Plug-and-Play Macroscopes: Custom Tools for Data Analysis, Modeling, and Visualization

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

With special thanks to the members at the Cyberinfrastructure for Network Science Center, Mapping Science exhibit map makers and advisory board members, and the VIVO team.

THE CHALLENGES OF VISUALISING BIOLOGICAL DATA
Workshop run by the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the Arts and Humanities Research Council (AHRC)
The Grand by Thistle, Bristol, BS1 2EL
November 16-17, 2010
Th Wh l B i C t l

Drew Berry

http://www.home-2009.com

http://www.malarialifecycle.com

The Whole Brain Catalog:
http://wholebraincatalog.org

Drew Berry
All three care deeply about

1. Data,
2. Existing expertise and insight needs, and
3. Are able to acquire the resources it takes to

• Spent months/years to deeply understand the problem/possible solutions.
• Render data optimally for the human perceptual-cognitive system – given our current understanding of human perception/cognition and technology.

The result are insightful yet perceptually stunning, intellectually stimulating, and emotion provoking imagery.
Today's massive amounts of streaming data cannot be rendered by hand.

How to use computers to “envision” biomedical science?

How to combine data mining and visualization algorithms to explore and communicate biomedical science?
Plug-and-Play Macroscope Design Using OSGi/CIShell

- CIShell ([http://cishell.org](http://cishell.org)) is an open source software specification for the integration and utilization of datasets, algorithms, and tools.
- It extends the Open Services Gateway Initiative (OSGi) ([http://www.osgi.org](http://www.osgi.org)), a standardized, component oriented, computing environment for networked services widely used in industry since more than 10 years.
- Specifically, CIShell provides “sockets” into which existing and new datasets, algorithms, and tools can be plugged using a wizard-driven process.
1.) Type of Analysis vs. Level of Analysis

Exemplified in Biomedicine

2.) Needs-Driven Workflow Design

using a modular data acquisition/analysis/modeling/visualization pipeline as well as modular visualization layers.

Implementation in different plug-and-play tools/CIs

<table>
<thead>
<tr>
<th>Statistical Analysis/Profiling</th>
<th>Micro/Individual (1-100 records)</th>
<th>Meso/Local (101–10,000 records)</th>
<th>Macro/Global (10,000 &lt; records)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Analysis (When)</td>
<td>Funding portfolio of one individual</td>
<td>Mapping topic bursts in 20-years of PNAS</td>
<td>113 Years of Physics Research</td>
</tr>
<tr>
<td>Geospatial Analysis (Where)</td>
<td>Career trajectory of one individual</td>
<td>Mapping a states intellectual landscape</td>
<td>PNAS Publications</td>
</tr>
<tr>
<td>Topical Analysis (What)</td>
<td>Base knowledge from which one grant draws.</td>
<td>Knowledge flows in Chemistry research</td>
<td>VxOrd/Topic maps of NIH funding</td>
</tr>
<tr>
<td>Network Analysis (With Whom?)</td>
<td>NSF Co-PI network of one individual</td>
<td>Co-author network</td>
<td>NSF’s core competency</td>
</tr>
</tbody>
</table>
### Type of Analysis vs. Level of Analysis

<table>
<thead>
<tr>
<th>Statistic Analysis/Profiling</th>
<th>Micro/Individual (1-100 records)</th>
<th>Meso/Local (101-10,000 records)</th>
<th>Macro/Global (10,000 &lt; records)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual person and their expertise profiles</td>
<td>Larger labs, centers, universities, research domains, or states</td>
<td>All of NSF, all of science, all of USA, all of society</td>
<td></td>
</tr>
<tr>
<td>Temporal Analysis (When)</td>
<td>Funding portfolio of one individual</td>
<td>Mapping topic bursts in 20-years of PNAS</td>
<td>113 Years of PNAS Research</td>
</tr>
<tr>
<td>Geospatial Analysis (Where)</td>
<td>Career trajectory of one individual</td>
<td>Mapping a state’s intellectual landscape</td>
<td>PNAS, VxOrd/Topic Maps, NIH funding</td>
</tr>
<tr>
<td>Topical Analysis (What)</td>
<td>Base knowledge from which one grant draws.</td>
<td>Knowledge flows in Chemistry research</td>
<td>NIH’s Funding Map</td>
</tr>
<tr>
<td>Network Analysis (With Whom?)</td>
<td>NSF Co-PI network of one individual</td>
<td>Co-author network NIH’s core competency</td>
<td>NIH’s Collaborative Network</td>
</tr>
</tbody>
</table>

### Mapping Indiana’s Intellectual Space

**Geospatial/Network Analysis**

2001-2006, BioMed, IN Scope

*Academic-Industry collaborations and knowledge diffusion*
Mapping Topic Bursts

Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.


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


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 access high quality data and expertise and 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 institutions?

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.
This map highlights the research collaborations of the Chinese Academy of Sciences with locations in China and countries around the world. The large geographic map shows the research collaborations of all CAS institutes. Each smaller geographic map shows the research collaborations by the CAS researchers in one province-level administrative division. Collaborations between CAS researchers are not included in the data. On each map, locations are colored on a logarithmic scale by the number of collaborations from red to yellow. The darkest red is 3,395 collaborations by all of CAS with researchers in Beijing. Also, flow lines are drawn from the location of focus to all locations collaborated with. The width of the flow line is linearly proportional to the number of collaborations with the locations it goes to, with the smallest flow lines representing one collaboration and the largest representing differing amounts on each geographic map.

**Geospatial Analysis**

*World, Chinese Academy of Science*

**Collaboration and knowledge diffusion via co-author networks**

**Individual Co-PI Network**

*Ke & Börner, (2006)*

**Temporal/Network Analysis**

*2001-2006, US, InfoVis Scope*

**Evolving project-PI networks**
Mapping the Evolution of Co-Authorship Networks

Temporal/Network Analysis
1986-2004, US, InfoVis Scope
Evolving co-author networks and author impact

Legend
Nodes = Authors
Node color = Number of papers published
Node size = Number of citations
Edges = Co-authorship relations
Edge color = Year of first co-authorship
Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams

Research question:
• Is science driven by prolific single experts or by high-impact co-authorship teams?

Contributions:
• New approach to allocate citational credit.
• Novel weighted graph representation.
• Visualization of the growth of weighted co-author network.
• Centrality measures to identify author impact.
• Global statistical analysis of paper production and citations in correlation with co-authorship team size over time.
• Local, author-centered entropy measure.

113 Years of Physical Review

http://scimaps.org/dev/map_detail.php?map_id=171
Bruce W. Herr II and Russell Dubon (Data Mining & Visualization), Elisha F. Hardy (Graphic Design), Shashikant Penumarthi (Data Preparation) and Katy Börner (Concept)
Topical Composition and Knowledge Flow Patterns in Chemistry Research for 1974 and 2004
Kevin W. Boyack, Katy Börner, & Richard Klavans (2007)

Chemistry - Biology Interface

Temporal/Network Analysis
1974-2004, US, NSF Chemistry Funding Scope
Mapping the main structure, topic evolution, and knowledge diffusion

Latest ‘Base Map’ of Science

➢ 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

Funding patterns of the US Department of Energy (DOE)

Science map applications: Identifying core competency

Funding Patterns of the National Science Foundation (NSF)
Science map applications: Identifying core competency

Funding Patterns of the National Institutes of Health (NIH)

Topic/Network Analysis
Funding Scope
Map main structure of science and funding profiles
Interactive Science Map of NIH Funding

Topic/Network Analysis
2007, US, NIH Funding Scope
Map main structure of NIH funding and institutes’ funding profiles


Structure of the Remaining Talk

1.) Type of Analysis vs. Level of Analysis
Exemplified in Biomedicine

2.) Needs-Driven Workflow Design
using a modular data acquisition/analysis/modeling/visualization pipeline as well as modular visualization layers.
Implementation in different plug-and-play tools/CIs
2.) **Needs-Driven Workflow Design** using a modular data acquisition/analysis/modeling/visualization pipeline as well as modular visualization layers.


---

**Computational Scientometrics**

**Cyberinfrastructures**

- **Scholarly Database**: 25 million scholarly records
  - [http://sdb.slis.indiana.edu](http://sdb.slis.indiana.edu)

- **VIVO Research Networking**
  - [http://vivoweb.org](http://vivoweb.org)

- **Information Visualization Cyberinfrastructure**
  - [http://iv.slis.indiana.edu](http://iv.slis.indiana.edu)

- **Network Workbench Tool & Community Wiki**
  - [http://nwb.slis.indiana.edu](http://nwb.slis.indiana.edu)

- **Science of Science (Sci²) Tool and CI Portal**
  - [http://sci.slis.indiana.edu](http://sci.slis.indiana.edu)

- **Epidemics Cyberinfrastructure**
  - [http://epic.slis.indiana.edu](http://epic.slis.indiana.edu)
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 February 2009, the tool provides more 169 plugins that support the preprocessing, analysis, modeling, and visualization of networks. More than 50 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 65,000 times since December 2006.


Computational Proteomics

What relationships exist between protein targets of all drugs and all disease-gene products in the human protein–protein interaction network?


Figure 2 Drug-target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are shown in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.
Computational Epidemics
Forecasting (and preventing the effects of) the next pandemic.


Sci² Tool – “Open Code for S&T Assessment”

OSGi/CIShell powered tool with NWB plugins and many new scientometrics and visualizations plugins.

Horizontal Time Graphs

Sci² Tool: Algorithms
See https://nwb.slis.indiana.edu/community

**Preprocessing**
- Extract Top N/\% Records
- Extract Top N Records
- Normalize Text
- Slice Table by Line
- Extract Top Nodes
- Extract Nodes Above or Below Value
- Delete Isolates
- Extract Top Edges
- Extract Edges Above or Below Value
- Remove Self Loops
- Trim by Degree
- MST-Pathfinder Network Scaling
- Fast Pathfinder Network Scaling
- Snowball Sampling (in nodes)
- Node Sampling
- Edge Sampling
- Symmetrize
- Dichotomize
- Multiparite Joining
- Geocoder
- Extract ZIP Code

**Modeling**
- Random Graph
- Watts-Strogatz
- Small World
- Barabási-Albert Scale-Free
- TRL

**Analysis**
- Network Analysis Toolkit (NAT)
  - Node Degree
  - Degree Distribution
  - K-Nearest Neighbor (Java)
  - Watts-Strogatz Clustering Coefficient
  - Watts-Strogatz Clustering Coefficient over K
  - Diameter
  - Average Shortest Path
  - Shortest Path Distribution
  - Node Betweenness Centrality
  - Weak Component Clustering
  - Global Connected Components
  - Extract K-Core
  - Annotate K-Coreness
  - HITS

**Visualization**
- GnuPlot
- GUESS
- Image Viewer
- Radial Tree/Graph (prefuse alpha)
- Radial Tree/Graph with Annotation (prefuse beta)
- Tree View (prefuse beta)
- Tree Map (prefuse beta)
- Force Directed with Annotation (prefuse beta)
- Fruchterman-Reingold with Annotation (prefuse beta)
- DrL (VsOrd)
- Specified (prefuse beta)

**Scientometrics**
- Remove ISI Duplicate Records
- Remove Rows with Multitudinous Fields
- Detect Duplicate Nodes
- Update Network by Merging Nodes

**NEW:**
- Database support for ISI and NSF data.

---

See https://nwb.slis.indiana.edu/community

---

**Textual**
- Burst Detection

**Visualization**
- Horizontal Line Graph
- Circular Hierarchy
- Geo Map (Circle Annotation Style)
- Geo Map (Colored-Region Annotation Style)
- Science Map (Circle Annotation)

**Scientometrics**
- Extract Directed Network
- Extract Paper Citation Network
- Extract Author Paper Network
- Extract Co-Citation Network
- Extract Word Co-Occurrence Network
- Extract Co-Author Network
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network
- Extract Document Co-Citation Network
Mapping Transdisciplinary Tobacco Use Research Centers Publications

Compare R01 investigator based funding with TTURC Center awards in terms of number of publications and evolving co-author networks.

Zass & Börner, forthcoming.

Interactive Science Map of NIH Funding


MEDLINE Publication Output by The National Institutes of Health (NIH) Using Nine Years of ExPORTER Data

Katy Börner, Nianli Ma, Joseph R. Biberstine, Cyberinfrastructure for Network Science Center, SLIS, Indiana University, Robin M. Wagner, Rediet Berhane, Hong Jiang, Susan E. Ivey, Katrina Pearson and Carl McCabe, Reporting Branch, Division of Information Services, Office of Research Information Systems, Office of Extramural Research, Office of the Director, National Institutes of Health (NIH), Bethesda, MD.

Where Are the Academic Jobs? Interactive Exploration of Job Advertisements in Geospatial and Topical Space

Angela Zoss, Michael Connover, Katy Börner (2010)

Visualization of Job Postings

http://cns-nd3.slis.indiana.edu/mapjobs/geo
Computational Scientometrics
Cyberinfrastructures

Scholarly Database: 25 million scholarly records
http://sdb.slis.indiana.edu

VIVO Research Networking
http://vivoweb.org

Information Visualization Cyberinfrastructure
http://iv.slis.indiana.edu

Network Workbench Tool & Community Wiki
http://nwb.slis.indiana.edu

Science of Science (Sci²) Tool and CI Portal
http://sci.slis.indiana.edu

Epidemics Cyberinfrastructure
http://epic.slis.indiana.edu/

Scholarly Database: Web Interface
http://sdb.slis.indiana.edu

Supports federated search of 25 million publication, patent, grant records.
Results can be downloaded as data dump and (evolving) co-author, paper-citation networks.

Register for free access at http://sdb.slis.indiana.edu
Since March 2009:

Users can download networks:
- Co-author
- Co-investigator
- Co-inventor
- Patent citation and tables for burst analysis in NWB.
VIVO: A Semantic Approach to Creating a National Network of Researchers (http://vivoweb.org)

- Semantic web application and ontology editor originally developed at Cornell U.
- Integrates research and scholarship info from systems of record across institution(s).
- Facilitates research discovery and cross-disciplinary collaboration.

Soon:
- Simplify reporting tasks, e.g., generate biosketch, department report.


VIVO Users and Needs

- Faculty/Researchers
  - Customize profile created via feeds; find potential collaborators, “people like me”; discovery via high search rankings; info on activity of colleagues...
- Students
  - Create profiles; easily find mentors + collaborators; locate facilities, events, funding opportunities...
- Administrators
  - Quickly find cross-disciplinary expertise (research area; geography); centralize public data from diverse sources; easily repurpose information for consumers; improve faculty collaboration within or across departments and institutions...
- Funding, donor, legislative agencies
  - Discover projects, grants, expertise (e.g. for review panels; targets for funding)...
- General public
  - Find expertise, learn about research in a region/institution...
VIVO Data Providers & Users

- Eagle-i (“enabling resource discovery” U24 award)
- Federal agencies – NIH (NIH RePORTER), NSF, USDA, ...
- Search Providers – Google, Bing, Yahoo, ...
- Professional Societies – AAAS, ...
- Publishers/vendors – PubMed, Elsevier, Collexis, ISI...
- Semantic Web community – DERI, ...
- Consortia of schools – SURA, CTSA...
- Producers, consumers of semantic web-compliant data
VIVO & Linked Open Data
2010 National VIVO Conference August 12&13, NYC
http://conferences.dce.ufl.edu/vivo

VIVO makes high coverage, high quality data from systems of record
• available online
• for free, and
• in machine readable format.

VIVO ontology is aligned with many existing Web 2.0 and scholarly ontologies to ease interoperability.

http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-07-14_colored.png

Download Data

General Statistics
• 36 publication(s) from 2001 to 2010 (CSV File)
• 80 co-author(s) from 2001 to 2010 (CSV File)

Co-Author Network (GraphML File)

Save as Image (PNG file)

Tables
• Publications per year (CSV File)
• Co-authors (CSV File)

36 publication(s) from 2001 to 2010 *(CSV File)*

80 co-author(s) from 2001 to 2010 *(CSV File)*

<table>
<thead>
<tr>
<th>Year</th>
<th>Count Co-Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>1 Chan C.</td>
</tr>
<tr>
<td>2002</td>
<td>3 Chen C., McManus T., Feing Y.</td>
</tr>
<tr>
<td>2003</td>
<td>2 Chen C., Beynek K.W.</td>
</tr>
<tr>
<td>2004</td>
<td>17 Senapati A., Penumarthi S., Thaku S., Senamurthi R., Maru J.T., Shrin R.M., Mera K., Moor K.A.</td>
</tr>
</tbody>
</table>

Co-author network *(GraphML File)*

Save as Image *(PNG file)*
Publications per year *(CSV File)*, see top file.

Co-authors *(CSV File)*

<table>
<thead>
<tr>
<th>Co-Author</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andenken O.</td>
<td>1</td>
</tr>
<tr>
<td>Andenken N.</td>
<td>1</td>
</tr>
<tr>
<td>Ben-Meel Z.</td>
<td>1</td>
</tr>
<tr>
<td>Blackwell A.</td>
<td>1</td>
</tr>
<tr>
<td>Beynek K.W.</td>
<td>4</td>
</tr>
<tr>
<td>Benavente M.</td>
<td>1</td>
</tr>
<tr>
<td>Bradbury D.</td>
<td>1</td>
</tr>
<tr>
<td>Burkhard R.A.</td>
<td>1</td>
</tr>
<tr>
<td>Chen C.</td>
<td>5</td>
</tr>
</tbody>
</table>
Science is global. World view of VIVO activity.
Web site visits are aggregated at the country level.

Shown are the
- Number of people profiles in the 7 different VIVO installation sites plus CAS and U Melbourne.
- Email contacts by data and service providers as well as institutions interested to adopt VIVO.
- The number of visitors on http://vivoweb.org

Circles are area size coded using a logarithmic scale.
VIVO 1.0 source code was publicly released on April 14, 2010.  
87 downloads by June 11, 2010.  
The more institutions adopt VIVO, the more high quality data will be available to understand, navigate, manage, utilize, and communicate progress in science and technology.

Computational Scientometrics

References


http://www.pnas.org/content/vol101/suppl_1/


http://scimaps.org/atlas
Mapping Science Exhibit – 10 Iterations in 10 years

http://scimaps.org

The Power of Maps (2005)
The Power of Forecasts (2007)
Science Maps for Economic Decision Makers (2008)
Science Maps for Science Policy Makers (2009)
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 72 venues on four continents. Currently at
- NSF, 10th Floor, 4201 Wilson Boulevard, Arlington, VA
- Marston Science Library, University of Florida, Gainesville, FL
- Center of Advanced European Studies and Research, Bonn, Germany
- Science Train, Germany.

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

Illuminated Diagram Display

Questions:
• Who is doing research on what topic and where?
• What is the ‘footprint’ of interdisciplinary research fields?
• What impact have scientists?

Contributions:
• Interactive, high resolution interface to access and make sense of data about scholarly activity.
Science Maps in “Expedition Zukunft” science train visiting 62 cities in 7 months
12 coaches, 300 m long
Opening was on April 23rd, 2009 by German Chancellor Merkel

http://www.expedition-zukunft.de
**Clickstream Map of Science**

This is the first map created from large-scale, worldwide, scholarly usage data. It visualizes the individual intellectual movements from one journal to another in their online navigation behavior.

The MUSE platform provides unique aggregated research data on how users navigate the more than 18,000 fully searchable scholarly journals in the platform. This map is based on the clickstream data of our academic users during a six-month period. It aggregates data from all the journals in MUSE, which are divided into 17 subsources.

The map is an online interactive tool that allows users to zoom in or out and explore the connections of related fields.

---

**Chemical Research & Development Powers the U.S. Innovation Engine**

The Council for Chemical Research (CCR) is a summit for the U.S. Congress and government policy makers with interests in the regulatory, research, and development (R&D) needs of the chemical industry. CCR develops the scientific and public policy agenda to ensure that chemical research and development are well funded and globally competitive through its annual 12-13 day-long, two-phase research strategy. To meet full leverage of typically limited access to policy makers, CCR develops the scientific agenda, as a science partner that helps guide the direction of funding decisions. Policy and public outreach conducted by CCR reflects the needs of chemical research and development in fields that benefit from chemical research and development.

---
