



cyberinfrastructure for  
NETWORK SCIENCE CENTER  
School of Library and Information Science | Indiana University Bloomington



## Introducing the Science of Science (Sci2) Tool to the Reporting Branch at the National Institutes of Health

Presented to the Science of Science Measurement Workshop  
December 3, 2010

Robin M. Wagner\*  
Katy Börner\*\*

National Institutes of Health\*  
Indiana University\*\*



## Research Team

Cyberinfrastructure for Network  
Science Center  
School of Library and Information  
Science  
Indiana University

Katy Börner, PhD  
Nianli Ma, MS  
Joseph R. Biberstine, BS

Reporting Branch  
Office of Extramural Research  
Office of the Director  
National Institutes of Health (NIH)

Robin M. Wagner, PhD, MS  
Rediet Berhane, MUPPD  
Hong Jiang, PhD  
Susan E. Ivey, MA  
Katrina Pearson  
Carl McCabe, PhD





## Two Key Contributions

- Discussion of **Socio-Technical Challenges** when introducing science of science tools to an agency
  - What context, insight needs exist?
  - How to select the best tool (and improve it continuously)?
  - How to best transfer expertise—tutorials or close collaboration?
- Answering **Research Questions** with the new tools
  - What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved from 2001-2009?
  - What is the time lag between NIH grant awards being made and papers being published and what is the probability distribution for the number of papers per project?



3



## Two Key Contributions

- Discussion of **Socio-Technical Challenges** when introducing science of science tools to an agency
  - What context, insight needs exist?
  - How to select the best tool (and improve it continuously)?
  - How to best transfer expertise—tutorials or close collaboration?
- Answering **Research Questions** with the new tools
  - What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved from 2001-2009?
  - What is the time lag between NIH grant awards being made and papers being published and what is the probability distribution for the number of papers per project?



4



## Background and Motivation

- Scholars and policy makers have long sought to evaluate the long-term societal impacts of research
- This task is particularly daunting for large portfolios
  - Large portfolios may be linked to thousands of researchers and millions of research outputs, outcomes and impacts, appearing in multiple and often unlinked data sources and databases
  - Data sources may be inconsistent, inaccurate or incomplete
- Increased digitization of scientific information, improved electronic search and linkage tools and capabilities, and new methods and tools have created new opportunities to evaluate large research enterprises

U.S. federal government has mandated,  
*“Agencies should support the development and use of “science of science policy” tools that can improve management of their R&D portfolios and better assess the impact of their science, technology, and innovation investments.”*

Orszag et al., 2010



5



## Available Scientometrics Tools

- Many tools are available to analyze, model, and visualize publication, patent, funding or other science and technology datasets
- Highly specialized tools, e.g.,
  - BibExcel and Publish or Perish support bibliometric data acquisition and analysis
  - HistCite and CiteSpace address specific needs, from studying the history of science to identifying scientific research frontiers
- More general tools, e.g.,
  - Science and Technology Dynamics Toolbox provides many algorithms commonly used in scientometrics research and bridges to other tools
  - Pajek and UCINET are very versatile, powerful network analysis tools widely used in social network analysis
  - Cytoscape is optimized for visualizing biological network data

For review of 20 scientometrics tools, see  
<http://ivl.slis.indiana.edu/km/pub/2010-borner-et-al-nwb.pdf>



6



# Expanding Visualization Tool Capabilities

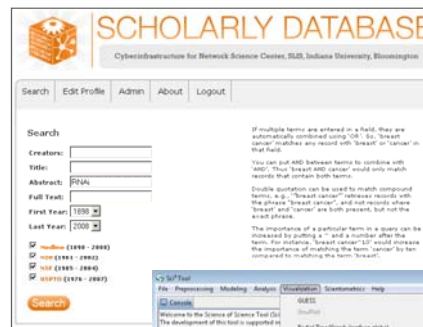
- Cyberinfrastructure for Network Science (CNS) Center
  - Conducts research on structure/dynamics of science for 10 years
  - Curates international Mapping Science exhibit (<http://scimaps.org>)
  - Develops large scale scholarly databases and open source tools to study science by scientific means
- CNS Center has developed the Science of Science (Sci2) and other tools, with significant advantages
  - Based on open source, free software
  - Contain some of the most advanced analysis algorithms
  - Use industry standard, Open Services Gateway Initiative, to build modular software so new algorithms can be easily added by non-computer scientists, tailored to specific agency needs
  - Support data preprocessing, e.g., data cleaning, de-duplication, filtering, and network extraction, essential for high quality analyses
  - Generate easy to read visualizations, many with fixed reference systems, automatic legend design, and audit trail documentation
  - Have extensive publically available documentation



# Using the Scholarly Database and the Sci2 Tool

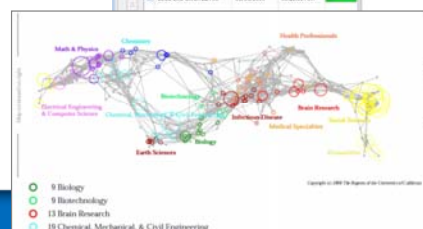
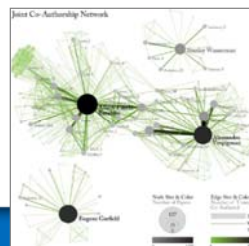
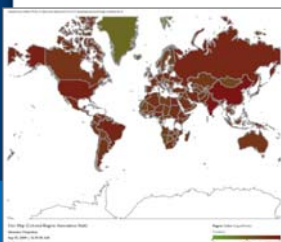
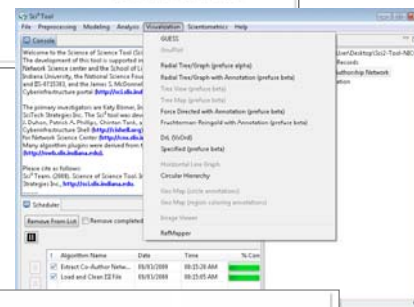
## Scholarly Database

Supports free cross-search and bulk download of 25 million MEDLINE papers, USPTO patents, NSF and NIH awards (<http://sdb.slis.indiana.edu>)



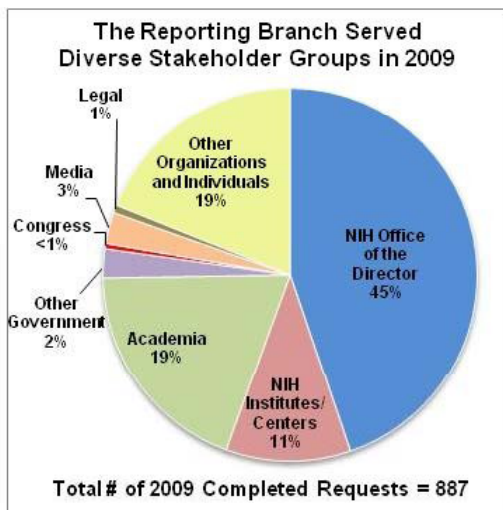
## Science of Science (Sci2) Tool

This NSF SciSIP funded, OSGi/CIShell powered tool with 150+ algorithm plug-ins and is compatible with Epidemics, NWB, and TextTrend.org tools (<http://sci2.slis.indiana.edu>)





## Bringing Sci2 Tool to NIH Reporting Branch



- Branch conducts analyses of NIH-supported research projects and investigators to support NIH policy development and to communicate the impact of NIH's research investment,  $\approx$  \$30 billion/year
- Branch sought new visualization tools to provide new insights into how NIH-supported research and investigators contribute to biomedical knowledge and improving health
- Branch invited Dr. Börner to NIH for one month (July 2010) to provide training and collaborate on research



9



## 12 Tutorials in 12 Days at NIH

- |   |                            |
|---|----------------------------|
| 1. Science of Science Research  | <b>1<sup>st</sup> Week</b> |
| 2. Information Visualization  |                            |
| 3. CShell Powered Tools: Network Workbench and Science of Science (Sci2) Tool |                            |
| 4. Temporal Analysis—Burst Detection  | <b>2<sup>nd</sup> Week</b> |
| 5. Geospatial Analysis and Mapping  |                            |
| 6. Topical Analysis & Mapping   |                            |
| 7. Network Analysis   | <b>3<sup>rd</sup> Week</b> |
| 8. Network Analysis cont.   |                            |
| 9. Extending the Sci2 Tool  |                            |
| 10. Using the Scholarly Database at IU  | <b>4<sup>th</sup> Week</b> |
| 11. VIVO National Researcher Networking                                       |                            |
| 12. Future Developments   |                            |



10



## Questions Federal Agencies Can Answer with Sci2 Tools

- How did the number of grants and total award dollars given to various fields of biomedical science change over time? ([Temporal Analysis](#))
- Where are agency research collaborators located worldwide? ([Geospatial Analysis](#))
- To what degree do agency-funded researchers publish in the areas in which they were funded to do research, and does this differ for more basic versus applied research? ([Topical Analysis](#))
- What are the co-author networks on publications citing agency funding? ([Network Analysis](#))
- In what areas of science does the agency pioneer funding and in which areas does it follow the initial funding by other agencies? ([Scholarly Database](#))



11



## First Post-Tutorial Collaboration: MEDLINE Publication Output by NIH

- Discussion of **Socio-Technical Challenges** when introducing science of science tools to an agency
  - What context, insight needs exist?
  - How to select the best tool (and improve it continuously)?
  - How to best transfer expertise—tutorials or close collaboration?
- Answering **Research Questions** with the new tools
  - What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved from 2001-2009?
  - What is the time lag between NIH grant awards being made and papers being published and what is the probability distribution for the number of papers per project?



12



## Methods

- Extracted public information on NIH grants using electronic tool, Research Portfolio Online Reporting Tools Expenditures and Results (RePORTER) on NIH RePORT website (<http://www.report.nih.gov>)
  - Includes MEDLINE publications whose authors cite NIH grant support and can be linked with automated tool, SPIRES
  - Chose all grants with budget start date in fiscal years 2001-2009 (10/1/2000-9/30/2009) and linked publications published in budget start date year or later (1/1/2001-12/31/2009)
  - For analyses of new grants, applied time lag of 3 months for those awarded in first 3 months of fiscal year (10/1-12/31)
- Evaluated data in 3 time periods examined individually and cumulatively: 2001-2003, 2004-2006, 2007-2009
  - To answer Q1, evaluated number and growth rate of publications linked to all grants by discipline over time, plotted on the University of California, San Diego (UCSD) Map of Science
  - To answer Q2, evaluated time lag between new grant awards and linked publications



13



## Methods (cont.)

- UCSD Map of Science
  - Map based on 7,200,000 publications in 16,000 journals, proceedings and series from Thomson Scientific and Scopus from 2001-2005
  - Contains 554 individual areas of science representing groups of journals comprising 13 major disciplines plus interdisciplinary “Multiple Categories”
  - Publications are plotted on map based on their journal names
  - Advantages
    - Most comprehensive, accurate base map of science at paper level
    - Stable base map enables comparing different analyses generated within or across different agencies
    - Avoids burden of having to create a new semantic topic space for each new analysis



14



## Results

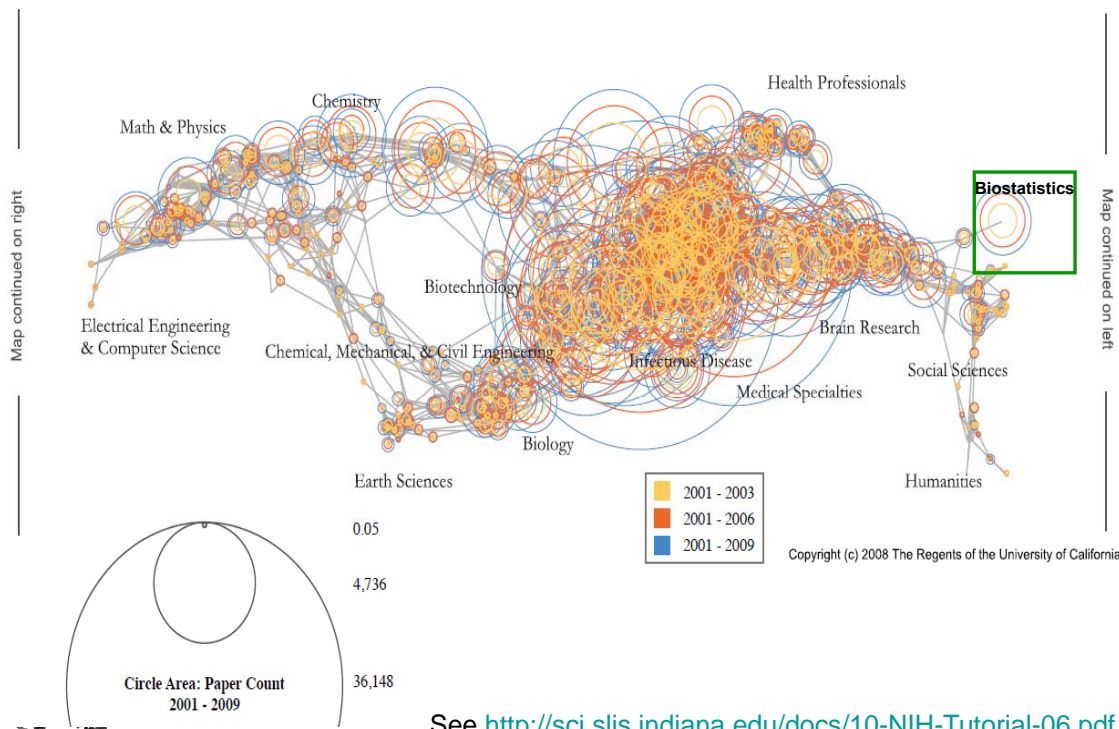
- 147,541 NIH grants (“base projects”) from 2001-2009
- 64% of projects (94,074) had at least 1 linked publication
- After applying time lags, identified 499,322 publications from all grants (Q1)
  - 122,660 papers published 1/1/2001-12/31/2003
  - 171,393 papers published 1/1/2004-12/31/2006
  - 205,269 papers published 1/1/2007-12/31/2009
- From new grant analyses (Q2), identified
  - 171,920 papers published 2001-2009 linked to 2001-2003 grants
  - 104,842 papers published 2004-2009 linked to 2004-2006 grants
  - 27,415 papers published 2007-2009 linked to 2007-2009 grants



15



## Cumulative Growth of Publications Citing NIH Grants Over Time by Scientific Area



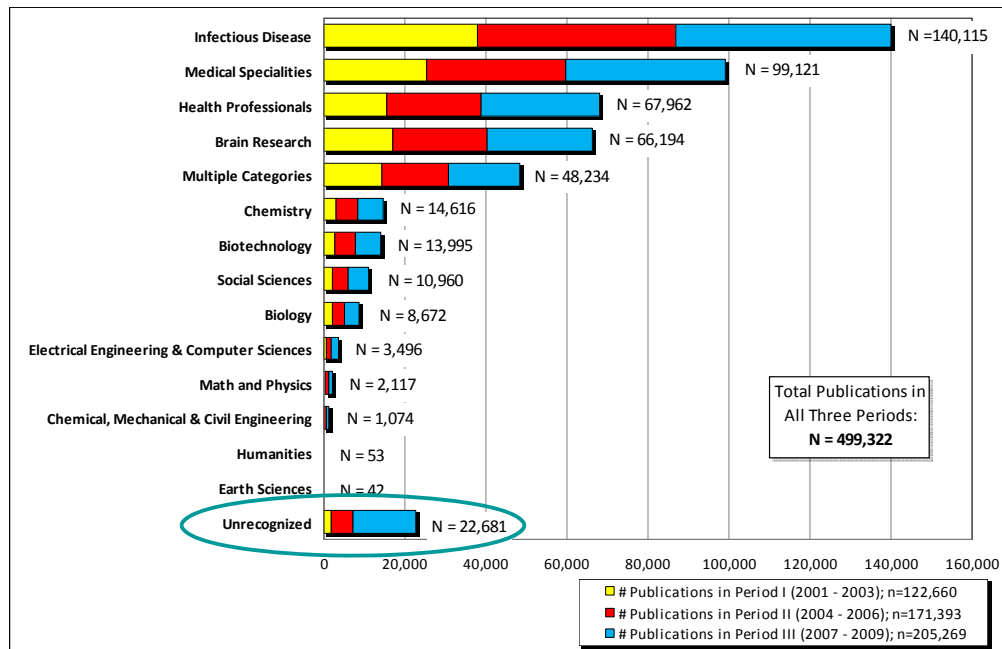
See <http://sci.slis.indiana.edu/docs/10-NIH-Tutorial-06.pdf>

16





# Publications Citing NIH Grant Support by Discipline, 2000-2009



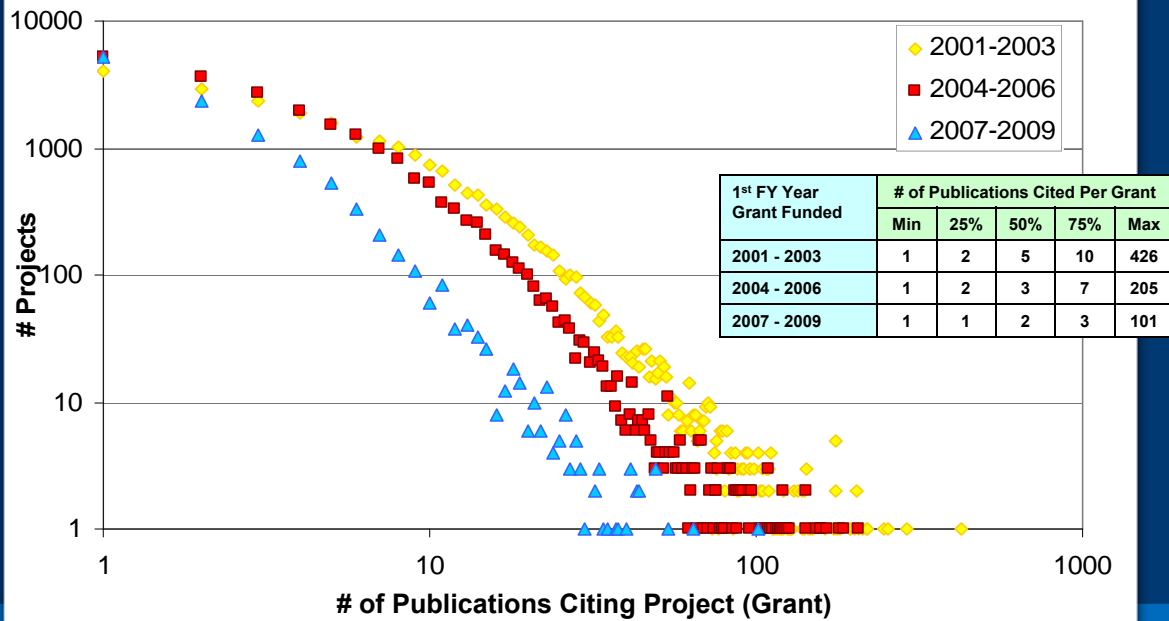
# Publication Growth by Discipline and Time Period

DISCIPLINE	TOTAL # Publications	Growth from Period I to II (2001-2003 to 2004-2006)	Growth from Period II to III (2004-2006 to 2007-2009)	Growth from Period I to III (2001-2009)
<b>Total</b>	<b>499,322</b>	<b>40%</b>	<b>&gt;&gt; 20%</b>	<b>67%</b>
Humanities	53	200%	61%	383%
Chemical, Mechanical, & Civil Engineering	1,074	119%	95%	327%
Math & Physics	2,117	110%	38%	190%
Electrical Engineering & Computer Science	3,496	98%	45%	187%
Social Sciences	10,960	72%	38%	137%
Biotechnology	13,995	77%	31%	132%
Chemistry	14,616	75%	22%	114%
Health Professionals	67,962	52%	24%	89%
Biology	8,672	48%	15%	69%
Medical Specialties	99,121	36%	14%	55%
Brain Research	66,194	36%	11%	52%
Earth Sciences	42	36%	7%	45%
Infectious Diseases	140,115	30%	9%	41%
Multiple Categories	48,234	17%	4%	22%
Unrecognized	22,681	208%	181%	764%





## Publications Citing New NIH Grants Increased with Time from Initial Award



19



## Discussion

- Analyses provide insight into the dynamics of knowledge outputs associated with NIH support
  - NIH leadership can use these results to better understand the behavior of NIH-supported scientists, informing the development of future policies, e.g., NIH public access policy
- Most frequent publications in Infectious Diseases, Medical Specialties, Health Professionals, and Brain Research disciplines coincide well with NIH's large investments in grants in these areas
- NIH's contribution to scientific knowledge, measured by publication outputs, increased over last decade, but growth rate was higher in 2004-2006 than in 2007-2009, compared to preceding time period
  - Likely associated with doubling of NIH budget from 1999-2003, which increased # of grants awarded by NIH, from 43,259 in 2000 to a peak of 52,789 in 2004
  - After 2003, NIH's budget -- and # of annually awarded grants -- remained approximately level, which might account for the slower growth rate of publications in 2007-2009



20



## Discussion (cont.)

- NIH-supported investigators are efficient producers of research knowledge
  - Amongst new grants which generated publications and had enough years of follow-up to observe the majority of publication outputs (5 years), about 2/3 were cited by papers published within the first 3 years of funding
- Limitations
  - UCSD map of science (based on 2001-2005 journals) may not include emerging fields of science, and precludes mapping publications from newer journals (map update is in preparation)
  - More recent grants have not had sufficient follow up time to generate all expected publications
  - Could not ascertain publications with missing, incomplete or incorrect grant number citations



21



## Two Key Contributions

- Discussion of **Socio-Technical Challenges** when introducing science of science tools to an agency
  - What context, insight needs exist?
  - How to select the best tool (and improve it continuously)?
  - How to best transfer expertise—tutorials or close collaboration?
- Answering **Research Questions** with the new tools
  - What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved from 2001-2009?
  - What is the time lag between NIH grant awards being made and papers being published and what is the probability distribution for the number of papers per project?



22



## Suggestions for Introducing Sci2 Skills and Tools to Agencies

- Federal agency (Reporting Branch) perspective
  - Intense tutorial schedule allowed frequent access to tools and resident scholar, but condensed a semester's material into one month, making it challenging to absorb the material due to competing Branch duties and heavy workload
  - More time was needed to learn how to “read” these novel visualizations, e.g., networks, which are unknown to many
  - Other agencies embarking on a similar training arrangement, might consider arranging for a semester sabbatical visit
- Resident scholar perspective
  - 12 days is short time to become acquainted with new colleagues, adapt to a different work culture, obtain security clearance, gain access to and understand internal agency data, and to develop, test, and document new workflows and algorithm plug-ins that address agency-specific needs



## Suggestions for Introducing Sci2 Skills and Tools to Agencies (cont.)

- These tools can be highly useful to agencies that do not opt for intensive training
  - Other governmental agencies and private foundations have started to use the Sci2 Tool
  - As organizations vary on data access, missions, and cultures, each is applying tools to suit its own needs and questions
  - Some agencies have awarded small contracts for developing new specific functionality in the tools, resulting in new plug-ins, many freely shareable with the larger user community, detailed documentation of new functionality and workflows, and dissemination of new insights via peer-reviewed publications
  - Several agencies have independently published peer reviewed papers on insights gained using the new tools





## Questions?

### Dr. Katy Börner

Victor H. Yngve Professor of  
Information Science  
Director, Cyberinfrastructure for Network  
Science Center, and Information Visualization  
Laboratory  
School of Library and Information Science  
Indiana University  
Bloomington, IN

[katy@indiana.edu](mailto:katy@indiana.edu)

<http://info.slis.indiana.edu/~katy>

### Dr. Robin M. Wagner

Chief, Reporting Branch  
Division of Information Services  
Office of Research Information Systems  
Office of Extramural Research  
Office of the Director  
National Institutes of Health  
Bethesda, MD

[wagnerr2@mail.nih.gov](mailto:wagnerr2@mail.nih.gov)

[http://grants.nih.gov/grants/oer\\_offices/reporting  
branch\\_brochure.pdf](http://grants.nih.gov/grants/oer_offices/reporting_branch_brochure.pdf)

This work is funded by the School of Library and Information Science and the Cyberinfrastructure for Network Science center at Indiana University, the National Science Foundation under Grant No. [SBE-0738111](#), and a James S. McDonnell Foundation grant.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



INDIANA UNIVERSITY  
SCHOOL OF LIBRARY  
AND INFORMATION SCIENCE



James S. McDonnell Foundation

