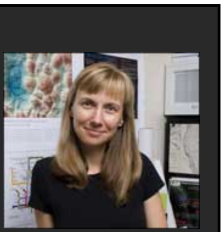
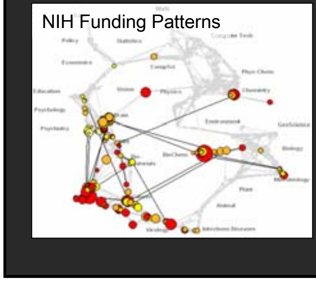
Mapping the Structure and Evolution of Scholarly Knowledge:

Data (Integration) Issues

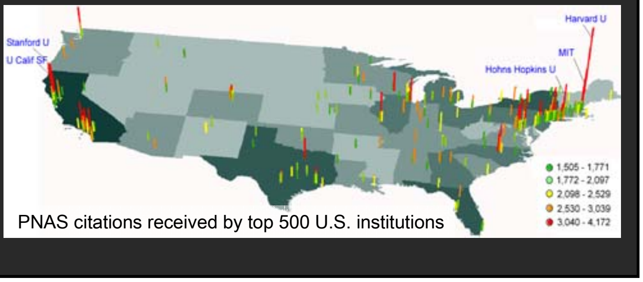


Dr. Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu

Workshop on Scholarly Databases & Data Integration
Indiana University, Bloomington, IN, Aug. 29 & 30, 2006



NIH Funding Patterns



PNAS citations received by top 500 U.S. institutions

1. Dream Tools for Scholarly Knowledge Management

2. Challenges
3. Opportunities

1. **Dream Tools for Scholarly Knowledge Management**

2. Challenges
3. Opportunities

Dream Tools for Scholarly Knowledge Management

Tools we developed for ourselves and our clients

Information Visualization Laboratory Management System

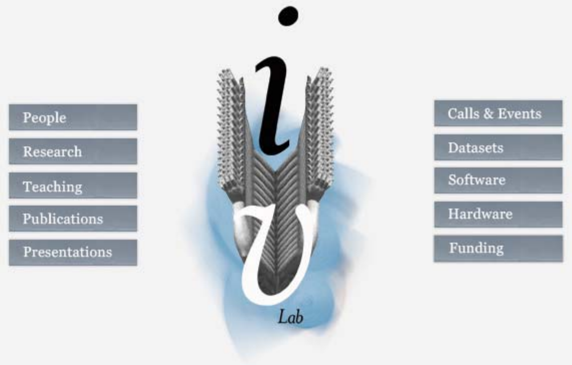
Taxonomy Visualization/ Validation System

Coupling Geospatial and Topic Space

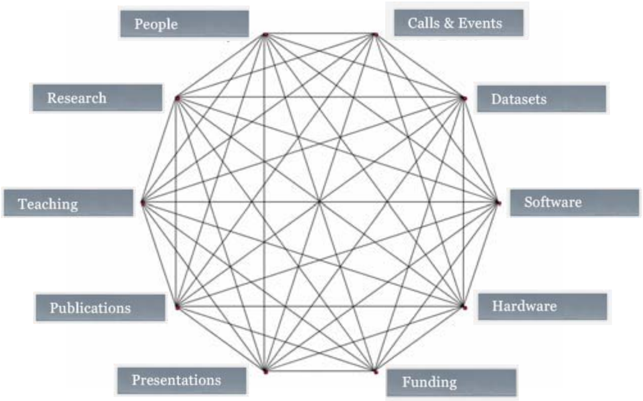
W. Bradford Paley
Scientific Mapmaker
Digital Image Design Incorporated
Dept. of Computer Science, Columbia University

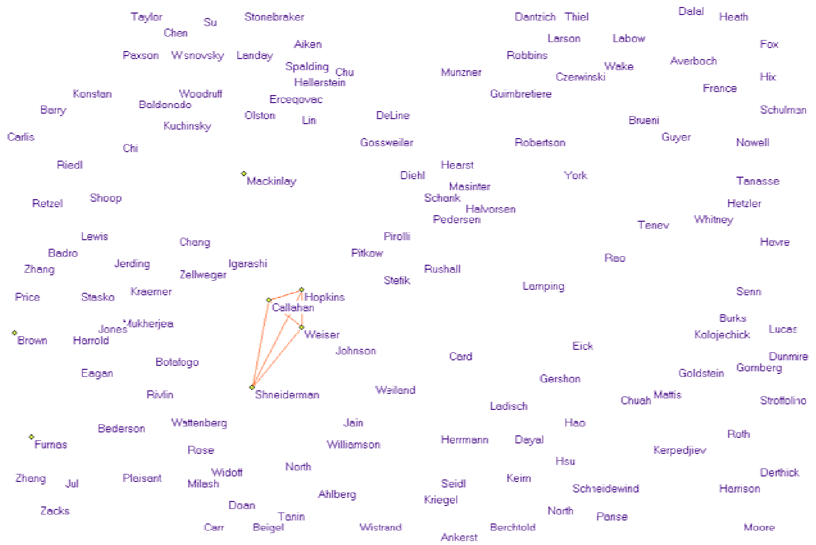
Science Maps for Kids

Information Visualization Laboratory
Management System



https://ivl.slis.indiana.edu





Legend

Nodes ~ Authors
Node area size ~ Number of papers publishad
Node color ~ Number of citations

Edges ~ Co-authorship relations
Edge color ~ Year of first co-authorship

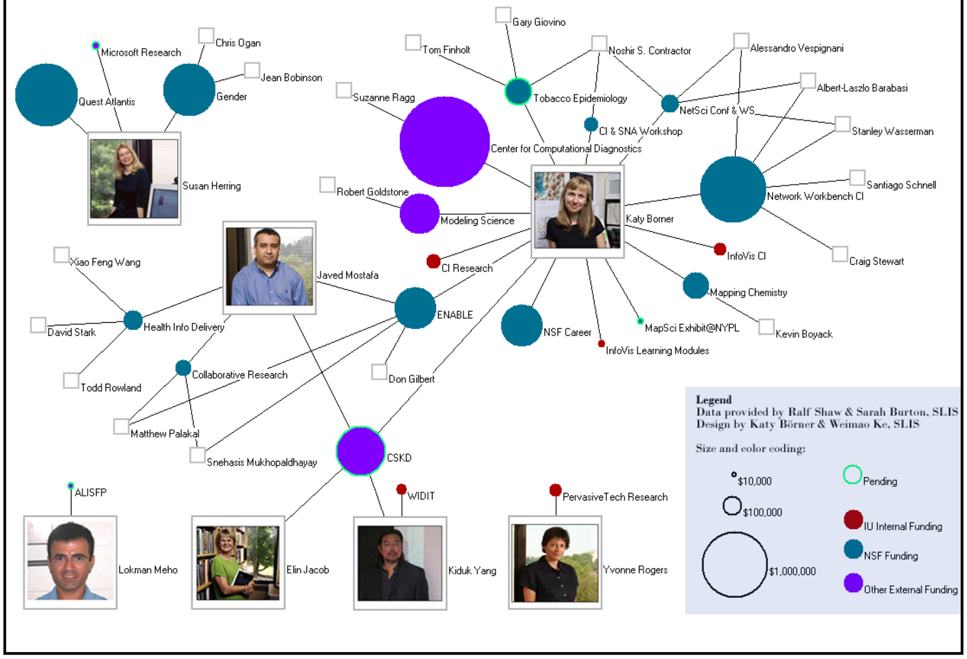
Node Color Code	
0 - 9	🟨
10 - 19	🟨
20 - 29	🟧
30 - 39	🟧
40 - 49	🟥
50 -	🟥

Edge Color Code	
86 - 90	🟧
91 - 95	🟩
96 - 00	🟩
01 -	🟩

Displayed Year: 1988

Mapping the Evolution of Co-Authorship Networks
Weimao Ke, Lalitha Visvanath & Katy Börner
InfoVis Lab @ Indiana University
2004

Awarded & Pending Funding of SLIS Faculty Members in January 2006



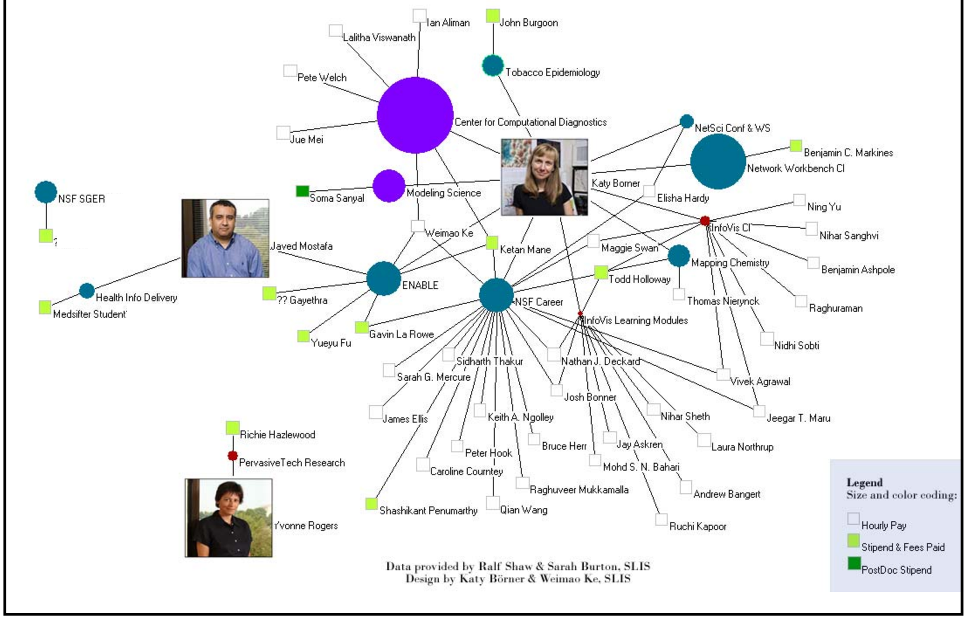
Legend
Data provided by Ralf Shaw & Sarah Burton, SLIS
Design by Katy Börner & Weimao Ke, SLIS

Size and color coding:

○ $10,000
○ $100,000
○ $1,000,000

○ Pending
● IU Internal Funding
● NSF Funding
● Other External Funding

Student Support by SLIS Faculty Members in January 2006

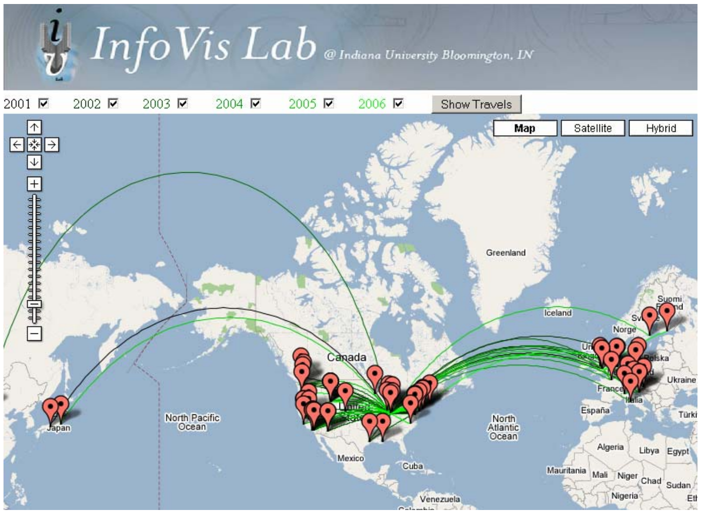


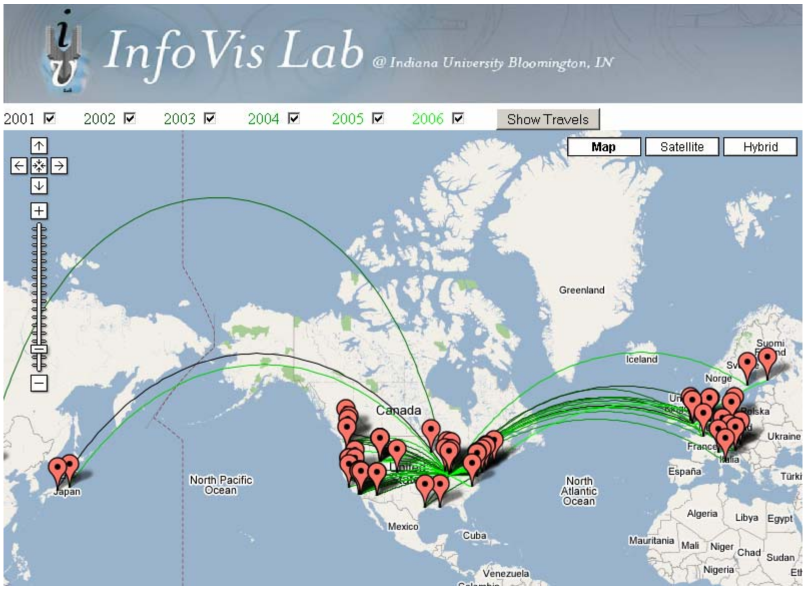
Legend
Size and color coding:

□ Hourly Pay
🟩 Stipend & Fees Paid
🟩 PostDoc Stipend

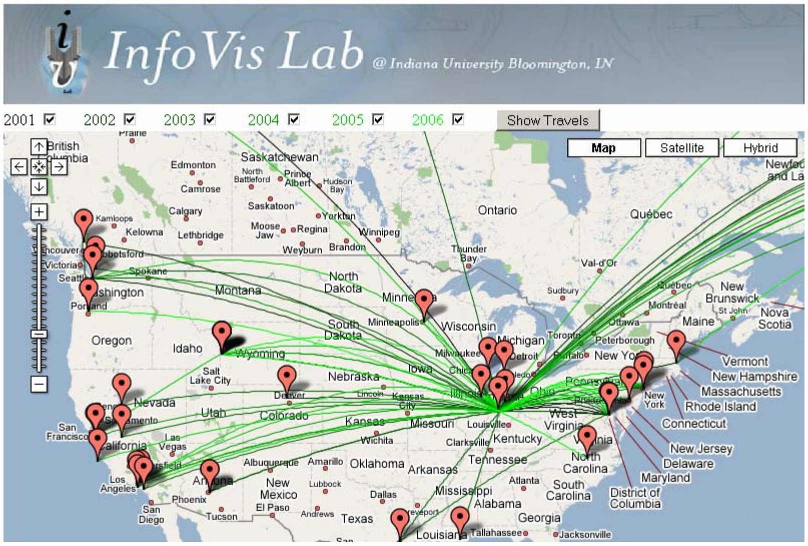
Data provided by Ralf Shaw & Sarah Burton, SLIS
Design by Katy Börner & Weimao Ke, SLIS

KM: Katy's Travels in 2000-2006

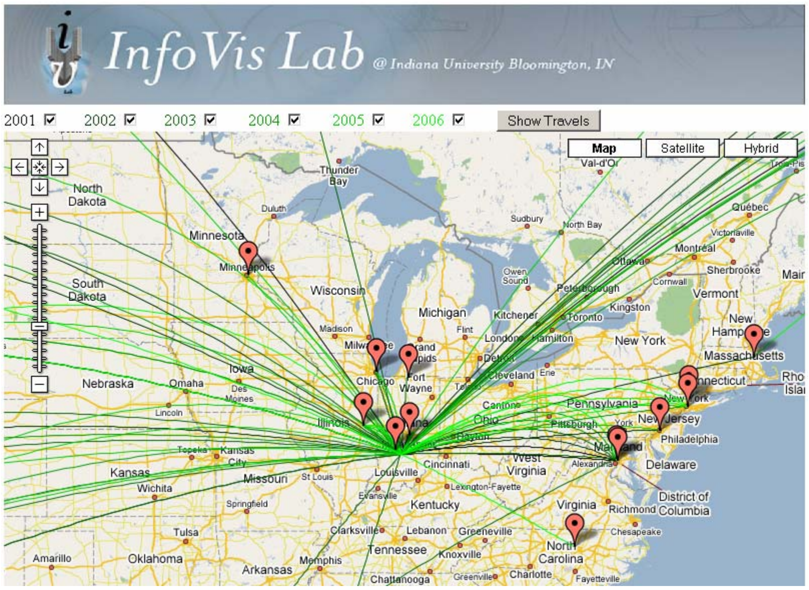




By Thomas Neirynck, 2007.

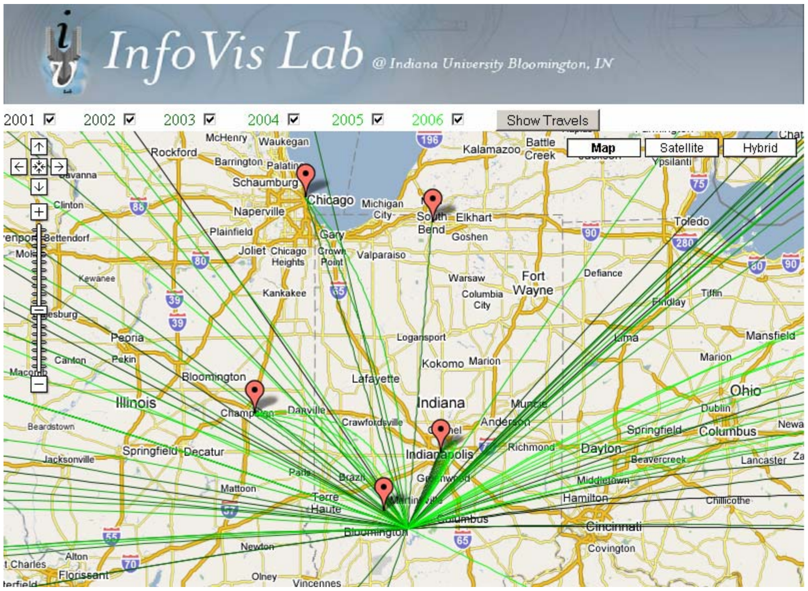


By Thomas Neirynck, 2007.



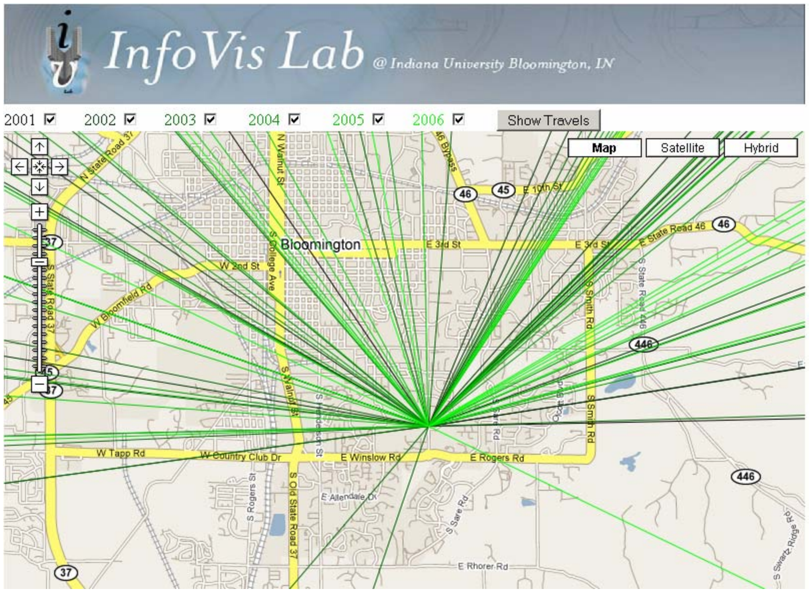
2001 ☑ 2002 ☑ 2003 ☑ 2004 ☑ 2005 ☑ 2006 ☑

By Thomas Neirynck, 2007.



2001 ☑ 2002 ☑ 2003 ☑ 2004 ☑ 2005 ☑ 2006 ☑

By Thomas Neirynck, 2007.



2001 ☑ 2002 ☑ 2003 ☑ 2004 ☑ 2005 ☑ 2006 ☑ Show Travels

By Thomas Neirynck, 2007.

MARC records, Wiki authors ↕ People

Science in the City ↕ Calls & Events

Courses at IU ↕ Research

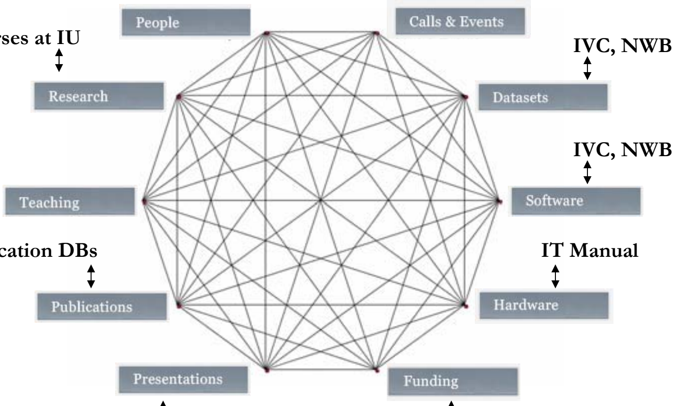
IVC, NWB data ↕ Datasets

IVC, NWB code ↕ Software

Teaching

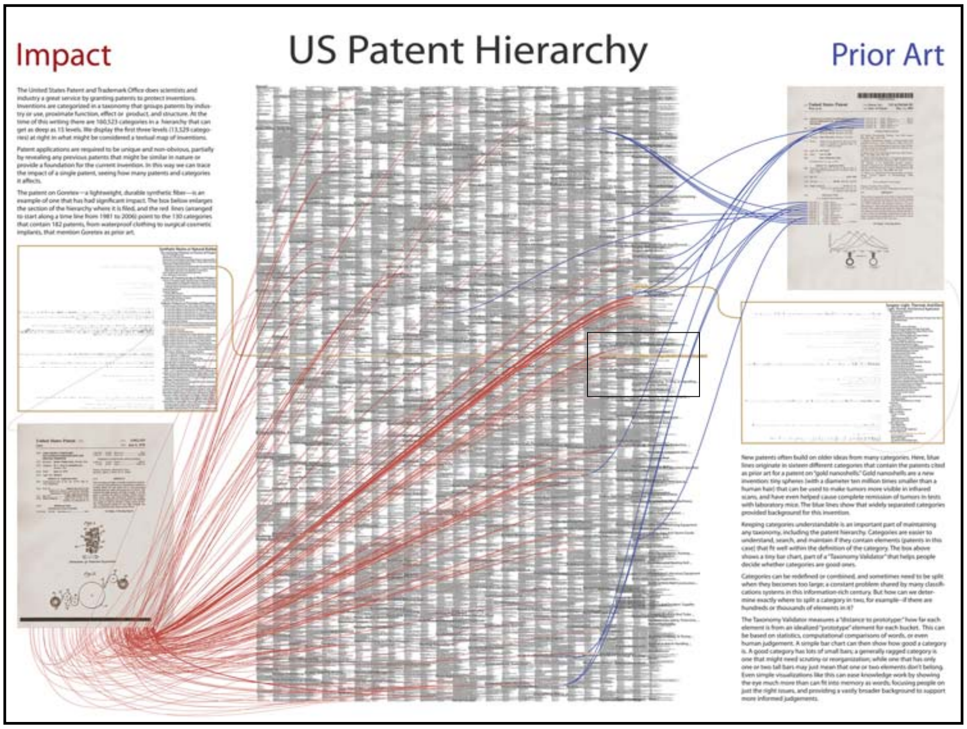
Publication DBs ↕ Publications

IT Manual ↕ Hardware



Presentations ↕ **NetSci, IM2 Conference talks**

Funding ↕ **NSF, NIH awards data, IU awards**



US Patent Hierarchy

Impact

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.

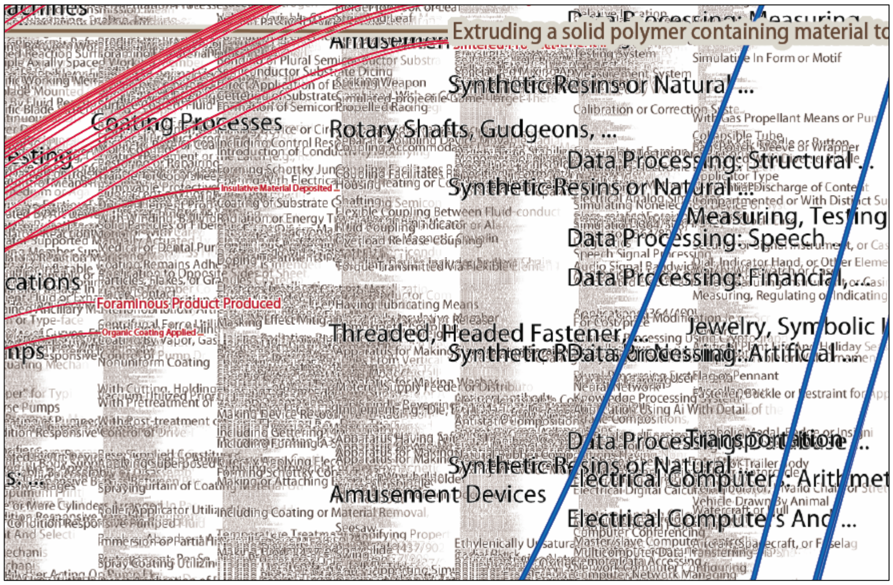
Prior Art

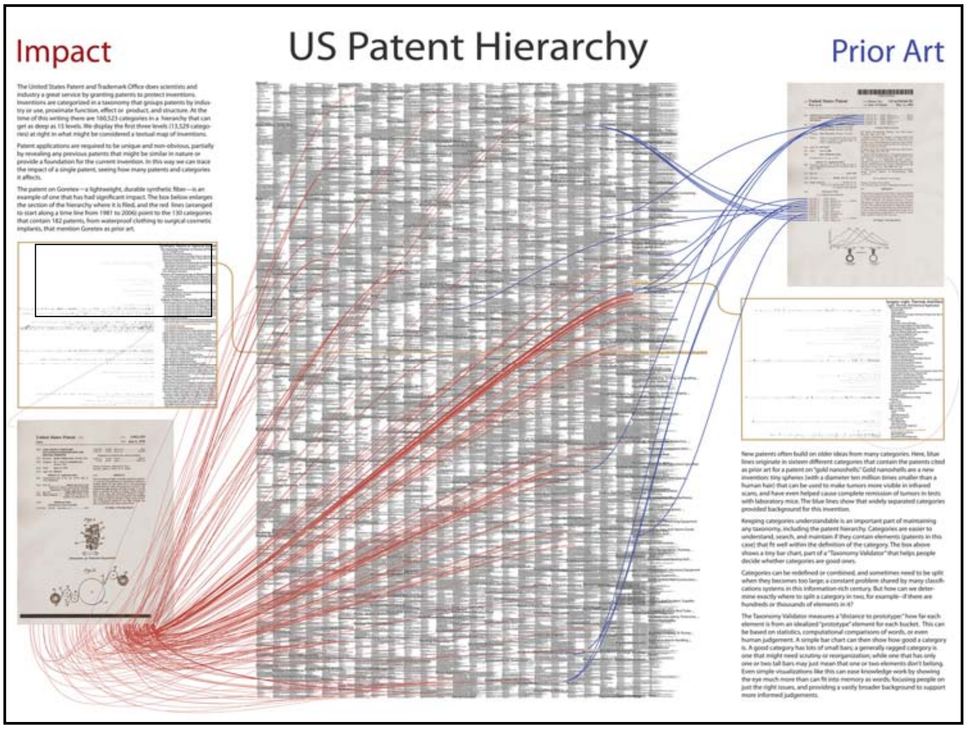
New patents often build on older ideas from many categories. Here, blue lines originate in sixteen different categories that contain the patents cited as prior art for a patent on "gold nanoshells." Gold nanoshells are a new invention: tiny spheres (with a diameter ten million times smaller than a human 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.


Extruding a solid polymer containing material to
Synthetic Resins or Natural ...
Coating Processes
Rotary Shafts, Gudgeons, ...
Data Processing: Structural ...
Synthetic Resins or Natural ...
Measuring, Testing
Data Processing: Speech
Data Processing: Financial, ...
Foraminous Product Produced
Organic Coating Applied
Insulative Material Deposited
Threaded, Headed Fastener, ...
Jewelry, Symbolic
Synthetic R
Data Processing: Artificial
Data Processing
Transportation
Synthetic Resins or Natural
Electrical Computers: Arithmet
Amusement Devices
Electrical Computers And ...
Simulating In Form or Motif
Calibration or Correction System
With Gas Propellant Means or Pur
Collapsible Tube
Discharge of Content
Compartmented or With Distinct Su
Speech Signal Processing
Audio Signal Bandwidth
Measuring, Regulating or Indicating
Having Lubricating Means
Flexible Coupling Between Fluid-conduct
Nonuniform Coating
With Cutting, Holding
With Pretreatment of
With Post-treatment
Spraying
Including Coating or Material Removal
Seesaw
Ethylenically Unsaturated
Vehicle Drawn By Animal
Watercraft or Hull
Trailer Body
Computer Conferencing
Multicomputer Data Transferring
Neural Network
Knowledge Processing
Buckle or Restraint for
Pennant
Electrical-Digital Calculating Computer
Game Target
Projectile Game
Propelled Racing
Plural Semiconductor Substrates
Dicing



US Patent Hierarchy

Impact

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.

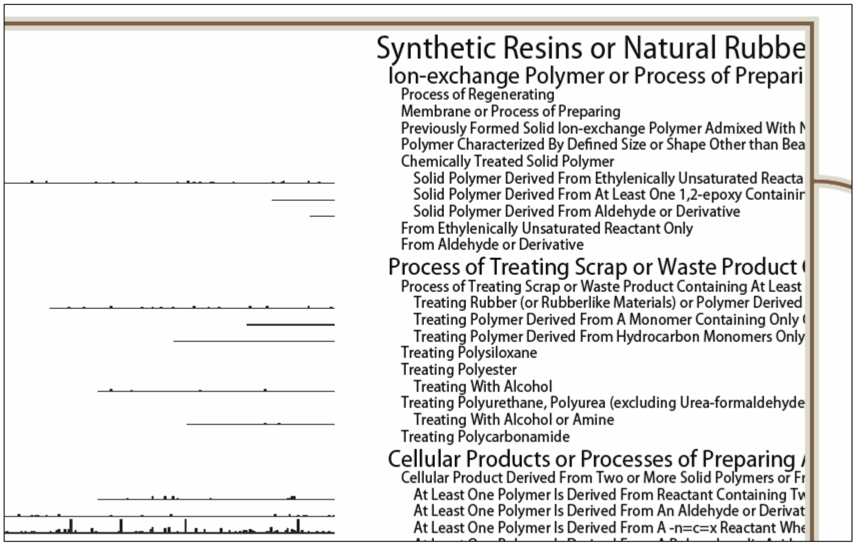
Prior Art

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.



Synthetic Resins or Natural Rubbe
Ion-exchange Polymer or Process of Prepari
Process of Regenerating
Membrane or Process of Preparing
Previously Formed Solid Ion-exchange Polymer Admixed With N
Polymer Characterized By Defined Size or Shape Other than Bea
Chemically Treated Solid Polymer
Solid Polymer Derived From Ethylenically Unsaturated Reacta
Solid Polymer Derived From At Least One 1,2-epoxy Containir
Solid Polymer Derived From Aldehyde or Derivative
From Ethylenically Unsaturated Reactant Only
From Aldehyde or Derivative
Process of Treating Scrap or Waste Product (
Process of Treating Scrap or Waste Product Containing At Least
Treating Rubber (or Rubberlike Materials) or Polymer Derived
Treating Polymer Derived From A Monomer Containing Only (
Treating Polymer Derived From Hydrocarbon Monomers Only
Treating Polysiloxane
Treating Polyester
Treating With Alcohol
Treating Polyurethane, Polyurea (excluding Urea-formaldehyde
Treating With Alcohol or Amine
Treating Polycarbonamide
Cellular Products or Processes of Preparing /
Cellular Product Derived From Two or More Solid Polymers or Fr
At Least One Polymer Is Derived From Reactant Containing Tw
At Least One Polymer Is Derived From An Aldehyde or Derivat
At Least One Polymer Is Derived From A -n=c=x Reactant Whe

Places & Spaces: Mapping Science *exhibit, see also* *http://scimaps.org.*

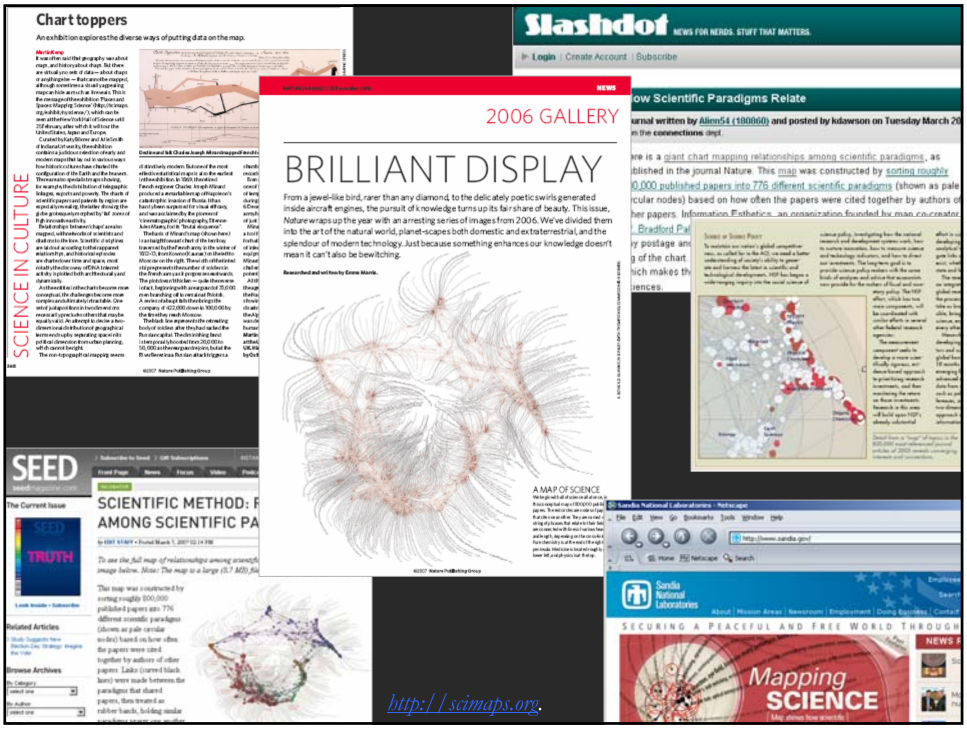


Chart toppers

An exhibition explores the diverse ways of putting data on the map.

Martin Kemp

It was often said that geography was about maps, and history about chaps. But there are virtually no sets of data — about chaps or anything else — that cannot be mapped, although sometimes a visually appealing map can hide as much as it reveals. This is the message of the exhibition 'Places and Spaces: Mapping Science' (http://scimaps.org/exhibit/nysci/), which can be seen at the New York Hall of Science until 25 February, after which it will tour the United States, Japan and Europe.

Curated by Katy Börner and Julie Smith of Indiana University, the exhibition contains a judicious selection of early and modern maps that lay out in various ways how historic cultures have charted the configuration of the Earth and the heavens. There are also specialist maps showing, for example, the distribution of telegraphic linkages, exports and poverty. The charts of scientific papers and patents by region are especially revealing, the latter showing the globe grotesquely morphed by 'hot' zones of high innovative activity.

Relationships between 'chaps' are also mapped, with networks of scientists and disciplines to the fore. Scientific disciplines are laid out according to their apparent relationships, and historical episodes are charted over time and space, most notably the discovery of DNA. Internet activity is plotted both architecturally and dynamically.

As the entities in the charts become more conceptual, the challenges become more complex and ultimately intractable. One set of juxtapositions in two dimensions necessarily precludes others that may be equally valid. An attempt to devise a two-dimensional distribution of geographical terms ends up by separating space in its political dimension from urban planning, which cannot be right.

The non-topographical mapping seems distinctively modern. But one of the most effective statistical maps is also the earliest in the exhibition. In 1869, the retired French engineer Charles Joseph Minard produced a remarkable map of Napoleon's catastrophic invasion of Russia. It has hardly been surpassed for visual efficacy, and was acclaimed by the pioneer of 'cinematographic' photography, Etienne-Jules Marey, for its "brutal eloquence".

The basis of Minard's map (shown here) is a straightforward chart of the territory traversed by the French army in the winter of 1812–13, from Kowno (Kaunas) on the left to Moscow on the right. The width of the tinted strip represents the number of soldiers in the French army as it progresses eastwards. The plot doesn't thicken — quite the reverse in fact, beginning with a vanguard of 33,000 men branching off to remain at Polotsk. A series of abrupt falls then brings the company of 422,000 down to 100,000 by the time they reach Moscow.

The black line represents the retreating body of soldiers after they had sacked the Russian capital. The diminishing band is temporarily boosted from 20,000 to 50,000 as the vanguard rejoins, but at the River Berezina a Russian attack triggers a

Dedicated folk: Charles Joseph Minard mapped French

SCIENCE IN CULTURE

248

©2007 Nature Publishing Group

NATURE|Vol 444|21/28 December 2006

NEWS

2006 GALLERY

BRILLIANT DISPLAY

From a jewel-like bird, rarer than any diamond, to the delicately poetic swirls generated inside aircraft engines, the pursuit of knowledge turns up its fair share of beauty. This issue, *Nature* wraps up the year with an arresting series of images from 2006. We've divided them into the art of the natural world, planet-scapes both domestic and extraterrestrial, and the splendour of modern technology. Just because something enhances our knowledge doesn't mean it can't also be bewitching.

Researched and written by Emma Marris.

A MAP OF SCIENCE

©2007 Nature Publishing Group

Slashdot NEWS FOR NERDS. STUFF THAT MATTERS.

▶ Login | Create Account | Subscribe

How Scientific Paradigms Relate

journal written by Alien54 (180860) and posted by kdawson on Tuesday March 20

from the connections dept.

There is a giant chart mapping relationships among scientific paradigms, as published in the journal Nature. This map was constructed by sorting roughly 800,000 published papers into 776 different scientific paradigms (shown as pale circular nodes) based on how often the papers were cited together by authors of other papers. Information Esthetics, an organization founded by map co-creator W. Bradford Paley … by postage and … of the chart … which makes th… sciences.

SEED seedmagazine.com

Subscribe to Seed | Gift Subscriptions

Front Page | News | Focus | Video | Podcast

The Current Issue

Look Inside • Subscribe

Related Articles

Browse Archives

By Category: select one

By Author: select one

SCIENTIFIC METHOD: R… AMONG SCIENTIFIC PA…

by SEED STAFF • Posted March 7, 2007 02:14 PM

To see the full map of relationships among scientific … image below. Note: The map is a large (8.7 MB) file

This map was constructed by sorting roughly 800,000 published papers into 776 different scientific paradigms (shown as pale circular nodes) based on how often the papers were cited together by authors of other papers. Links (curved black lines) were made between the paradigms that shared papers, then treated as rubber bands, holding similar paradigms nearer one another

Sandia National Laboratories - Netscape

File Edit View Go Bookmarks Tools Window Help

http://www.sandia.gov/

Sandia National Laboratories

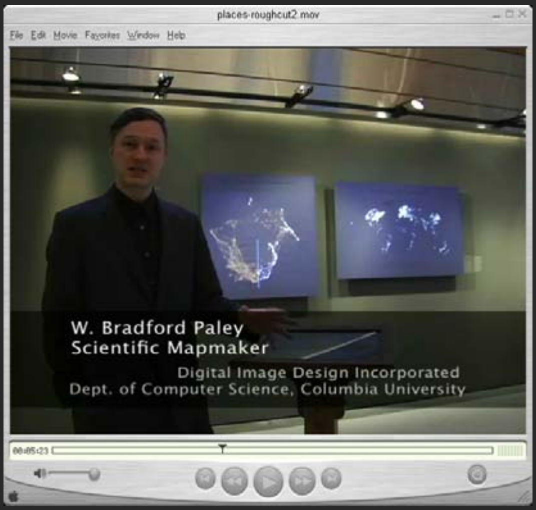
About | Mission Areas | Newsroom | Employment | Doing Business | Contact

SECURING A PEACEFUL AND FREE WORLD THROUGH

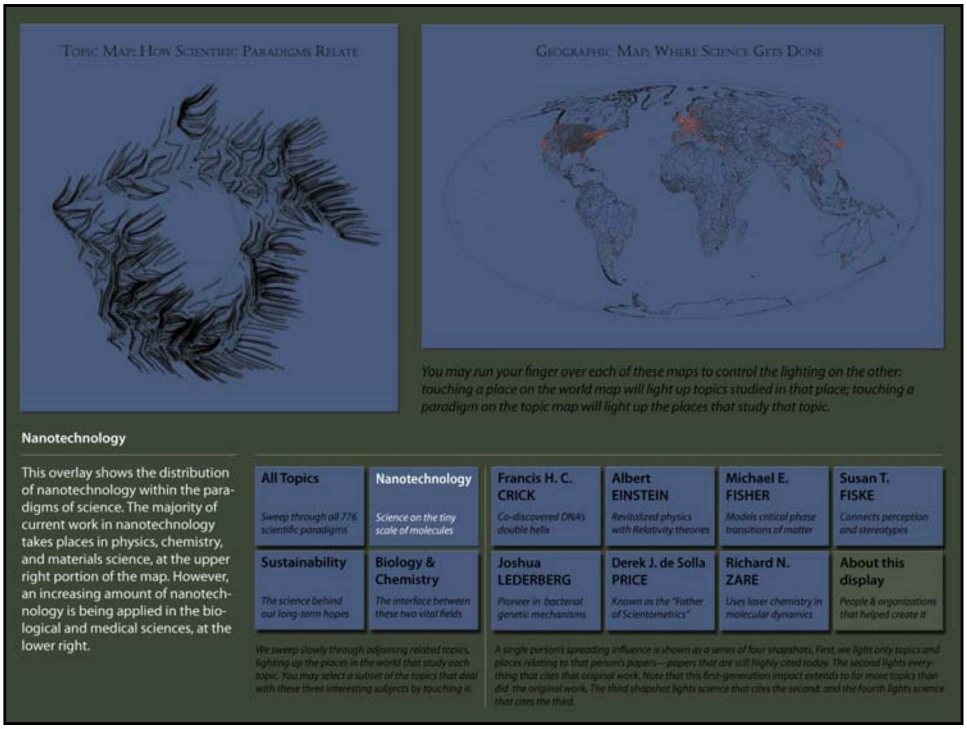
Mapping SCIENCE

NEWS

http://scimaps.org.


places-roughcut2.mov
File Edit Movie Favorites Window Help
W. Bradford Paley
Scientific Mapmaker
Digital Image Design Incorporated
Dept. of Computer Science, Columbia University
00:05:23

"Places & Spaces: Mapping Science"
on display at the NYPL Science, Industry, and Business Library
Madison/34th, New York City
April 3rd - August 31st, 2006.


Topic Map: How Scientific Paradigms Relate
Geographic Map: Where Science Gets Done

You may run your finger over each of these maps to control the lighting on the other: touching a place on the world map will light up topics studied in that place; touching a paradigm on the topic map will light up the places that study that topic.

Nanotechnology

This overlay shows the distribution of nanotechnology within the paradigms of science. The majority of current work in nanotechnology takes places in physics, chemistry, and materials science, at the upper right portion of the map. However, an increasing amount of nanotechnology is being applied in the biological and medical sciences, at the lower right.

All Topics *Sweep through all 776 scientific paradigms*

Nanotechnology *Science on the tiny scale of molecules*

Sustainability *The science behind our long-term hopes*

Biology & Chemistry *The interface between these two vital fields*

We sweep slowly through adjoining related topics, lighting up the places in the world that study each topic. You may select a subset of the topics that deal with these three interesting subjects by touching it.

Francis H. C. CRICK *Co-discovered DNA's double helix*

Albert EINSTEIN *Revitalized physics with Relativity theories*

Michael E. FISHER *Models critical phase transitions of matter*

Susan T. FISKE *Connects perception and stereotypes*

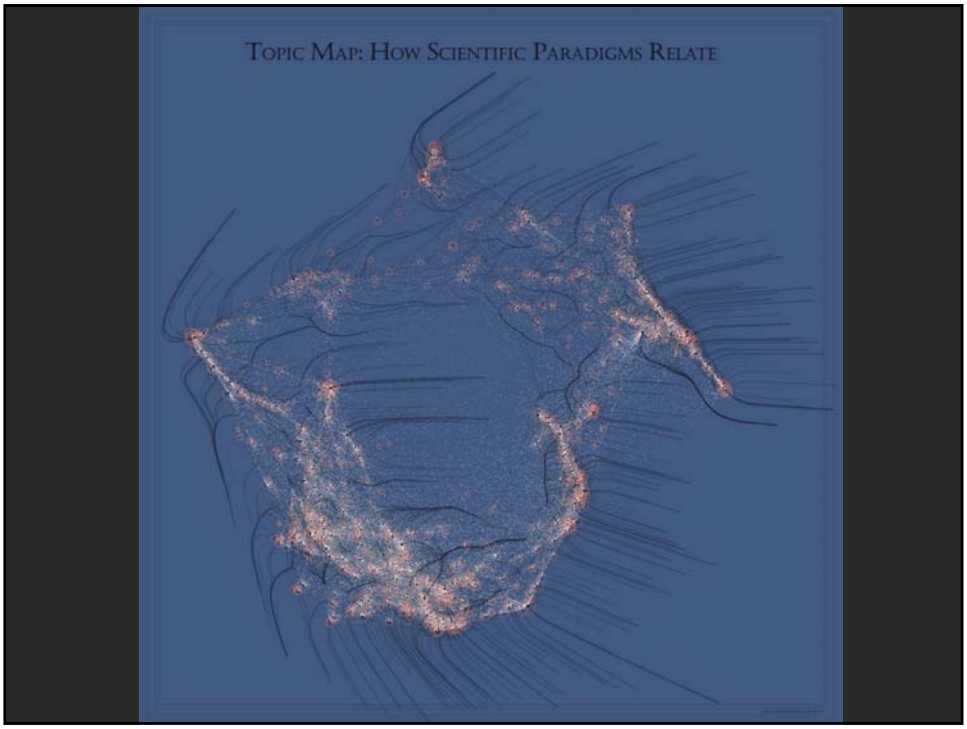
Joshua LEDERBERG *Pioneer in bacterial genetic mechanisms*

Derek J. de Solla PRICE *Known as the "Father of Scientometrics"*

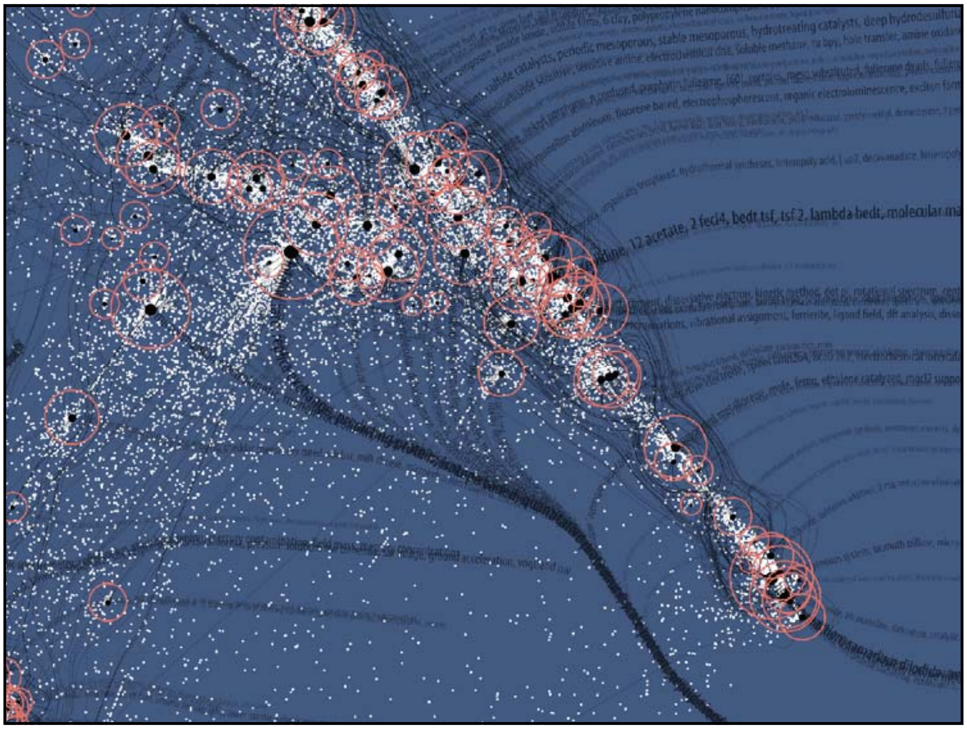
Richard N. ZARE *Uses laser chemistry in molecular dynamics*

About this display *People & organizations that helped create it*

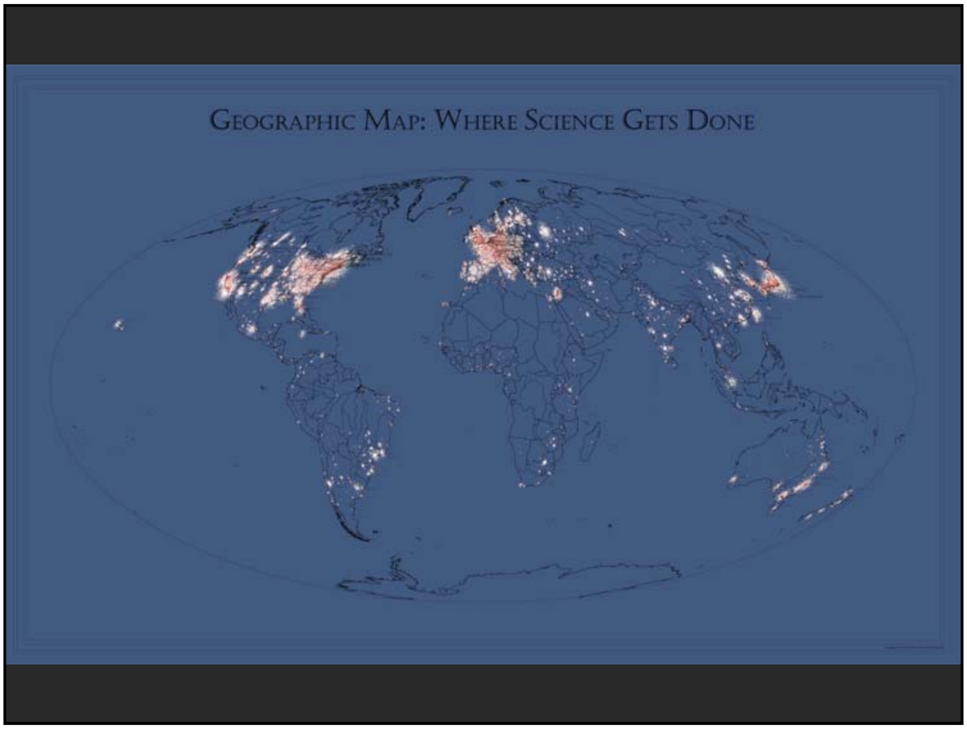
A single person's spreading influence is shown as a series of four snapshots. First, we light only topics and places relating to that person's papers—papers that are still highly cited today. The second lights everything that cites that original work. Note that this first-generation impact extends to far more topics than did the original work. The third shapshot lights science that cites the second; and the fourth lights science that cites the third.



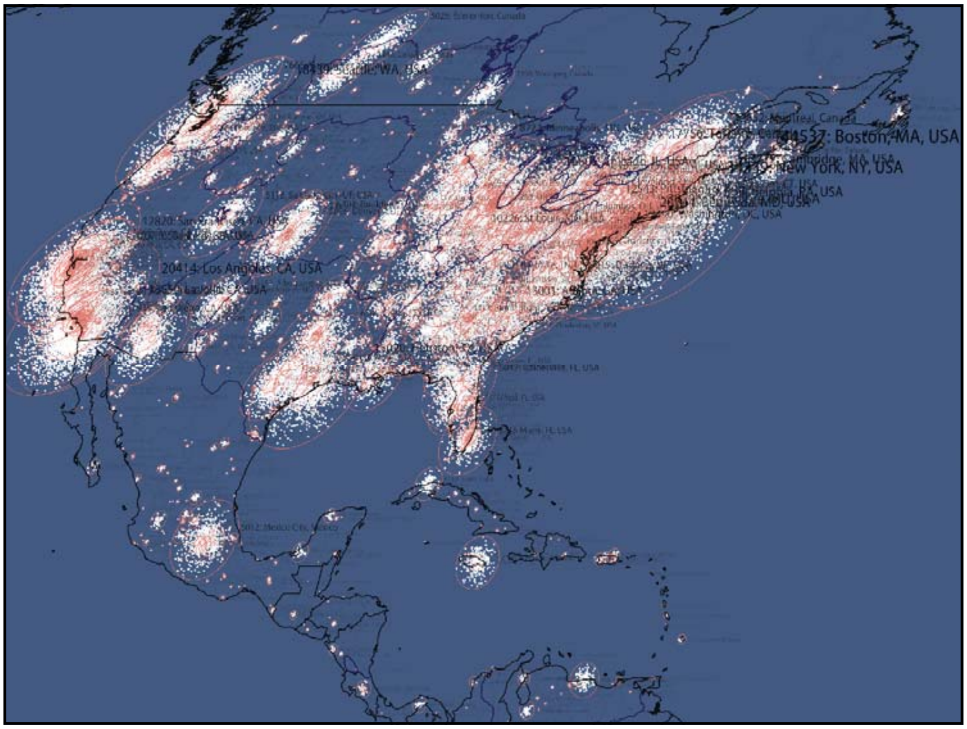
Topic Map: How Scientific Paradigms Relate

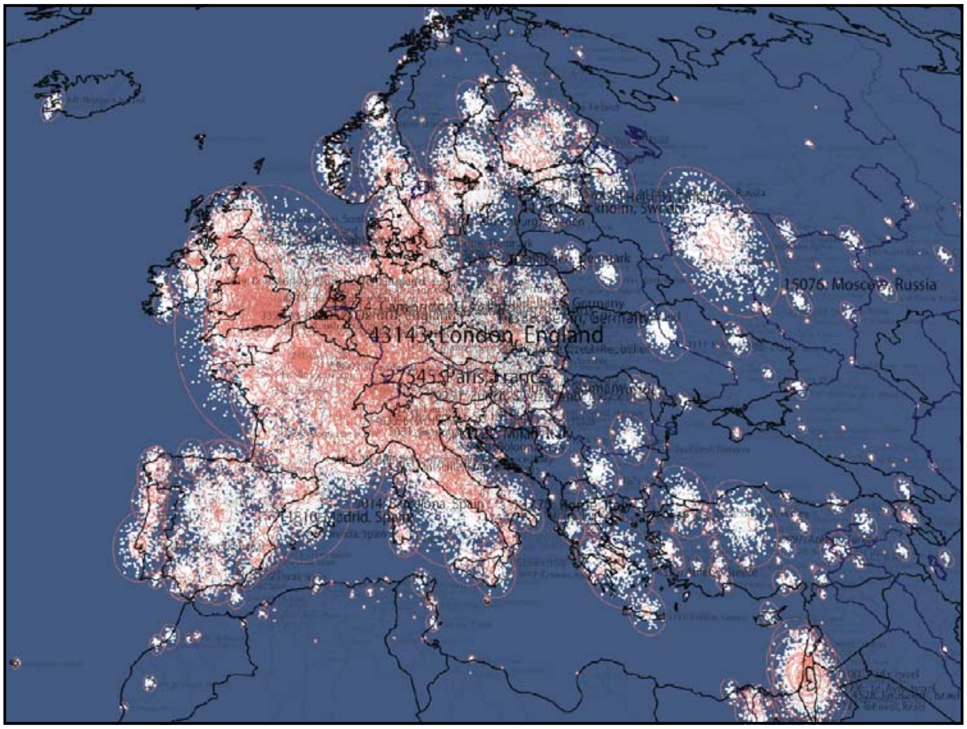


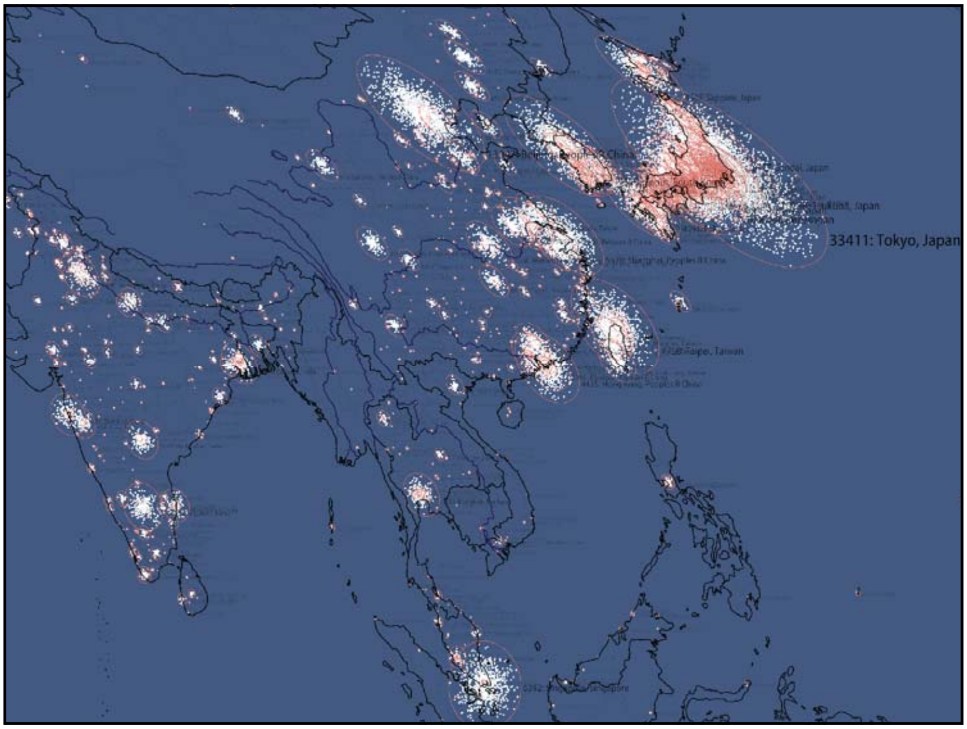
6 clay, polypropylene nanocomposite
periodic mesoporous, stable mesoporous, hydrotreating catalysts, deep hydrodesulfuri
soluble methane, ru bpy, hole transfer, amine oxida
sulfide catalysts
fullerene, [60], pyrrolidine, meso substituted, fullerene dyads, fuller
fluorene based, electrophosphorescent, organic electroluminescence, exciton form
hydrothermal syntheses
12 acetate, 2 fecl4, bedt tsf, tsf 2, lambda bedt, molecular ma
dissociative electron, kinetic method, rotational spectrum
vibrational assignment, ferrierite, ligand field, dft analysis
ethylene catalyzed, mo02 suppo
acceleration

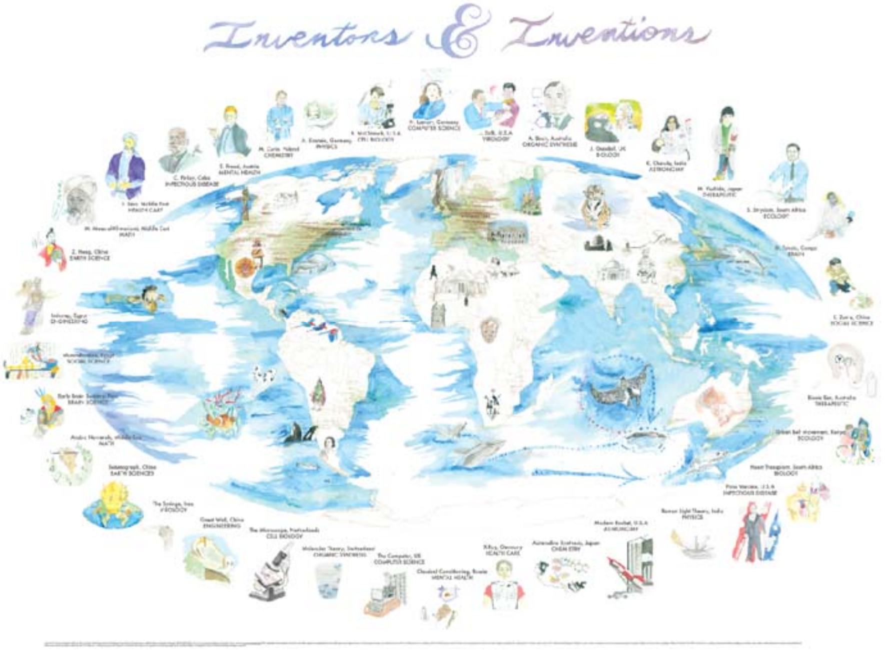


Geographic Map: Where Science Gets Done




15076: Moscow, Russia
43143: London, England
27545: Paris, France
Madrid, Spain


33411: Tokyo, Japan
Taipei, Taiwan



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.)

Inventors & Inventions

Hands-On Science Maps for Kids, by Fileve Palmer (Painting), Julie Smith (Data Acquisition), Elisha Hardy and Katy Börner (Graphic Design), BLOOMINGTON, IN, 2006. Courtesy of Indiana University. Learn more at www.scimaps.org. This map plots the locations of where scientific papers were published: each light green dot represents 10 or fewer papers; they are scattered around the exact location for visibility, within a labelled green circle whose size is proportional to the number of papers published in that place. The base map is part of an "Illuminated diagram" display which used a computer and two projectors, projecting spots of light on the prints to highlight different kinds of scientific research (on a sibling map of scientific paradigms) and the areas in the world where such science was performed. Base map research by Kevin Boyack and Dick Klavans, cartography by John Burgoon, data from Thompson ISI, graphics and typography by W. Bradford Paley. Copyright © 2006 W. Bradford Paley, all rights reserved.

Inventors

Hands-On Science Maps for Kids, by Fileve Palmer (Painting), Julie Smith (Data Acquisition), Elisha Hardy and Katy Börner (Graphic Design), BLOOMINGTON, IN, 2006. Courtesy of Indiana University. Learn more at www.scimaps.org. This map plots the locations of where scientific papers were published: each light green dot represents 10 or fewer papers; they are scattered around the exact location for visibility, within a labelled green circle whose size is proportional to the number of papers published in that place. The base map is part of an "Illuminated diagram" display which used a computer and two projectors, projecting spots of light on the prints to highlight different kinds of scientific research (on a sibling map of scientific paradigms) and the areas in the world where such science was performed. Base map research by Kevin Boyack and Dick Klavans, cartography by John Burgoon, data from Thompson ISI, graphics and typography by W. Bradford Paley. Copyright © 2006 W. Bradford Paley, all rights reserved.

Inventors & Inventions

ASTRONOMY
PHYSICS
ENGINEERING
COMPUTER SCIENCE
MATH
SOCIAL SCIENCE
PHYSICAL CHEMISTRY
BRAIN
MENTAL HEALTH
EARTH SCIENCES
ORGANIC THEORY
ECOLOGY
BIOLOGY
HEALTH CARE
THERAPEUTIC RESEARCH
CELL BIOLOGY
VIRUSES
INFECTIOUS DISEASE

Hands-On Science Maps for Kids, by Fileve Palmer (Painting), Julie Smith (Data Acquisition), Elisha Hardy and Katy Börner (Graphic Design), BLOOMINGTON, IN, 2006. Courtesy of Indiana University. Learn more at www.scimaps.org. This map plots the locations of where scientific papers were published: each light green dot represents 10 or fewer papers; they are scattered around the exact location for visibility, within a labeled green circle whose size is proportional to the number of papers published in that place. The base map is part of an "Illuminated diagram" display which used a computer and two projectors, projecting spots of light on the prints to highlight different kinds of scientific research (on a sliding map of scientific paradigms) and the areas in the world where such

Inventors

ASTRONOMY
PHYSICS
ENGINEERING
COMPUTER SCIENCE
MATH
SOCIAL SCIENCE
PHYSICAL CHEMISTRY
BRAIN
MENTAL HEALTH
EARTH SCIENCES
ORGANIC THEORY
ECOLOGY
BIOLOGY
HEALTH CARE
THERAPEUTIC RESEARCH
CELL BIOLOGY
VIRUSES
INFECTIOUS DISEASE

Hands-On Science Maps for Kids, by Fileve Palmer (Painting), Julie Smith (Data Acquisition), Elisha Hardy and Katy Börner (Graphic Design), BLOOMINGTON, IN, 2006. Courtesy of Indiana University. Learn more at www.scimaps.org. This map plots the locations of where scientific papers were published: each light green dot represents 10 or fewer papers; they are scattered around the exact location for visibility, within a labeled green circle whose size is proportional to the number of papers published in that place. The base map is part of an "Illuminated diagram" display which used a computer and two projectors, projecting spots of light on the prints to highlight different kinds of scientific research (on a sliding map of scientific paradigms) and the areas in the world where such

1. Dream Tools for Scholarly Knowledge Management

2. **Challenges**
3. Opportunities

Challenges - Data Collection & Integration

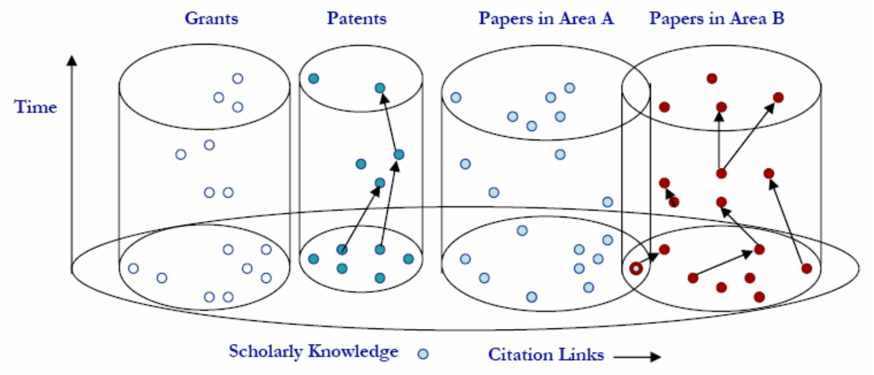
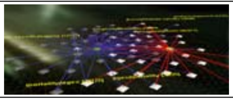


Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type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

Börner, K. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) Visualizing the Semantic Web, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.



Challenges - Semantic Mining / Integration

Works well if the records are written in similar styles, using similar formatting and conventions, are of similar length, etc.

Works less well if applied to interdisciplinary or multi-lingual datasets because

- Words, e.g., ‘prototype’, have very different meanings in computer science, biology, psychology, architecture, etc.
- Paper titles are frequently used to demonstrate the creativity of authors, e.g., “All you ever wanted to know about x”, “A unifying theory of x”.
- Author supplied keywords are useful to identify an author but not to find similar papers. Note: Controlled vocabularies/thesauri work well.

Humans might simply be too different and too creative to produce proper raw material that can be analyzed using existing text mining and data mining algorithms.

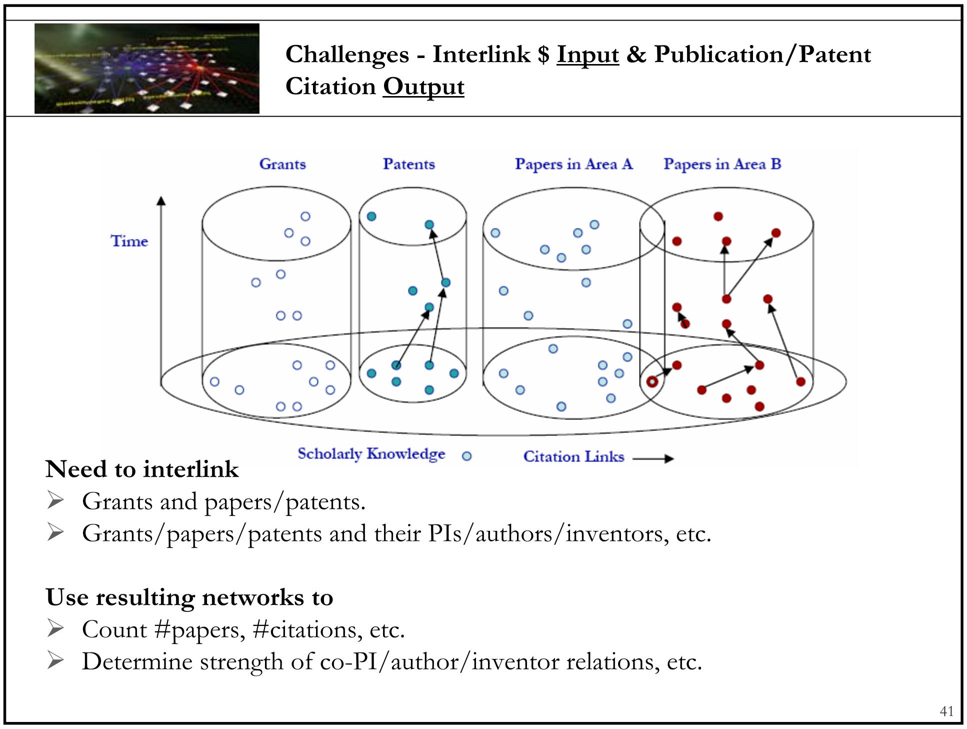
Challenges - Link Traversal & Link Mining

Link Search: Google

(Citation) Link Traversal in One Database: *Thomson Scientific, Google Scholar,* and *CiteSeer* already support citation link traversal. The *Proceedings of the National Academy of Sciences of the United States of* America (PNAS) online interface <http://www.pnas.org>. provides citation maps that shows articles citing or being cited by a selected article.

(Citation) Link Traversal Across Databases: The *Library Without Walls* project <http://library.lanl.gov/lww/> at the *Los Alamos National Laboratory* interlinks major publication databases and supporting citation based search across different holdings that have citation linkages.

(Co-Author) Link Traversal: Some digital libraries such as the citation indexes published by *Thomson Scientific*, *DBLP Bibliography Server* <http://www.sigmod.org/dblp/db/>, and *ACM Digital Library* <http://portal.acm.org> provide information on co-authorships. Services comprise a listing of all papers by an author, a listing of all co-authors for one author, and co-author link traversal.



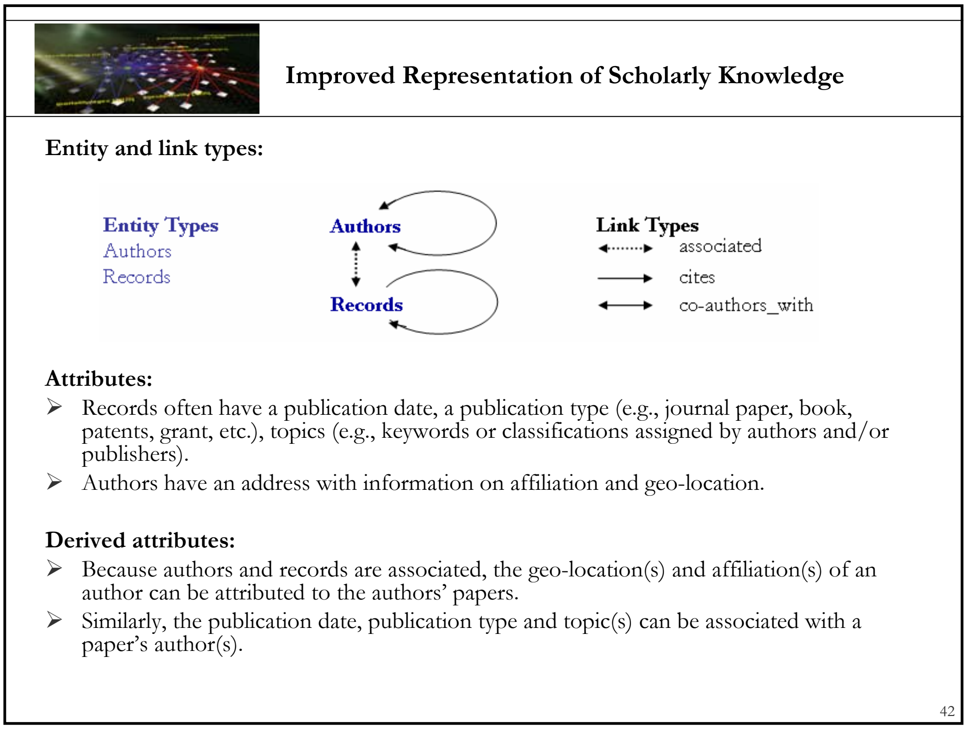
Challenges - Interlink $ Input & Publication/Patent Citation Output

Need to interlink

- Grants and papers/patents.
- Grants/papers/patents and their PIs/authors/inventors, etc.

Use resulting networks to

- Count #papers, #citations, etc.
- Determine strength of co-PI/author/inventor relations, etc.



Improved Representation of Scholarly Knowledge

Entity and link types:

Entity Types
Authors
Records

Link Types

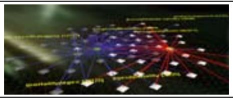
◂·······▸	associated
⟶	cites
⟷	co-authors_with

Attributes:

- Records often have a publication date, a publication type (e.g., journal paper, book, patents, grant, etc.), topics (e.g., keywords or classifications assigned by authors and/or publishers).
- Authors have an address with information on affiliation and geo-location.

Derived attributes:

- Because authors and records are associated, the geo-location(s) and affiliation(s) of an author can be attributed to the authors' papers.
- Similarly, the publication date, publication type and topic(s) can be associated with a paper's author(s).



Improved Representation of Scholarly Knowledge makes possible

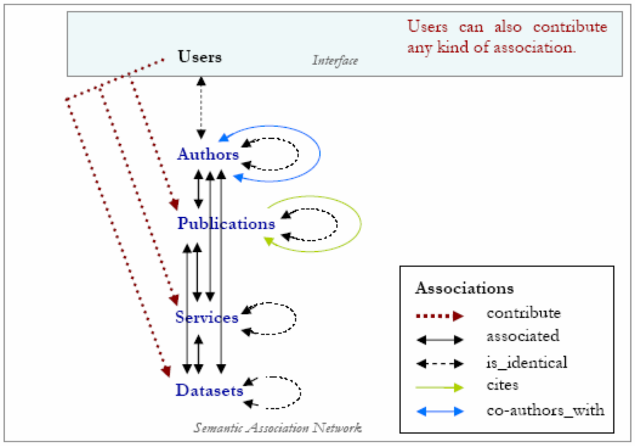
Statistics:

- Number of papers, grants, co-authorships, citation (over time) per author.
- Bursts of activity (#citations, #$, #patents, #collaborators, etc.).
- Changes of topics and geo-locations for authors and their institutions over time.

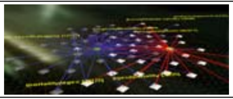
Visualizations:

- Geospatial and topical distribution of funding input & research output.
- Structure and evolution of research topics.
- Evolving research areas (e.g., based on young yet highly cited papers).
- Diffusion of information, people, $s over geospatial and topic space.

Semantic Association Networks



Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) Visualizing the Semantic Web, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.



Open Questions

- ➢ Interoperability: How to add more and more databases?
- ➢ Interlinkage: OAI works. What standard would work best for unique and persistent identifiers for authors/institutions/ countries/journals/geolocations/etc.?
  Will 95% automatic and 5% manual data cleaning work?
- ➢ How to add databases/services while in production?
- ➢ How to exploit peer-to-peer architectures?
- ➢ How to resolve proprietary/political issues?

CIShell
cyberinfrastructure shell

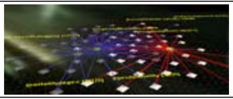
45

1. Dream Tools for Scholarly Knowledge Management

2. Challenges

3. Opportunities

46



Opportunities for Mapping Science

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)**.

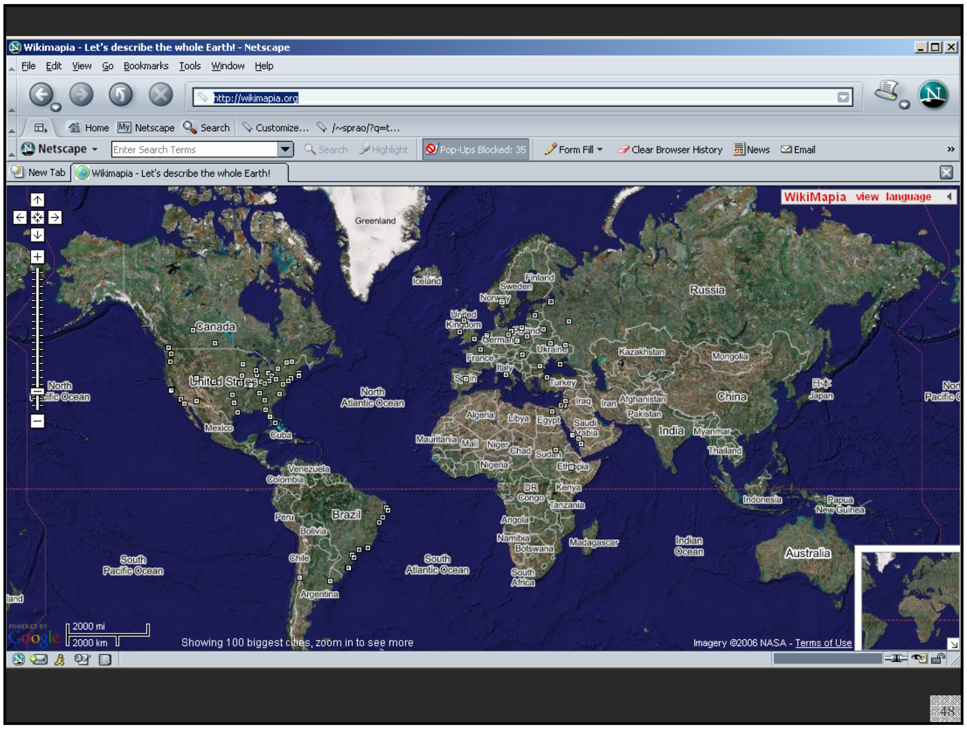
Advantages for Publishers

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

For Society

- Dramatically improved access to scientific knowledge and expertise.

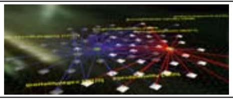
47



Wikimapia - Let's describe the whole Earth! - Netscape

http://wikimapia.org

48



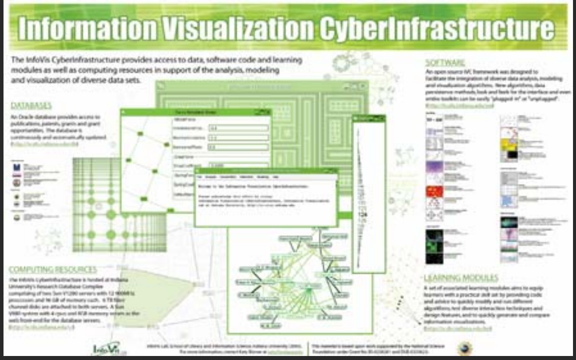
Opportunities for Modeling Science

- Dynamic science and technology indicators (emerging research frontiers, evolving networks, trends, feedback loops).
- Evolution of scientific communities/fields. Capacity limit to knowledge/ skills knowable by individual researchers.
- Interplay of competition and collaboration.
- Evolution of fields – birth, growth, mature, decline.
- Interactions among fields. Optimal interdisciplinary collaborations?
- Comparison of different funding models, e.g., few large vs. many small grants, teach the field how to fish or give them fish?
- Impact of publishing/collaboration/funding mechanisms on the dynamics of fields.
- Diffusion of people, ideas, skills, etc.
- How to best communicate modeling results/insights?

Ultimate goal:

Learn how to best increase, diffuse, and utilize our collective scholarly knowledge.

49


Information Visualization CyberInfrastructure
The InfoVis CyberInfrastructure provides access to data, software code and learning modules as well as computing resources in support of the analysis, modeling and visualization of diverse data sets.
DATABASES
SOFTWARE
COMPUTING RESOURCES
LEARNING MODULES
InfoVis

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

CIShell
cyberinfrastructure shell

NetworkWorkbench
A Workbench for Network Scientists

NSF SBC Sun microsystems HPNAP High Performance Network Applications Program 21st Century Research & Technology Fund

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

50

Building Market Places not Cathedrals

- ➢ Requires the design & implementation of 'software glue' that can interlink datasets and algorithms written in different languages using different data formats.
- ➢ The smaller the glue or 'CI Shell', the more likely it can be maintained.
- ➢ Dataset and algorithm 'plugins' are provided by application

Cyberinfrastructure Shell (CIShell)

http://cishell.org

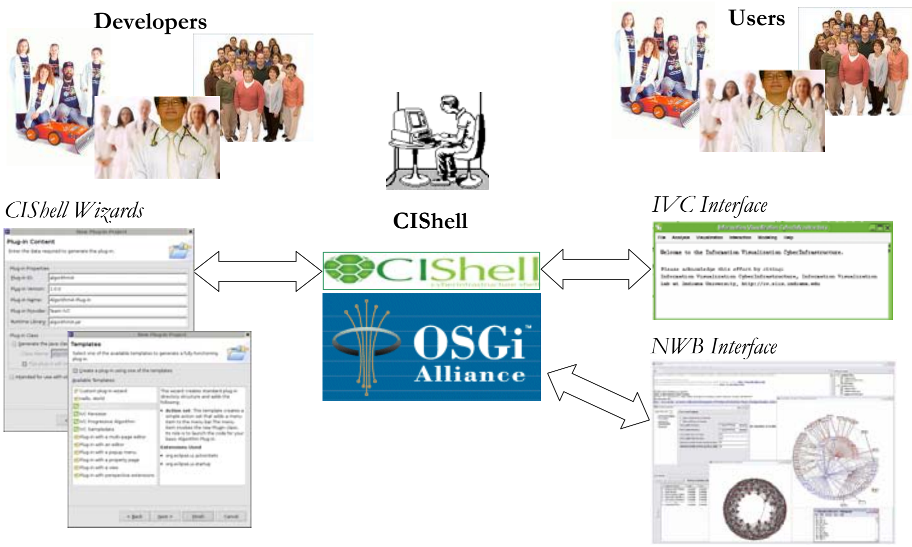
CIShell is an 'empty shell' that supports

- ➢ Easy integration of new datasets and algorithms by algorithm developers and
- ➢ Easy usage of algorithms by algorithm users.

Its plug-and-play architecture supports the integration and utilization of diverse

- ➢ Datasets, e.g., stored in files, databases, streaming data.
- ➢ Algorithms, e.g., data processing, analysis, modeling, visualization.
- ➢ Interfaces, e.g., remote services, scripting engines, peer-to-peer clients.
- ➢ Services, e.g., workflow support, scheduler.

CIShell – Needs of Algorithm Developers & Users



Developers

Users

CIShell

CIShell Wizards

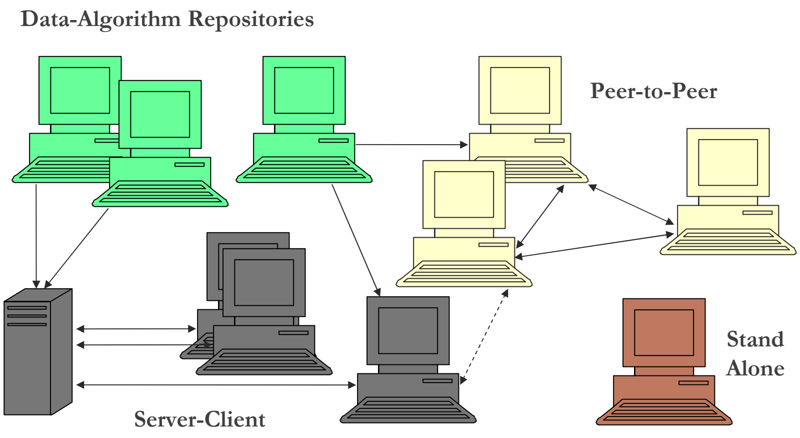
IVC Interface

Welcome to the Information Visualization CyberInfrastructure.

Please acknowledge this effort by citing:
Information Visualization CyberInfrastructure, Information Visualization Lab at Indiana University, http://iv.slis.indiana.edu

NWB Interface

CIShell – Deployment



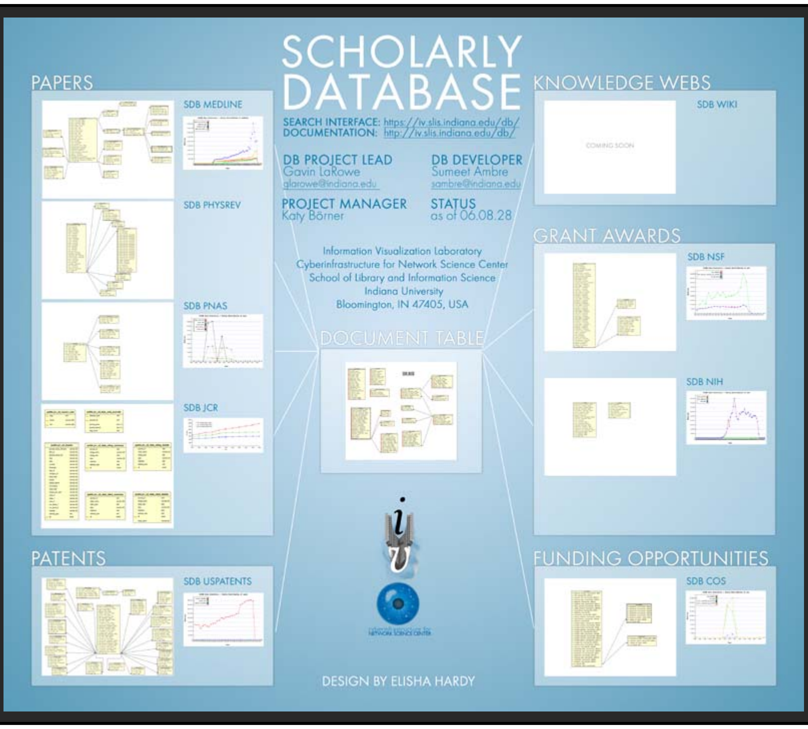
Data-Algorithm Repositories

Peer-to-Peer

Stand Alone

Server-Client

CIShell applications can be deployed as distributed data and algorithm repositories, stand alone applications, peer-to-peer architectures, and server-client architectures.

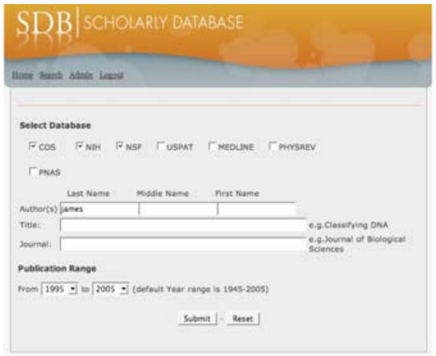


SDB | SCHOLARLY DATABASE

Scholarly Database: Web Interface

Search across publications, patents, grants.

Download records and/or (evolving) co-author, paper-citation networks.



Select Database

☑ COS ☑ NIH ☑ NSF □ USPAT □ MEDLINE □ PHYSREV □ PNAS

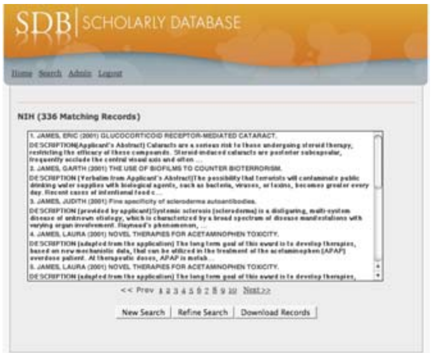
	Last Name	Middle Name	First Name
Author(s)	james		

Title: e.g.Classifying DNA

Journal: e.g.Journal of Biological Sciences

Publication Range

From 1995 to 2005 (default Year range is 1945-2005)

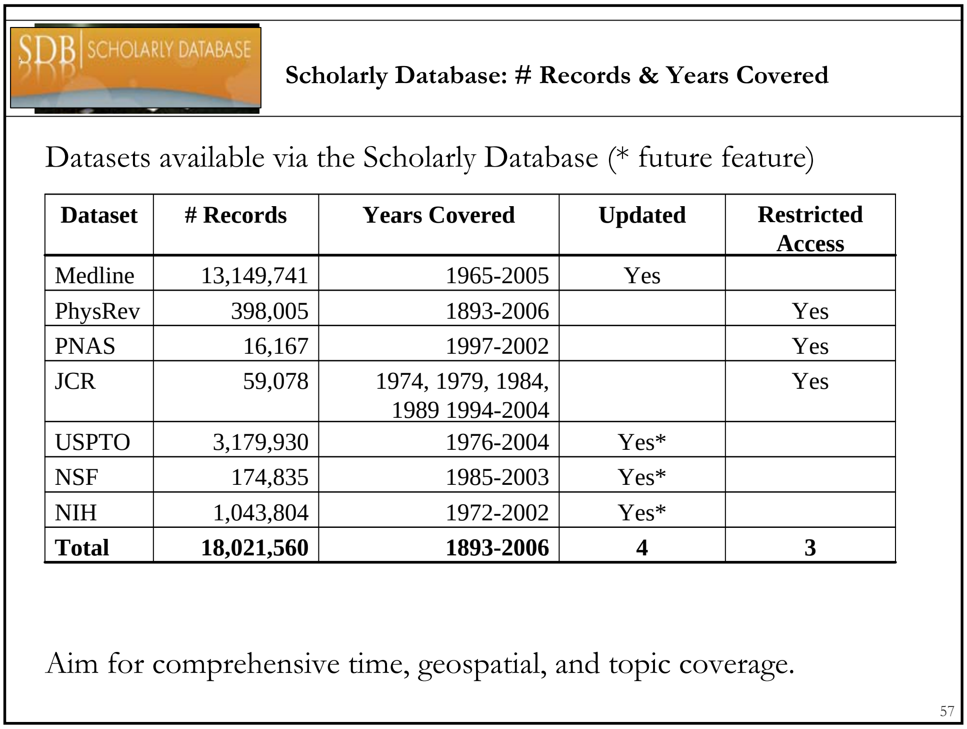


NIH (336 Matching Records)

1. JAMES, ERIC (2001) GLUCOCORTICOID RECEPTOR-MEDIATED CATARACT.
DESCRIPTION(Applicant's Abstract) Cataracts are a serious risk to those undergoing steroid therapy, restricting the efficacy of these compounds. Steroid-induced cataracts are posterior subcapsular, frequently occlude the central visual axis and often ...
2. JAMES, GARTH (2001) THE USE OF BIOFILMS TO COUNTER BIOTERRORISM.
DESCRIPTION (Verbatim from Applicant's Abstract)The possibility that terrorists will contaminate public drinking water supplies with biological agents, such as bacteria, viruses, or toxins, becomes greater every day. Recent cases of intentional food c...
3. JAMES, JUDITH (2001) Fine specificity of scleroderma autoantibodies.
DESCRIPTION (provided by applicant)Systemic sclerosis (scleroderma) is a disfiguring, multi-system disease of unknown etiology, which is characterized by a broad spectrum of disease manifestations with varying organ involvement. Raynaud's phenomenon, ...
4. JAMES, LAURA (2001) NOVEL THERAPIES FOR ACETAMINOPHEN TOXICITY.
DESCRIPTION (adapted from the application) The long term goal of this award is to develop therapies, based on new mechanistic data, that can be utilized in the treatment of the acetaminophen (APAP) overdose patient. At therapeutic doses, APAP is metab...
5. JAMES, LAURA (2001) NOVEL THERAPIES FOR ACETAMINOPHEN TOXICITY.
DESCRIPTION (adapted from the application) The long term goal of this award is to develop therapies,

<< Prev 1 2 3 4 5 6 7 8 9 10 Next >>

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



SDB | SCHOLARLY DATABASE

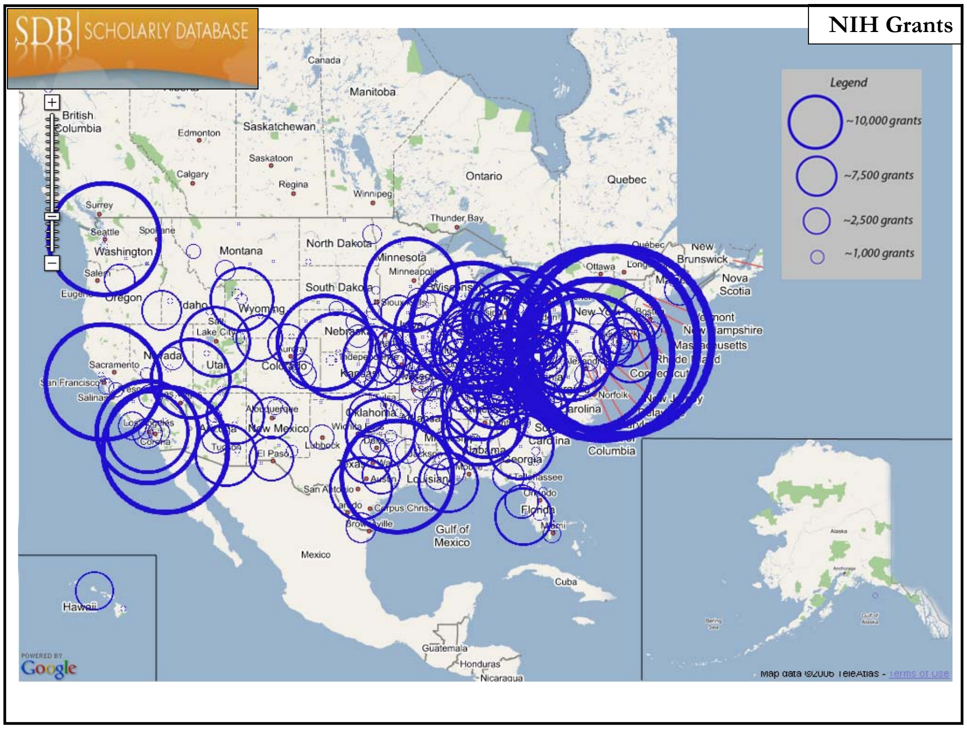
Scholarly Database: # Records & Years Covered

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.

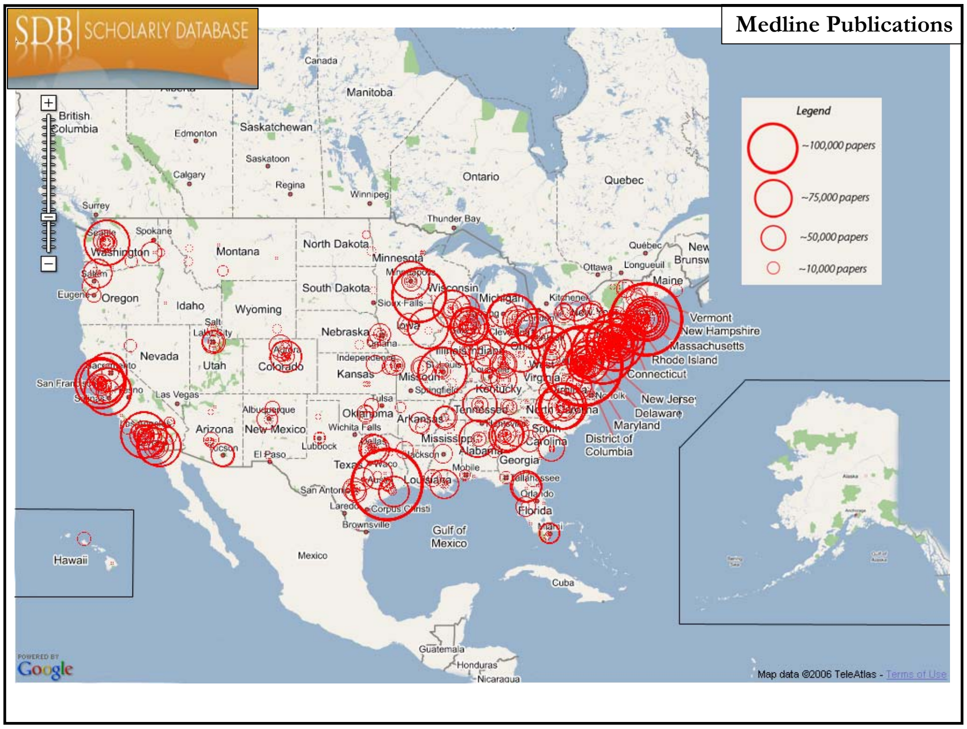
57



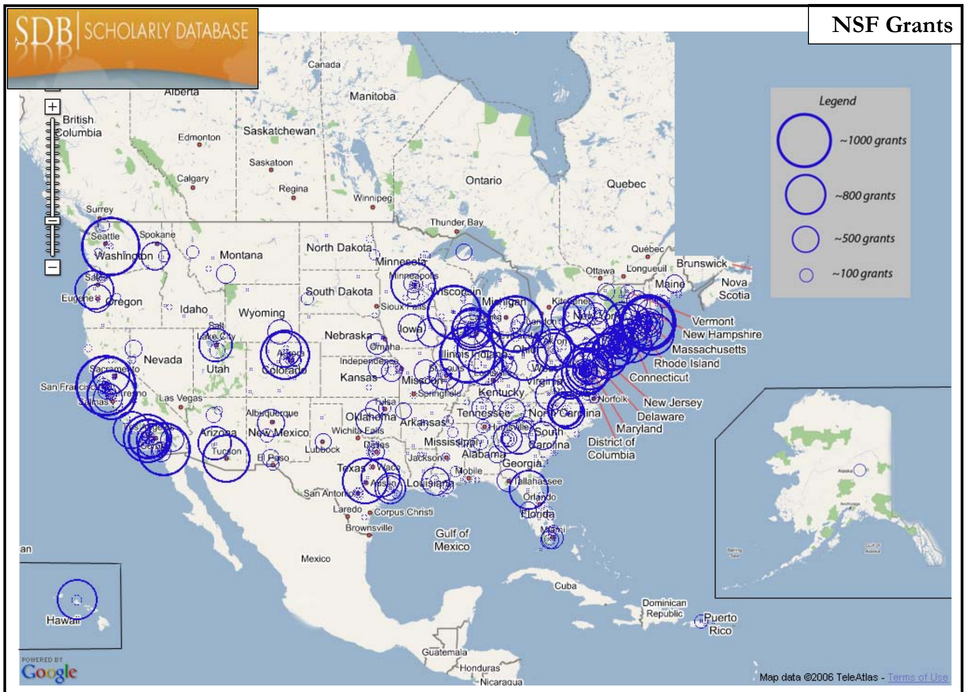
SDB | SCHOLARLY DATABASE

NIH Grants

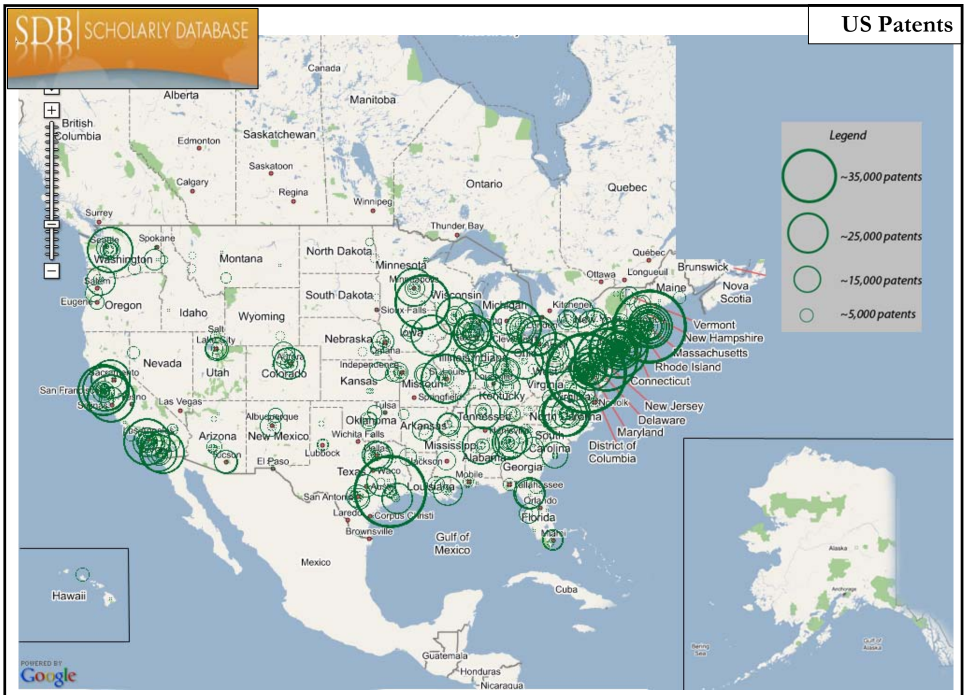
Medline Publications



NSF Grants



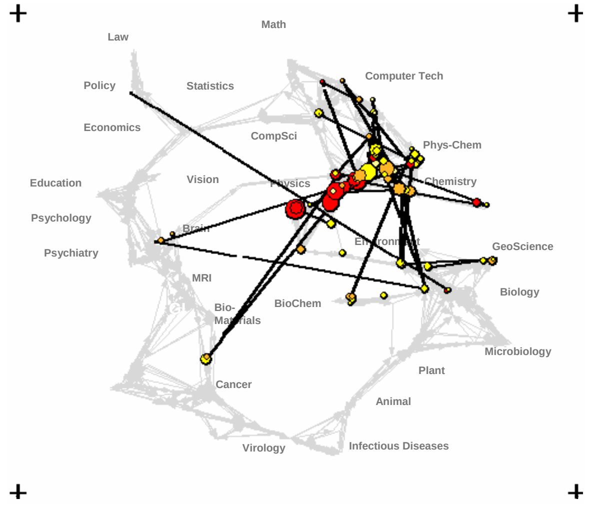
US Patents



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

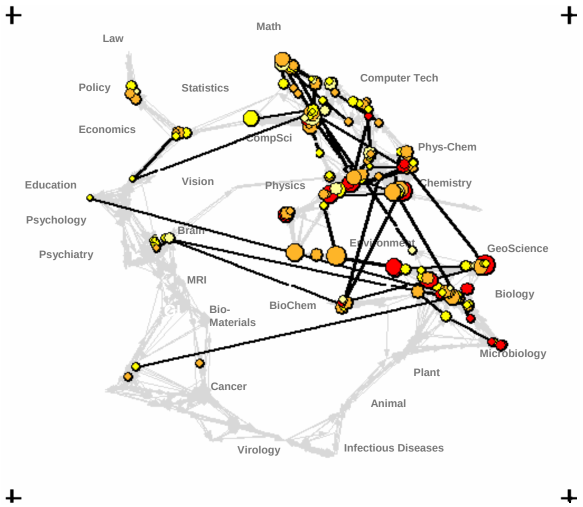
Funding patterns of the US Department of Energy (DOE)



Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

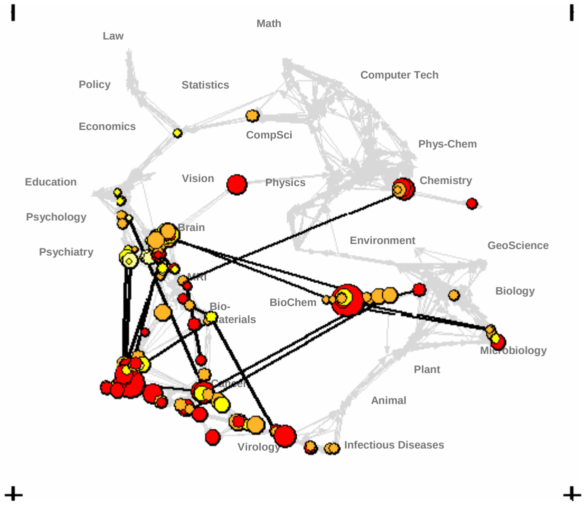
Funding Patterns of the National Science Foundation (NSF)

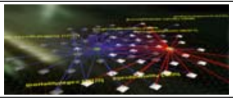


Science map applications: Identifying core competency

Kevin W. Boyack & Richard Klavans, unpublished work.

Funding Patterns of the National Institutes of Health (NIH)





References

- Bruce Herr, Weixia Huang, Shashikant Penumarthy, Katy Börner. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. Submitted to William S. Bainbridge (Ed.) Progress in Convergence. Annals of the New York Academy of Sciences.
- Börner, Katy. Mapping All of Science: How to Collect, Organize and Make Sense of Mankind's Scholarly Knowledge and Expertise. Accepted for *Environment and Planning B*, Special Issue on *Mapping Humanity's Knowledge and Expertise in the Digital Domain*.
- 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.
- Holloway, Todd, Božicevic, Miran and Börner, Katy. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. Accepted for *Complexity*. Also available as cs.IR/0512085.
- Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. In Vladimir Geroimenko & Chaomei Chen (eds.) *Visualizing the Semantic Web*, Springer Verlag, 2nd Edition, chapter 11, pp. 183-198.
- Boyack, Kevin W., Klavans, R. and Börner, Katy. (2005). Mapping the Backbone of Science. *Scientometrics,* 64(3), 351-374.
- Hook, Peter A. and Börner, Katy. (2005) Educational Knowledge Domain Visualizations: Tools to Navigate, Understand, and Internalize the Structure of Scholarly Knowledge and Expertise. In Amanda Spink and Charles Cole (eds.) *New Directions in Cognitive Information Retrieval*. Springer-Verlag, Netherlands, chapter 5, pp. 187-208.
- Börner, Katy, Dall'Asta, Luca, Ke, Weimao and Vespignani, Alessandro. (April 2005) Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*, special issue on *Understanding Complex Systems*, 10(4): pp. 58 - 67. Also available as cond-mat/0502147.
- Ord, Terry J., Martins, Emília P., Thakur, Sidharth, Mane, Ketan K., and Börner, Katy. (2005) Trends in animal behaviour research (1968-2002): Ethoinformatics and mining library databases. *Animal Behaviour*, 69, 1399-1413. Supplementary Material.
- Mane, Ketan K. and Börner, Katy. (2004). Mapping Topics and Topic Bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5287-5290. Also available as cond-mat/0402380.
- Börner, Katy, Maru, Jeegar and Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl_1):5266-5273. Also available as cond-mat/0311459.