Building a Science Observatory: Research, Tools, and Maps

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

With special thanks to the members at the Cyberinfrastructure for Network Science Center, the NWB team, the Sci2 team, the EpiC team, and the VIVO Collaboration

Network Science Talk at IU
December 8, 2014

NSF Workshop Report on "Knowledge Management and Visualization Tools in Support of Discovery"
Börner, Bettencourt, Gerstein, and Uzzo (Eds.)
(http://vw.cns.iu.edu/cdi2008/whitepaper.html)
published in Dec 2009 argues for a

- A decentralized, free “Scholarly Database” to keep track, interlink, understand and improve the quality and coverage of Science and Technology (S&T) relevant data. (see also page 76 and 77 in Appendix D)

- A “Science Marketplace” that supports the sharing of expertise and resources and is fueled by the currency of science: scholarly reputation. (see page 74 in Appendix D) This marketplace might also be used by educators and the learning community to help bring science to the general public and out of the “ivory tower”. (see page 89 in Appendix D)

- A “Science Observatory” that analyzes different datasets in real-time to assess the current state of S&T and to provide an outlook for their evolution under several (actionable) scenarios. (see page 72 in Appendix D)
“Validate Science [of Science Results and] Maps” to understand and utilize their value for communicating science studies and models across scientific boundaries, but also to study and communicate the longitudinal (1980-today) impact of funding on the science system. (see page 81 in Appendix D)

An easy to use, yet versatile, “Science Telescope” to communicate the structure and evolution of science to researchers, educators, industry, policy makers, and the general public at large. (see page 87 in Appendix D) The effect of this (and other science portals) on education and science perception needs to be studied in carefully controlled experiments. (see page 88 in Appendix D)

“Science of Science” studies are necessary to increase our understanding and support the formation of effective research and development teams. (see page 78 and 82 in Appendix D).

“Success Criteria” need to be developed that support a scientific calculation of S&T benefits for society. (see also page 88 in Appendix D)

A “Science Life” (an analog to Second Life) should be created to put the scientist’s face on their science. Portals to this parallel world would be installed in universities, libraries and science museums. (see page 80 in Appendix D)

Research – Multi-level, mixed methods approach to analyze and forecast S&T
Modeling Science Dynamics using
- multi-level,
- mixed methods, and
- multi-perspective models


Descriptive Models of Science
- Detect advances of scientific knowledge via "longitudinal mapping" (Garfield, 1994).
- Synthesis of specialty narratives from co-citation clusters (Small, 1986).
- Identify cross-disciplinary fertilization via "passages through science" (Small, 1999, 2000).
- Understand scholarly information foraging (Sandstrom, 2001).
- Knowledge discovery in un-connected terms (Swanson & Smalheiser, 1997).
- Determine areas of expertise for specific researcher, research group via "invisible colleges" (note that researchers self definition might differ from how field defines him/her) (Crane, 1972).
- Identify profiles of authors, also called CAMEOS, to be used to for document retrieval or to map an author's subject matter and studying his/her publishing career, or to map the social and intellectual networks evident in citations to and from authors and in co-authorships (White, 2001).
Descriptive Models of Science cont.

- Identification of scientific frontiers [http://www.science-frontiers.com/].
- *ISI’s Essential Science Indicators* [http://essentialscience.com/].
- Import-export studies (Stigler, 1994).
- Evaluation of 'big science' facilities using 'converging partial indicators' (Martin, 1996; Martin & Irvine, 1983).
- Input (levels of funding, expertise of scientists, facilities used) - output (publications, patents, Nobel prices, improved health, reduced environment insults, etc. - influenced by political, economic, financial, and legal factors studies (Kostroff & DelRio, 2001).
- Determine influence of funding on research output (Boyack & Borner, 2002).

- How to write highly influential paper (van Dalen & Henkens, 2001).

---

**The Global 'Scientific Food Web'**


**Contributions:**

Comprehensive global analysis of scholarly knowledge production and diffusion on the level of continents, countries, and cities.

Quantifying knowledge flows between 2000 and 2009, we identify global sources and sinks of knowledge production. Our knowledge flow index reveals, where ideas are born and consumed, thereby defining a global 'scientific food web'.

While Asia is quickly catching up in terms of publications and citation rates, we find that its dependence on knowledge consumption has further increased.
Process Models of Science

Can be used to predict the effects of
- Large collaborations vs. single author research on information diffusion.
- Different publishing mechanisms, e.g., E-journals vs. books on co-authorship, speed of publication, etc.
- Supporting disciplinary vs. interdisciplinary collaborations.
- Many small vs. one large grant on # publications, Ph.D. students, etc.
- Resource distribution on research output.
- ...

In general, process models provide a means to analyze the structure and dynamics of science -- to study science using the scientific methods of science as suggested by Derek J. deSolla Price about 40 years ago.
From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review

Existing (left) and proposed (right) funding systems. Reviewers in blue; investigators in red.
In the proposed system, all scientists are both investigators and reviewers: every scientist receives a fixed amount of funding from the government and discretionary distributions from other scientists, but each is required in turn to redistribute some fraction of the total they received to other investigators.

Current Model is Expensive:
If four professors work four weeks full-time on a proposal submission, labor costs are about $30k [1]. With typical funding rates below 20%, about five submission-review cycles might be needed resulting in a total expected labor cost of $150k. The average NSF grant is $128k per year. U.S. universities charge about 50% overhead (ca. $42k), leaving about $86k.
In other words, the four professors lose $150k-$86k= $64k of paid research time by obtaining a grant to perform the proposed research.

To add: Time spent by researchers to review proposals. In 2012 alone, NSF convened more than 17,000 scientists to review 53,556 proposals.

From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review


Assume

Total funding budget in year y is $t_y$
Number of qualified scientists is $n$

Each year,

the funding agency deposits a fixed amount into each account, equal to the total funding budget divided by the total number of scientists: $t_y/n$.
Each scientist must distribute a fixed fraction, e.g., 50%, of received funding to other scientists (no self-funding, COIs respected).

Result

Scientists collectively assess each others’ merit based on different criteria; they “fund-rank” scientists; highly ranked scientists have to distribute more money.

Example:

Total funding budget per year is 2012 NSF budget
Given the number of NSF funded scientists, each receives a $100,000 basic grant.
Fraction is set to 50%

In 2013, scientist $S$ receives a basic grant of $100,000 plus $200,000 from her peers, i.e., a total of $300,000.
In 2013, $S$ can spend 50% of that total sum, $150,000, on her own research program, but must donate 50% to other scientists for their 2014 budget.

Rather than submitting and reviewing project proposals, $S$ donates directly to other scientists by logging into a centralized website and entering the names of the scientists to donate to and how much each should receive.
From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review


Model Run and Validation:


It uses citations as a proxy for how each scientist might distribute funds in the proposed system.

Dataset: 37M articles from TR 1992 to 2010 Web of Science (WoS) database with 770M citations and 4,195,734 unique author names. The 867,872 names who had authored at least one paper per year in any five years of the period 2000–2010 were used in validation. For each pair of authors we determined the number of times one had cited the other in each year of our citation data (1992–2010).

NIH and NSF funding records from IU’s Scholarly Database provided 347,364 grant amounts for 109,919 unique scientists for that time period.

Simulation run begins in year 2000, in which every scientist was given a fixed budget of $8 = $100k. In subsequent years, scientists distribute their funding in proportion to their citations over the prior 5 years.

The model yields funding patterns similar to existing NIH and NSF distributions.
Different Stakeholder Groups and Their Needs

**Funding Agencies**
- Need to monitor (long-term) money flow and research developments, identify areas for future development, stimulate new research areas, evaluate funding strategies for different programs, decide on project durations, funding patterns.

**Scholars**
- Want easy access to research results, relevant funding programs and their success rates, potential collaborators, competitors, related projects/publications *(research push).*

**Industry**
- Is interested in fast and easy access to major results, experts, etc. Influences the direction of research by entering information on needed technologies *(industry-pull).*

**Advantages for Publishers**
- Need easy to use interfaces to massive amounts of interlinked data. Need to communicate data provenance, quality, and context.

**Society**
- Needs easy access to scientific knowledge and expertise.

---

**Scholars Have Different Roles/Needs**

**Researchers and Authors**—need to select promising research topics, students, collaborators, and publication venues to increase their reputation. They benefit from a global view of competencies, reputation and connectivity of scholars, hot and cold research topics and bursts of activity, and funding available per research area.

**Editors**—have to determine editorial board members, assign papers to reviewers, and ultimately accept or reject papers. Editors need to know the position of their journals in the evolving world of science. They need to advertise their journals appropriately and attract high-quality submissions, which will in turn increase the journal’s reputation and lead to higher quality submissions.

**Reviewers**—read, critique, and suggest changes to help improve the quality of papers and funding proposals. They need to identify related works that should be cited or complementary skills that authors might consider when selecting project collaborators.

**Teachers**—teach classes, train doctoral students, and supervise postdoctoral researchers. They need to identify key works, experts, and examples relevant to a topic area and teach them in the context of global science.

**Inventors**—create intellectual property and obtain patents, thus needing to navigate and make sense of research spaces as well as intellectual property spaces.

**Investigators**—scholars acquire funding to support students, hire staff, purchase equipment, or attend conferences. Here, research interests and proposals have to be matched with existing federal and commercial funding opportunities, possible industry collaborators and sponsors.

**Team Leads and Science Administrators**—many scholars direct multiple research projects simultaneously. Some have full-time staff, research scientists, and technicians in their laboratories and centers. Leaders need to evaluate performance and provide references for current or previous members; report the progress of different projects to funding agencies.
Tools – continuously identify, learn, advance, share code, e.g., via Plug-and-Play Macrosopes


Video and paper are at [http://www.scivee.tv/node/27704](http://www.scivee.tv/node/27704)
Designing “Dream Tools”

Many of the best micro-, tele-, and macrosopes are designed by scientists keen to observe and comprehend what no one has seen or understood before. Galileo Galilei (1564–1642) recognized the potential of a spyglass for the study of the heavens, ground and polished his own lenses, and used the improved optical instruments to make discoveries like the moons of Jupiter, providing quantitative evidence for the Copernican theory.

Today, scientists repurpose, extend, and invent new hardware and software to create “macrosopes” that may solve both local and global challenges.

CNS Macroscope tools empower me, my students, colleagues, and more than 130,000 others that downloaded them.

Macrosopes

Decision making in science, industry, and politics, as well as in daily life, requires that we make sense of data sets representing the structure and dynamics of complex systems. Analysis, navigation, and management of these continuously evolving data sets require a new kind of data-analysis and visualization tool we call a macroscope (from the Greek macros, or “great,” and skopein, or “to observe”) inspired by de Rosnay’s futurist science writings. Macrosopes provide a “vision of the whole,” helping us “synthesize” the related elements and enabling us to detect patterns, trends, and outliers while granting access to myriad details. Rather than make things larger or smaller, macrosopes let us observe what is at once too great, slow, or complex for the human eye and mind to notice and comprehend.
Plug-and-Play Macroscopes

Inspire computer scientists to implement software frameworks that empower domain scientists to assemble their own continuously evolving macroscopes, adding and upgrading existing (and removing obsolete) plug-ins to arrive at a set that is truly relevant for their work—with little or no help from computer scientists.

While microscopes and telescopes are physical instruments, macroscopes resemble continuously changing bundles of software plug-ins. Macroscopes make it easy to select and combine algorithm and tool plug-ins but also interface plug-ins, workflow support, logging, scheduling, and other plug-ins needed for scientifically rigorous yet effective work.

They make it easy to share plug-ins via email, flash drives, or online. To use new plugins, simply copy the files into the plug-in directory, and they appear in the tool menu ready for use. No restart of the tool is necessary. Sharing algorithm components, tools, or novel interfaces becomes as easy as sharing images on Flickr or videos on YouTube. Assembling custom tools is as quick as compiling your custom music collection.

Sharing Algorithms Across Disciplines

Different datasets/formats. Diverse algorithms/tools written in many programming languages.
Related Work

Google Code and SourceForge.net provide special means for developing and distributing software

- In August 2009, SourceForge.net hosted more than 230,000 software projects by two million registered users (265,957 in January 2011);
- In August 2009 ProgrammableWeb.com hosted 1,366 application programming interfaces (APIs) and 4,092 mashups (2,699 APIs and 5,493 mashups in January 2011)

Cyberinfrastructures serving large biomedical communities

- Cancer Biomedical Informatics Grid (caBIG) (http://cabig.nci.nih.gov)
- Biomedical Informatics Research Network (BIRN) (http://birn.bioinformatics.org)
- Informatics for Integrating Biology and the Bedside (i2b2) (https://www.i2b2.org)
- HUBzero (http://hubzero.org) platform for scientific collaboration uses
- myExperiment (http://myexperiment.org) supports the sharing of scientific workflows and other research objects.

Missing so far is a common standard for

- the design of modular, compatible algorithm and tool plug-ins (also called “modules” or “components”)
- that can be easily combined into scientific workflows (“pipeline” or “composition”),
- and packaged as custom tools.

OSGi & CShell

- CShell (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://osgi.org), a standardized, component oriented, computing environment for networked services widely used in industry since more than 10 years.
- Specifically, CShell provides “sockets” into which existing and new datasets, algorithms, and tools can be plugged using a wizard-driven process.
Easy Creation of Custom Tools

Common algorithm/tool pool
Easy way to share new algorithms
Workflow design logs
Custom tools

TexTrend

EpiC
Converters
Sci2
NWB

IS
CS
Bio
SNA
Phys
OSGi/CIShell Adoption

CIShell/OSGi is at the core of different CIs and a total of 169 unique plugins are used in the
- Information Visualization (http://iv.slis.indiana.edu),
- Network Science (NWB Tool) (http://nwb.slis.indiana.edu),
- Scientometrics and Science Policy (Sci2 Tool) (http://sci.slis.indiana.edu), and
- Epidemics (http://epic.slis.indiana.edu) research communities.

Most interestingly, a number of other projects recently adopted OSGi and one adopted CIShell:

Cytoscape (http://www.cytoscape.org) lead by Trey Ideker, UCSD is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data (Shannon et al., 2002).

Bruce visits Mike Smoot in 2009

Taverna Workbench (http://taverna.sourceforge.net) lead by Carol Goble, University of Mancheester, UK is a free software tool for designing and executing workflows (Hull et al., 2006). Taverna allows users to integrate many different software tools, including over 30,000 web services. Micahs, June 2010

MAE+ (https://wiki.ncsa.uiuc.edu/display/MAE/Home) managed by Shawn Hampton, NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research.

TEXTrend (http://www.textrend.org) lead by George Kampis, Eötvös University, Hungary develops a framework for the easy and flexible integration, configuration, and extension of plugin-based components in support of natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component.

As the functionality of OSGi-based software frameworks improves and the number and diversity of dataset and algorithm plugins increases, the capabilities of custom tools will expand.
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

Proportional Symbol Map

How To Read This Map

This proportional symbol map shows 52 U.S. states and other jurisdictions using the Albers equal-area conic projection with Alaska, Puerto Rico, and Hawaii inset. Each dataset record is represented by a circle centered at its geolocation. The area, interior color, and exterior color of each circle may represent numeric attribute values. Minimum and maximum data values are given in the legend.

Total Awards

How To Read This Map

This map is a visual representation of 554 sub-disciplines within 13 disciplines of science and their relationships to one another, shown as points and lines connecting those points respectively. Over top this visualization is drawn the result of mapping a dataset of journals to the underlying sub-disciplines of those journals context. Mapped sub-disciplines are shown with size relative to the number of matching journals and color from the discipline.
Map – effectively communicate the structure and dynamics of science to different stakeholders using (interactive) visualizations.

Mapping Science Exhibit on display at MEDIA X, Stanford University
Map of Scientific Collaborations from 2005-2009


Language Communities of Twitter

Eric Fischer - 2012
Clickstream Map of Science

This is the first map created from large-scale, worldwide, activity-unique data. A visualization of collective data from a particular event in history. The data is from online navigation analyses.


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

The Council for Chemical Research (CCR) recognizes the U.S. government's support of chemical research and development and the critical role of chemical science to our nation's economy and global competitiveness. The Council for Chemical Research (CCR) recognizes the U.S. government's support of chemical research and development and the critical role of chemical science to our nation's economy and global competitiveness.

INVESTMENT IN CHEMICAL SCIENCE R&D

$1 Billion FEDERAL GOVERNMENT

$5 Billion INDUSTRY FUNDING

$1B + $5 Billion CHEMICAL INDUSTRY

$10 Billion CHEMICAL INDUSTRY OPERATING INCOME

$40 Billion GROWTH IN GDP

600,000 JOBS-CREATED

U.S. ECONOMY

20 YEARS TIMELINE FROM CONCEPTION TO COMMERCIALIZATION

The Council for Chemical Research (CCR) recognizes the U.S. government's support of chemical research and development and the critical role of chemical science to our nation's economy and global competitiveness. The Council for Chemical Research (CCR) recognizes the U.S. government's support of chemical research and development and the critical role of chemical science to our nation's economy and global competitiveness.

Illuminated Diagram Display
on display at the Smithsonian in DC.
http://scimaps.org/exhibit_info/#ID
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
Places & Spaces Digital Display in North Carolina State’s brand new Immersion Theater
Science & Technology Forecasts @ Times Square in 2020

This is the only mockup in this slide show. Everything else is available today.

References


All papers, maps, tools, talks, press are linked from [http://cns.iu.edu](http://cns.iu.edu)

These slides will soon be at [http://cns.iu.edu/docs/presentations](http://cns.iu.edu/docs/presentations)

CNS Facebook: [http://www.facebook.com/cnscenter](http://www.facebook.com/cnscenter)

Mapping Science Exhibit Facebook: [http://www.facebook.com/mappingscience](http://www.facebook.com/mappingscience)