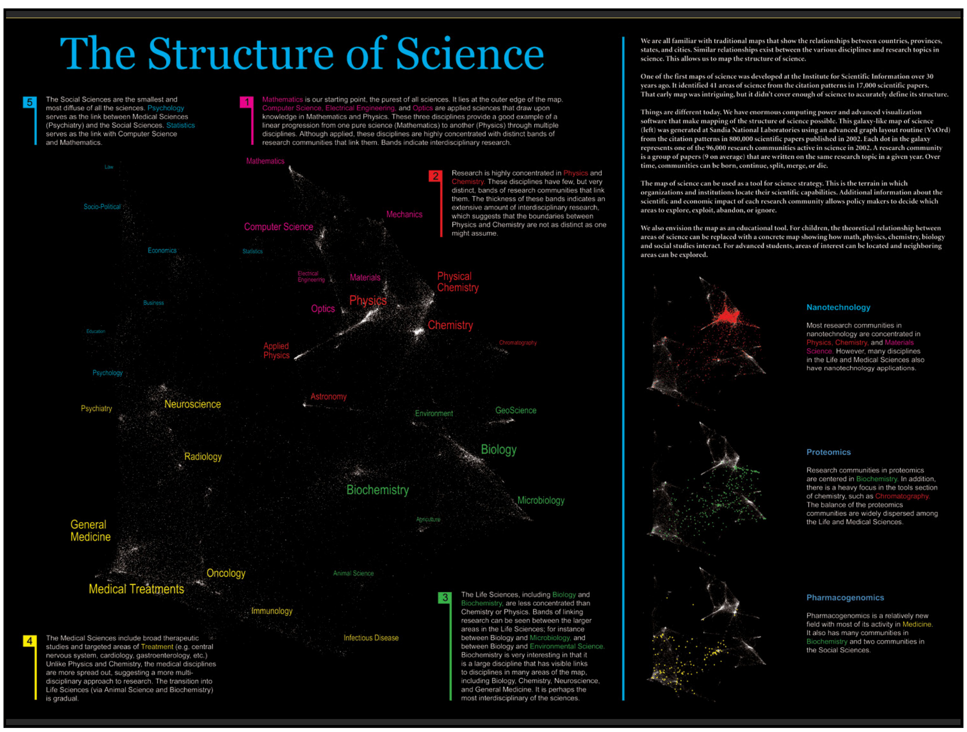


Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

Funding Patterns of the National Institutes of Health (NIH)

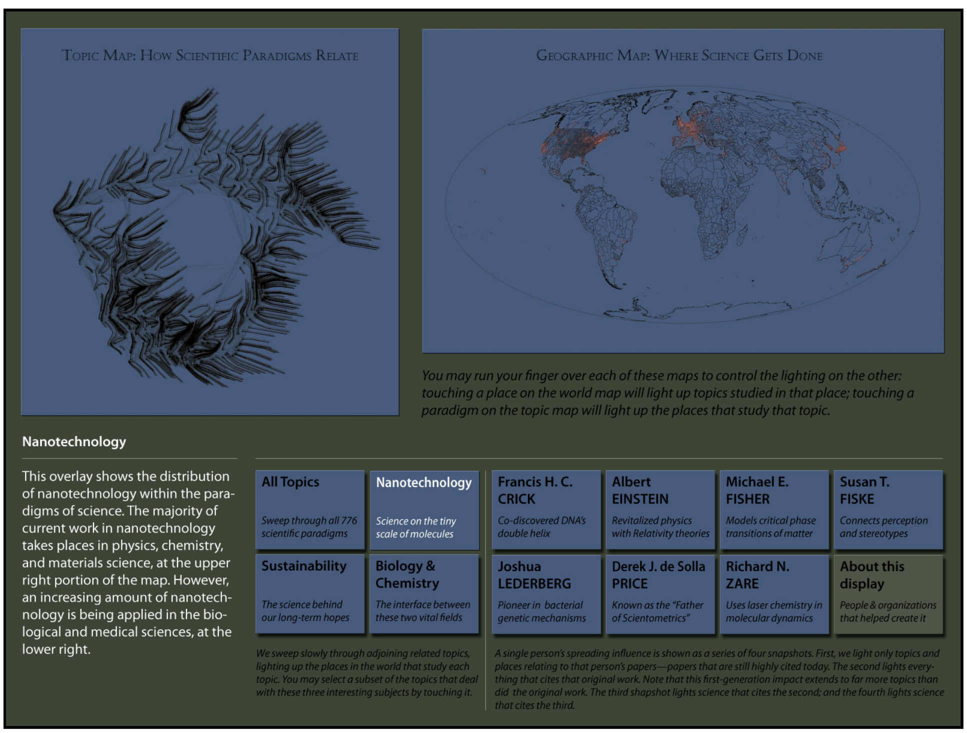


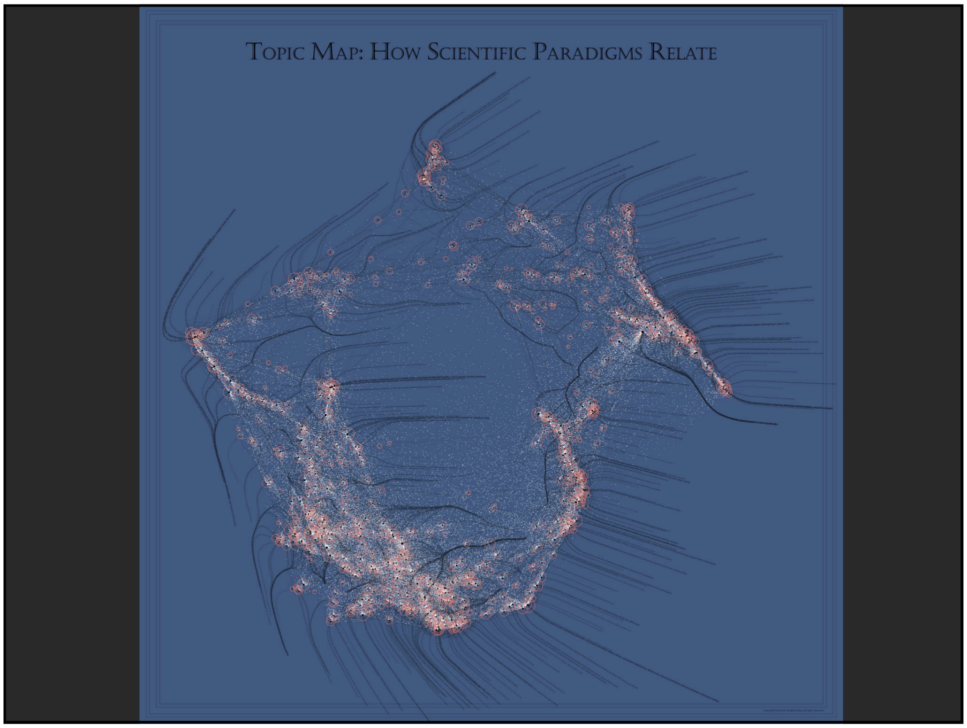


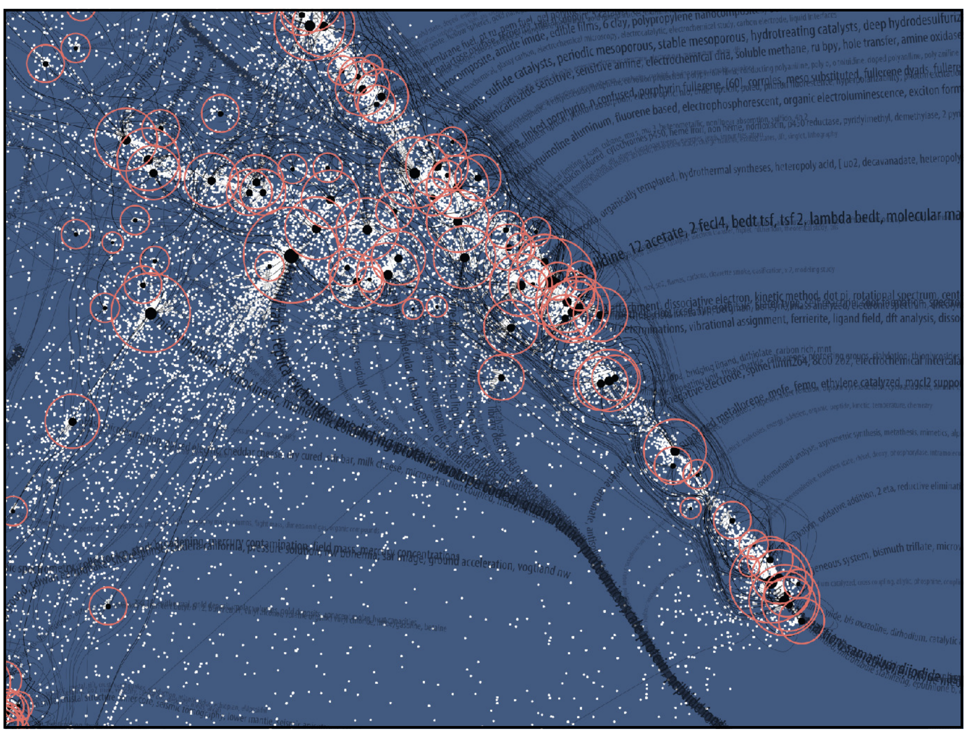
"Places & Spaces: Mapping Science" Illuminated Diagram Display

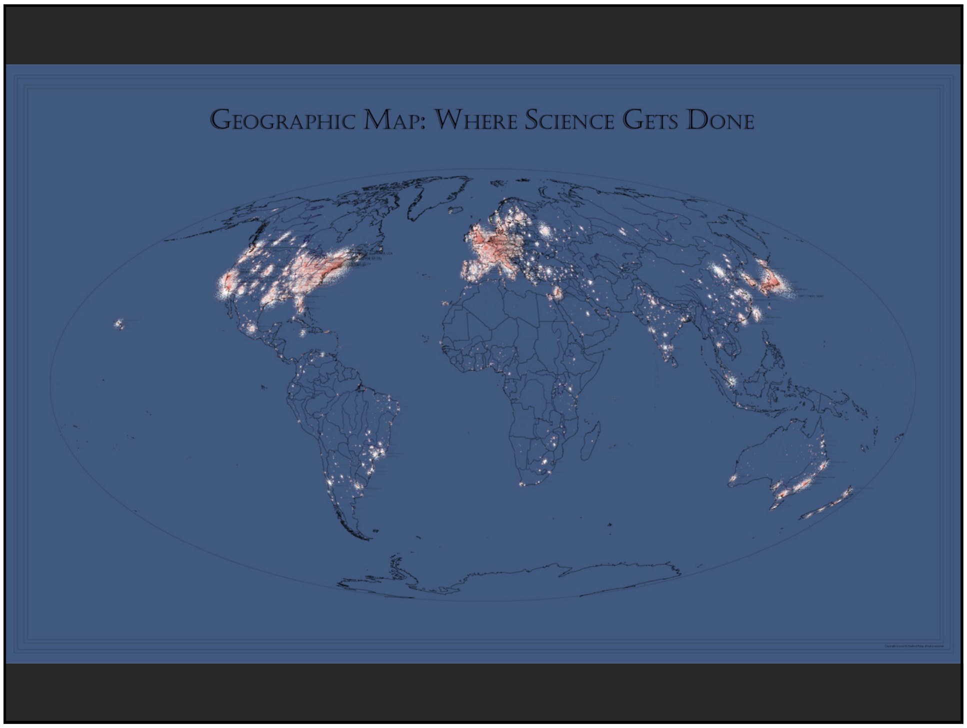
Was on display at the NYPL Science, Industry, and Business Library
Madison/34th, New York City, April 3rd - August 31st, 2006.

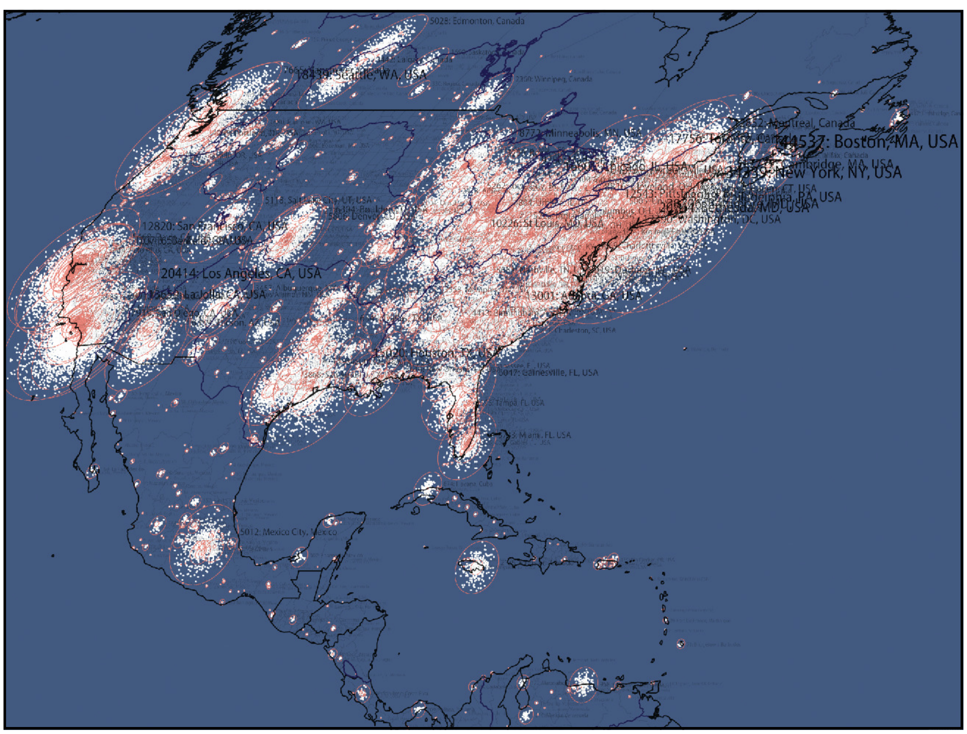
17

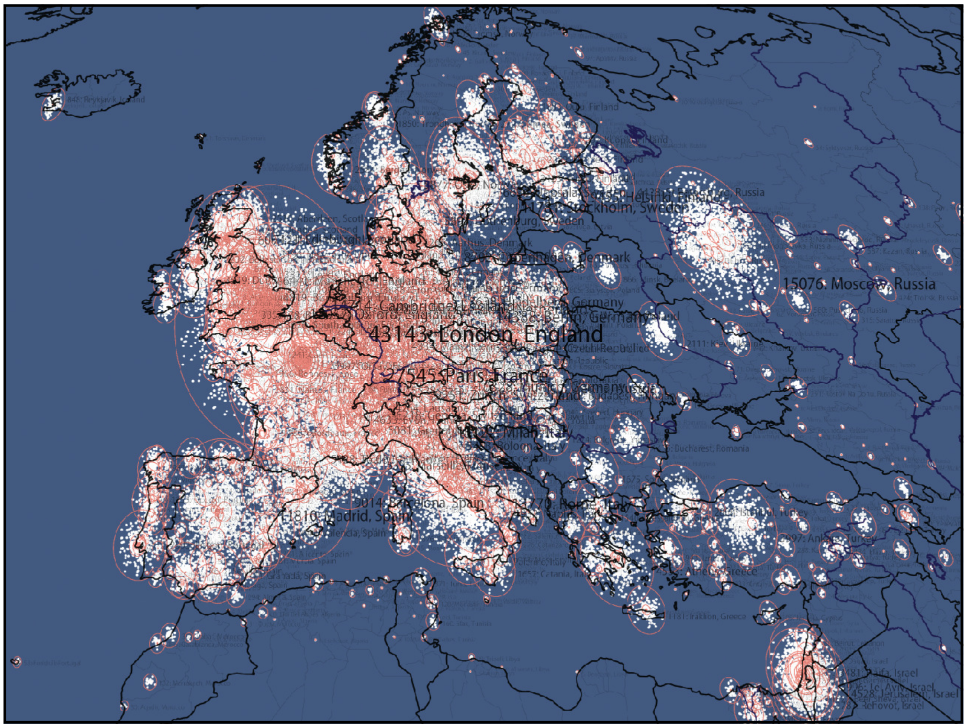


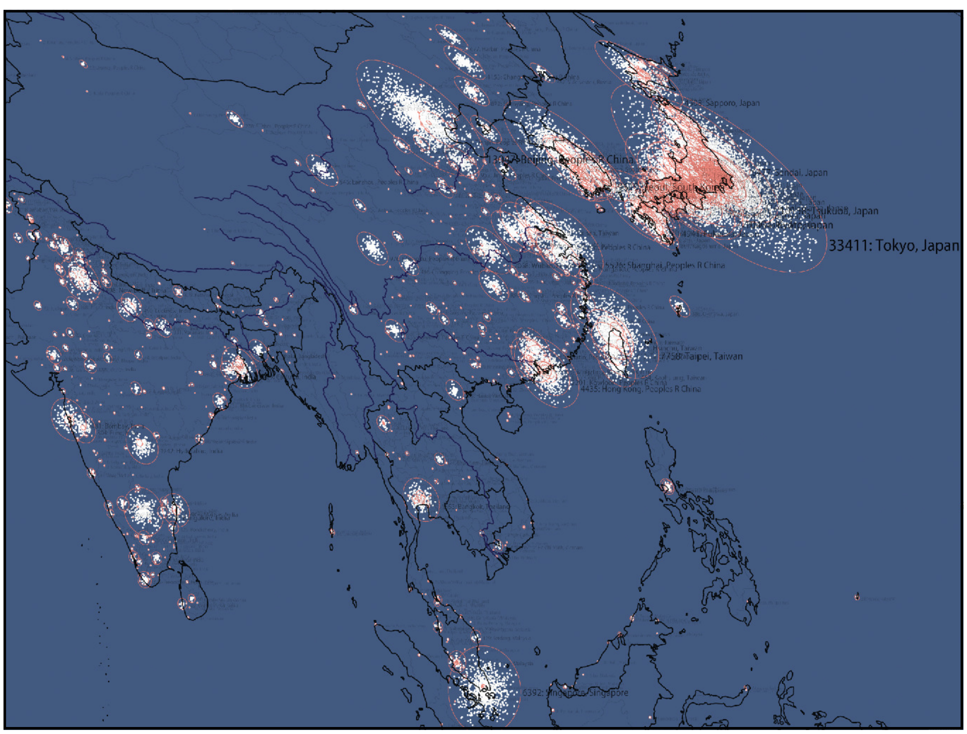

Topic Map: How Scientific Paradigms Relate


organically templated, hydrothermal syntheses, heteropoly acid, [uo2, decavanadate, heteropoly
12 acetate, 2 fecl4, bedt ttf, ttf 2, lambda bedt, molecular ma


Geographic Map: Where Science Gets Done


5028: Edmonton, Canada
44537: Boston, MA, USA
20414: Los Angeles, CA, USA
5012: Mexico City, Mexico


15076: Moscow, Russia
43143: London, England


33411: Tokyo, Japan

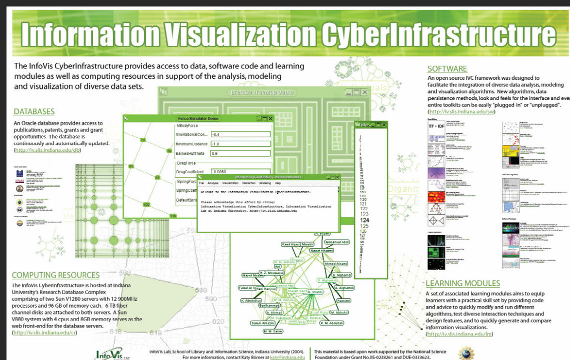
Science Puzzle Map for Kids by Fileve Palmer, Julie Smith, Elisha Hardy and Katy Börner, Indiana University, 2006.
(Base map taken from Illuminated Diagram display by Kevin Boyack, Richard Klavans, and W. Bradford Paley.)

NY HALL OF SCIENCE

For a New Take on Maps Chart Your Way to the New York Hall of Science
Places & Spaces: Mapping Science on view December 9, 2006 – February 25, 2007

Queens, N. Y. – Want to see science from above? Curious to see what impact one single person or invention can have? Keen to find pockets of innovation? Desperate for better tools to manage the flood of information? Or are you simply fascinated by maps? Then visit the *Places & Spaces: Mapping Science* exhibition, which aims to demonstrate the power of maps to navigate and make sense of physical places and abstract topic spaces.

Building Data/Software/Resources Bazaars not Cathedrals



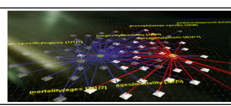
Cyberinfrastructure Shell
http://cishell.org

CAREER: Visualizing Knowledge Domains. NSF IIS-0238261 award (Katy Börner, $451,000) Sept. 03-Aug. 08.
http://iv.slis.indiana.edu/

NetworkWorkbench
A Workbench for Network Scientists

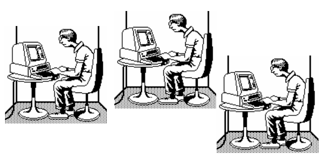
SEI: Network Workbench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. NSF IIS-0513650 award (Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert (Senior Personnel), $1,120,926) Sept. 05 - Aug. 08. http://nwb.slis.indiana.edu

29



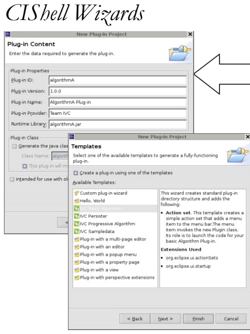
Serve Algorithms Developers & Users

Developers



Users

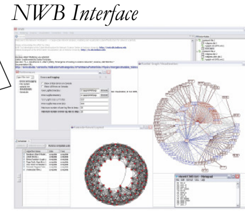
CIShell Wizards



CIShell

IVC Interface

NWB Interface



30