

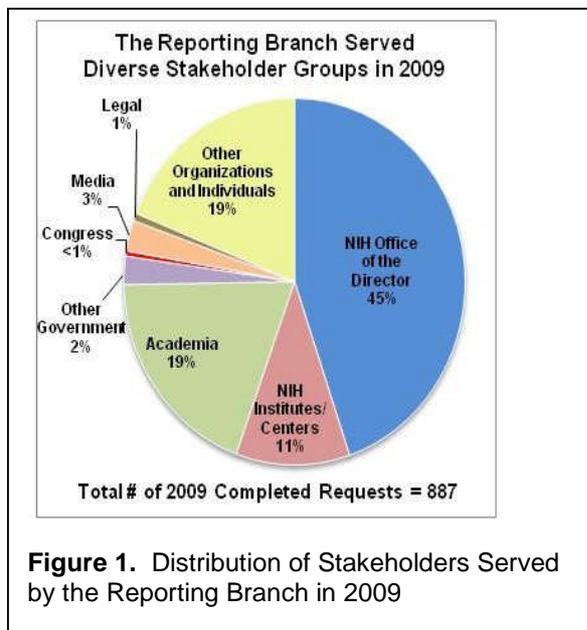
Introducing the Science of Science (Sci2) Tool to the Reporting Branch, Office of Extramural Research/Office of the Director, National Institutes of Health

Katy Börner, Nianli Ma, Joseph R. Biberstine, Cyberinfrastructure for Network Science Center, SLIS, Indiana University, Bloomington, IN 47401

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 20892

Abstract

Recent advances in the availability of high quality and high coverage of science and technology data as well as linkages among previously isolated data silos, e.g., funding and publications, in combination with advanced algorithms and methods to analyze these datasets make it possible to study science in a comprehensive, scientific fashion. This paper presents a set of novel science of science questions, reviews tools that can be readily applied to answer them, and reports challenges and opportunities that arise when introducing new tools to science policy practice. In addition, it presents results from an exemplary analysis that sought to answer two important questions using publically available data: (1) What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved from 2001-2009? (2) 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? We conclude with general suggestions on how to improve the transfer of academic expertise and tools to science policy practice.



Background and Motivation

Scholars and policy makers alike have long endeavored to evaluate the societal impacts of research ranging from individual research projects and programs to federal agency portfolios to the entire national research enterprise. While assessing the long-term impacts of a single research project or small portfolio has proven challenging enough, this task has proven particularly daunting for portfolios of increasing size. Large portfolios may be associated with thousands of researchers and millions of research outputs, outcomes and impacts, appearing in multiple and often unlinked data sources and databases. These data sources may contain inconsistent, inaccurate or incomplete data, further complicating research evaluations.

The rapidly increasing digitization of scientific information, improved electronic search and linkage tools and capabilities, and new methods and tools to synthesize, distill and display enormous amounts of relevant data have created new opportunities to evaluate large scientific research enterprises (Big Data, 2008).

Indeed, the U.S. federal government has recognized these new capabilities and has mandated that executive department and agencies, “should develop outcome-oriented goals for their science, technology, and innovation activities, establish timelines for evaluating the performance of these activities, and target investments toward high-performing programs in their budget submissions. 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). This mandate identified six areas in which agencies should focus their resources in fiscal year 2012 including, “promoting sustainable economic growth and job creation” and “defeating the most dangerous diseases and achieving better health outcomes for all while reducing health care costs.”

In this paper, we present lessons learned and first results achieved when introducing the National Science Foundation (NSF)-funded Science of Science (Sci2) Tool to the Reporting Branch of the Office of Extramural Research at the National Institutes of Health (NIH), the nation’s largest biomedical research agency (NIH, 2010a, 2010b). The NIH Reporting Branch conducts analyses of NIH-supported research projects and investigators to support NIH policy development and to communicate the impact of NIH’s research investments—about \$30 billion annually—to a wide range of audiences, including the extramural research community, Congress, the media, and the public (see Figure 1). The Branch completed 887 requests in 2009, including custom, ad-hoc analyses as well as mandatory, recurrent reports, many of which are published on NIH’s website (see <http://report.nih.gov>). These analyses addressed a wide range of questions including: (1) What is the distribution of awards and dollars by different grant mechanisms (e.g., research centers and individual research projects), NIH funding Institutes and Centers, awardee institutions, geography and Congressional districts? (2) What are the success rates of NIH grant applications and funding rates of NIH-supported investigators, and how do they differ by first time versus experienced investigator status, age, gender, and NIH funding ICs? and (3) What is the effect of NIH training support on subsequent success in obtaining a NIH major research grant award?

The Branch was seeking new tools and methods to make its analyses more accessible to a broader audience through visualizations. The Branch also wanted to learn techniques to provide new insights about how NIH-supported research and investigators contribute to the creation of biomedical knowledge and improve health. In particular, the Branch was interested in new analysis and visualization methods that could answer questions about NIH’s research collaborations and networks; and the inter-relationships between NIH funding, the scientific workforce and its institutions, research projects and topics, research outputs (e.g., publications and patents), and research impacts across different biomedical disciplines, time, and places.

The Branch became aware of Professor Börner’s work and tools in these areas, and invited her to provide training and collaborate on research via the Intergovernmental Personnel Act (IPA) Mobility Program. In July 2010, Professor Börner visited NIH for 12 days, developing and delivering 12 two-hour tutorials to teach the utility and usage of different data analysis and visualizations tool as well as data infrastructures. The tutorials were attended by about 20 individuals, including staff members of the Reporting Branch and other NIH colleagues.

Below, we provide an overview of the tutorial series, discuss promising analyses workflows, present the results of a first joint analysis on research topic coverage and topic evolution of NIH supported research, and list a set of suggestions on how to best transfer expertise and tools developed in academic settings into research policy practice.

12 Tutorials in 12 Days

The Cyberinfrastructure for Network Science (CNS) Center (<http://cns.slis.indiana.edu>) at Indiana University has conducted research on the structure and dynamics of science for 10

years; curates the international Mapping Science exhibit (<http://scimaps.org>); and develops large scale scholarly databases and open source tools for the study of science by scientific means. The CNS Center has adopted a novel approach to tool design, using the OSGi.org, an industry standard, combined with CShell.org to support the easy plug-and-play of algorithms and interfaces. This empowers non-computer science programmers to easily share their algorithms with an interdisciplinary user community, and permits subsets of existing algorithms to be bundled into custom tools. This approach was used to implement the NSF-sponsored *Network Workbench Tool* (NWB) for the study of large scale networks (<http://nwb.slis.indiana.edu>); and the NSF-supported *Science of Science Tool* (Sci2) (<http://sci2.slis.indiana.edu>) optimized for science of science research and practice.

Twelve tutorials were designed and taught in an attempt to communicate science of science research and tools to the Branch. The first three tutorials introduced science of science research, information visualization principles, and modular tools that support the continuous plug-and-play of new analysis and visualization algorithms as they become available.

The subsequent six tutorials used the NWB and the Sci2 Tool. Both tools come with extensive tutorials (Börner et al., 2009; Weingart et al., 2010) and are actively used in science of science research and practice. The tutorials taught and exemplified diverse temporal, geospatial, topical, and network analyses and visualizations of datasets at the micro (individual), meso (local), and macro (global) levels. A main teaching objective aimed to convey differences and commonalities of questions and corresponding analyses and visualizations. For example, “when” questions are commonly addressed via temporal analyses, “where” questions often involve the application of geospatial methods, “what” questions require topical analyses, “with whom” questions are often answered via network studies, and “how” questions help us understand the process by which a certain structure or dynamics came into existence and models are used to replicate and validate these processes. NIH data were used in hands-on sessions together with algorithm plug-ins specifically developed for this tutorial series.

Tutorials 10 and 11 discussed data infrastructures such as the *Scholarly Database* at Indiana University (<http://sdb.slis.indiana.edu>) that supports the federated search of 25 million publication, patent, and grant records (La Rowe, Ambre, Burgoon, Ke, & Börner, 2009) and the NIH-funded *VIVO National Network of Researchers* (<http://vivoweb.org>). The series concluded with an outlook of promising future development and collaboration opportunities in tutorial 12. All slides have been made freely available online at (<http://sci2.slis.indiana.edu>).

Tutorial participants provided structured feedback on the tutorials’ format; engaged in daily exercises on the utility of the presented analyses, visualizations, and workflows for their work; and participated in a brainstorming session and subsequent prioritization of promising analyses.

The comments and feedback provided by tutorial participants resulted in 44 concrete candidate analysis ideas from participants, relevant to understanding NIH’s research investments and impact. The brainstorming session generated an additional 30 analysis ideas for consideration. From the brainstorming session, about a third each of the proposed ideas involved topical (N=11) or network (N=10) analyses, while the remainder were temporal (N=6) or geospatial analyses (N=3); these proposals were nearly evenly divided between enhancements of the Branch’s previous analyses with the new visualization tools (N=16), and totally new research questions for the Branch that were difficult or impossible to answer without these new tools (N=14). These ideas ranged from highly specific inquiries on specific diseases to the broadest questions of NIH’s impact on the scientific community and health. Some examples, which could utilize the different methods taught and be applicable to other science-funding agencies, included:

- How did the number of grants and total award dollars given to various fields of biomedical science change over time? (Temporal Analysis such as burst analysis or

the separate analysis and animated visualization of multiple data time slices, see Tutorial #04)

- Where are agency research collaborators located worldwide? (Geospatial Analysis of the data density in different geospatial regions, data aggregation by congressional district or US state, see Tutorial #05)
- 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 such as term co-occurrence analysis or the simultaneous mapping of topics extracted from solicitations, funded awards, and resulting publications, see Tutorial #06)
- What are the co-author networks on publications citing agency funding? (Network Analysis such as the identification of highly connected nodes (e.g., authors) or nodes that have a bridging, also called gatekeeper role; the identification of information diffusion backbones; scientific communities, see Tutorial #07-09)
- 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 that supports cross-search of different publication, patent, funding databases and the download of data dumps in well documented data formats, see Tutorial #10)

The Reporting Branch and Professor Börner identified two particularly research questions relevant to NIH and conducted analyses to address them, which are presented in below.

First Post-Tutorial Collaborative Analysis: MEDLINE Publication Output by NIH

Research Questions. We sought to answer: (1) What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved from 2001-2009? (2) 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?

Methods. This funding input-paper output analysis is based on publically available data from the Research Portfolio Online Reporting Tools (RePORT) Expenditures and Results (RePORTER, see <http://projectreporter.nih.gov/reporter.cfm>), an electronic tool appearing on the NIH RePORT website (see <http://www.report.nih.gov>). RePORTER allows the public to search a repository of NIH-funded research projects—including grants, contracts and intramural projects—and resultant publications and patents. RePORTER also includes research projects from the Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention, Health Resources and Services Administration, Substance Abuse and Mental Health Services Administration, Food and Drug Administration, and the Department of Veterans Affairs. ExPORTER, online at http://projectreporter.nih.gov/exporter/ExPORTER_Catalog.aspx, enables users to download RePORTER data in bulk including linkage data between NIH *base awards* and MEDLINE publications. The automated linkage is accomplished with an electronic tool, SPIRES (see Jordan, 2007).

Note that *RePORTER Project Data* has full project number records, while *RePORTER Publications Link Tables* contain associated *base projects* to MEDLINE publications. A *base project* is defined as the same funded research project over its lifecycle, regardless of how many award transactions occur for that project. Unlike some other federal agencies such as the NSF, the NIH issues a separate record for each award transaction. In addition, if a grant has subprojects, each will have its own record. Thus, a base project will have many records in RePORTER, e.g., “P” grants are center grants with many subprojects over many years. For NIH grants, each transaction associated with a base project has: 1) a prefix which indicates the type

of award (e.g., new or continuation), 2) a unique identifier, comprised of the grant activity code (e.g., R01), administering NIH IC, and serial number, and 3) a suffix, which indicates the support year and application submission status (e.g., first submission or later resubmission) as well as whether the award is a supplement or a fellowship's institutional allowance. Those grants with subprojects have an additional unique subproject identifier. The unique identifier (described in 2) above) is the component which is used to identify the base project.

The data were further restricted to all NIH extramural grants, including new, renewal and continuation grants, with a budget start date in federal fiscal years of 2001 through 2009. Budget start date rather than award date was selected because the latter was not available for all years in the ExPORTER downloads. The nine years of data were then split into three three-year time periods using a disjoint approach: 2001-2003, 2004-2006, and 2007-2009, and a cumulative approach: 2001-2003, 2001-2006, and 2001-2009.

Publication output for these grants as well as time lags (in years) between the year of new (also known as type 1) grant awards and their resultant publications were examined for these three time slices. The University of California, San Diego (UCSD) Map of Science is used to associate papers with major fields of science. The UCSD base map was created using 7.2 million papers published in over 16,000 separate journals, proceedings, and series from Thomson Scientific and Scopus over the five year period from 2001 to 2005 resulting in a rather stable layout (Klavans & Boyack, 2007). The 554 individual areas of science in this map represent groups of journals that are further grouped into 13 general disciplines such as Mathematics and Physics (for the full list, see Table 1 and Figure 2). Note that there are 13 proper disciplines and one "Multiple Categories" category for journals, e.g., *Science* and *Nature*, which are interdisciplinary enough to fall under the umbrella of more than one discipline.

Using a Sci2 Tool algorithm, a set of papers is "science-located" by matching paper journal names to the 554 science map node journal names and size coding the science map nodes by the number of matching publications. Some journals are interdisciplinary enough to fall under the umbrella of more than one field. In this case, the journal is fractionally assigned to multiple nodes. For example, "Atmospheric Environment" is associated with 56 of the 554 fields. For each occurrence of "Atmospheric Environment", we increase the total for a node by one divided by 56 times a field specific weight for that journal, e.g., "Protein Science" by 0.015625, "Solar & Wind Power" by 0.03125, etc. There is no averaging, only sums.

Rationale for Tool Selection. We used the UCSD Map of Science because it is the most comprehensive and accurate base map of science in existence today. To create a new semantic topic space for each new analysis would be highly burdensome to analysts as they would first need to understand the topic space – the location of math, biology, physics, etc. -- before they could interpret any data overlay. Having a stable base map makes it possible to compare different analyses generated within or across different agencies.

More broadly, the Sci2 Tool was selected because it has significant advantages over other existing tools, which could be of great utility to other federal agencies. It is based on open source, free software, and contains some of the most advanced analysis algorithms available. As noted earlier, it uses OSGi, a major industry standard, to build modular software; new algorithms can be easily added by non-computer scientists, that are tailored to specific agency needs. It supports extensive data preprocessing, including data cleaning, de-duplication, filtering, and network extraction, which are essential for high quality analyses. It generates easy to read visualizations, many with fixed reference systems, automatic legend design, and audit trail documentation (e.g., for each visualization, data sets used and creator). There is extensive publically available documentation on how to use the tool for science of science research and science policy. Although currently only a desktop tool version is available, it could be extended

to be used as a web service. These attractive features have resulted in the tool being actively used in peer reviewed research and by three US agencies/private foundations.

Many other tools are available to analyze, model, and visualize publication, patent, funding or other science and technology datasets; we rejected these because they include some, but not all, of the advantages of the Sci2 Tool. Some tools are highly specialized, e.g., BibExcel and Publish or Perish support bibliometric data acquisition and analysis; HistCite and CiteSpace address specific insight needs -- from studying the history of science to identifying scientific research frontiers. Other tools are more general, e.g., the Science and Technology Dynamics Toolbox provides many algorithms commonly used in scientometrics research along with bridges to other tools. Pajek and UCINET are very versatile, powerful network analysis tools that are widely used in social network analysis. Cytoscape is optimized for visualizing biological network data. For a review of 20 tools that were originally developed in the social science, scientometrics, biology, geography, and computer science, see Börner et al., 2010.

Results. From ExPORTER, we downloaded all *RePORTER Project Data* for fiscal years 2001-2009 (updated on 05/10/2010) as well as *RePORTER Publications* from MEDLINE and *RePORTER Publications Link Tables* (files for fiscal years 2005-2010 and 2000-2004 were last updated on 03/09/2010 and 07/30/2010, respectively). Records from other agencies were excluded, resulting in 174,872 NIH base projects. This dataset was then further restricted to 163,246 extramural grant base projects, omitting intramural projects, inter-agency agreements, and contracts. Next, the dataset was restricted to 147,625 base projects with a budget start date of October 1, 2000 to September 30, 2009, corresponding to federal fiscal years of 2001 through 2009. Cross validation of the 147,625 base projects with internal data available to the NIH Reporting Branch revealed that 535 full project numbers (corresponding to 84 out of 147,625 base projects) could not be found in the NIH internal database and 701 full project numbers (corresponding to 147 base projects) from the NIH internal data could not be found in the NIH RePORTER data. The great majority of these mismatches were due to changes to the award records after the ExPORTER downloads were created (personal communication with Dr. James Onken, NIH, 10/05/2010). After excluding the 84 unmatched projects, the total number of base projects used in subsequent analyses of all grants was 147,541.

A total of 94,074 or 64% of these 147,541 base projects had 976,572 links to 543,440 MEDLINE papers. These papers were designated to publication year (publication month or day is not available for all publications). For each base project, only publications that had a publication year identical or later than its budget start date were included. For base projects with a budget start date between October 1-December 31, 2000, only papers published in 2001 and later were included. In addition, for new type 1 awards whose budget start dates fell between October 1-December 31, only publications beginning in the next calendar year and beyond were included (e.g., only publications from January 1, 2006 and later would be included for new awards with budget start dates of October 1-December 31, 2005). We applied a time lag for new projects awarded in the last quarter of the year because we assumed most publications from the same calendar year were published earlier in that year. The resulting 499,322 MEDLINE papers we evaluated were all published between January 1, 2001 to December 31, 2009. For these publications, there were 901,263 base project-MEDLINE linkages. We refer to this dataset as NIH-9FY.

To answer **(1) What fields of science are covered by publications that acknowledge NIH extramural grant funding and how have the fields evolved over 2001-2009?**, we divided the dataset into the following five time subsets, corresponding to the three study time periods, evaluated separately and cumulatively:

Name	Duration in Years (Dates)	# of Papers
I	01-03 (January 1, 2001 to December 31, 2003)	122,660
II	01-06 (January 1, 2001 to December 31, 2006)	294,053
III	01-09 (January 1, 2001 to December 31, 2009)	499,322
IV	04-06 (January 1, 2004 to December 31, 2006)	171,393
V	07-09 (January 1, 2007 to December 31, 2009)	205,269

Using datasets I-III, UCSD science map overlays were generated using information on the publication venue of each paper, representing cumulative growth in knowledge generation over the three study time periods. Here, publication venue refers to the outlet in which the paper was published, e.g., a journal name. The Sci2 Tool was applied to generate a topical map for each time slice using the UCSD science map. The UCSD science map overlays, given in Figure 2, show the evolution of biomedical research as represented by the scientific outputs (papers) that acknowledge NIH funding. A circle is drawn at the position of a scientific field, if this field has produced papers. The area size of the circle corresponds to the number of published papers.

The individual areas of science with the largest number of publications citing NIH support were *Clinical Cancer Research* with 36,143 papers followed by *Neuroscience; Molecular & Cellular* (26,467 papers), *Protein Science* (20,875 papers), and *Immunology* (18,882 papers). Some areas such as *BioStatistics* in the top right corner of the map experienced considerable growth as evidenced by the size increase from yellow to red to blue circles for that area (see Figure 2).

Scientific fields were further aggregated into 13 disciplines of science that are labeled and shown in Table 1. Custom code was used to extract numerical values for the number of publications in the 13 major disciplines and the 554 areas of science. Table 1 shows the publication counts as well as changes in the number of publications by major discipline over time in raw numbers and in percentages, and is based on datasets I, IV, and V.

From Table 1, it is evident that the NIH-supported research knowledge output, as measured by the total number of publications, has increased in each successive time period, but the growth rate was higher in 2004-2006 at 40% than in 2007-2009 at 20%, when compared to the immediately preceding time period. This same pattern of peak growth in 2004-2006 followed by diminished growth in the later time period was also observed for each of the 13 specific disciplines as well as the "Multiple Categories" field. It is likely that the doubling of the NIH budget from 1999 to 2003, which greatly increased the number of grants awarded by NIH -- from 43,259 in 2000 to a peak of 52,789 in 2004 -- contributed to the jump in publications in the 2004-2006 time period. After 2003, NIH's budget remained approximately level, as did the number of annually awarded grants, which might account for the slower rate of growth of publications in 2007-2009 (for NIH grant award totals in 2000-2009 see <http://report.nih.gov/FileLink.aspx?rid=569> accessed on 10/07/2010).

Two disciplines represented nearly half of the research outputs: Infectious Diseases included 140,115 or 28% of all publications, followed by Medical Specialties, which had 99,121 or 20% of the total. Health Professionals and Brain Research were associated with approximately the same number of publications over the nine years (67,962 and 66,194, respectively), but rate of growth for the former discipline was substantially higher than that for the latter over time. Multiple Categories, reflecting interdisciplinary research, represented another 48,234 or 10% of all publications. All the remaining disciplines, including the physical and social sciences and humanities, had far lower topic coverage as measured by the total number of publications. However, these low coverage topic areas had some of the greatest overall growth over 2001-2009 (e.g., the growth rate was 383% for the Humanities, 327% for Chemical, Mechanical and Civil Engineering, 190% for Math and Physics, and 187% for Electrical Engineering and Computer Science), and may reflect an increase in transdisciplinary

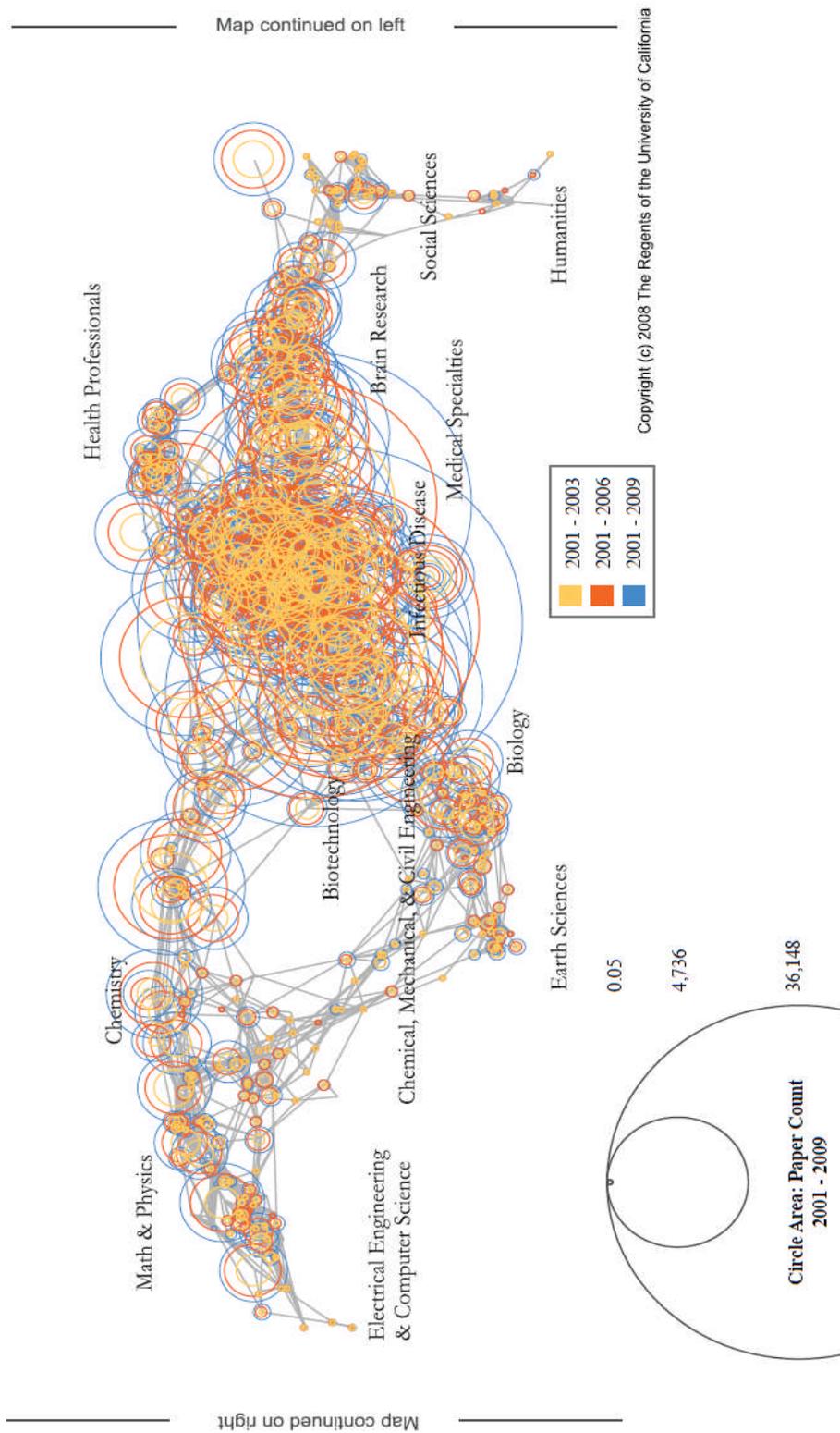


Figure 2: UCSD Map of Science with three data overlays for 2001-2003 (yellow), 2001-2006 (orange), and 2001-2009 (dark red). The largest circle represents 36,143 papers.

Table 1: Changes in the number of publications in the 13 major disciplines from 2001-2009.

Discipline	# Publications (01-03)	# Publications (04-06)	# Publications (07-09)	01-03 to 04-06 growth	04-06 to 07-09 growth	01-03 to 07-09 growth
Biology	2,079	3,073	3,520	48%	15%	69%
Biotechnology	2,750	4,865	6,380	77%	31%	132%
Brain Research	17,040	23,244	25,910	36%	11%	52%
Chemical, Mechanical, & Civil Engineering	144	315	615	%	95%	327%
Chemistry	2,989	5,232	6,395	75%	22%	114%
Earth Sciences	11	15	16	36%	7%	45%
Electrical Engineering & Computer Science	597	1,183	1,716	98%	45%	187%
Health Professionals	15,383	23,431	29,148	52%	24%	89%
Humanities	6	18	29	%	61%	383%
Infectious Diseases	37,825	49,002	53,288	30%	9%	41%
Math & Physics	353	741	1,023	%	38%	190%
Medical Specialties	25,351	34,409	39,361	36%	14%	55%
Social Sciences	2,154	3,697	5,109	72%	38%	137%
Multiple Categories	14,196	16,680	17,358	17%	4%	22%
Unrecognized	1,782	5,498	15,401	%	%	764%
Total	122,66	171,39	205,26	40%	20%	67%

biomedical research.

Of note, a non-trivial number of publications could not be mapped to any discipline, especially those appearing in the 2004-2006, and 2007-2009 literature. This was due, in part, to publications in new journals established after 2005, the last year of journals included in the UCSD science map. For example, 1816 *PLoS One*, 1670 *Biochem Biophys Acta*, 759 *Conf Proc IEEE Eng Med Biol Soc*, 453 *PLoS Genet*, and 439 *PLoS Pathog* could not be science-located on the UCSD science map.

To answer **(2) 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?**, we performed the following analysis.

The NIH-9FY dataset was restricted to new grant (Type 1) awards made in fiscal years 2001-2009 (October 1, 2000 to September 30, 2009). For each new grant, we identified associated MEDLINE papers that occurred in the year of earliest budget start date and subsequent calendar years through 2009 with one exception: for the subset of new awards whose budget

start date fell in the last three months of the calendar year (October 1-December 31). We excluded publications from that year because most publications from the same calendar year were presumed to have been published prior to last quarter of the year (in other words, we applied a time lag of one year for all projects awarded October 1-December 31). This nine year dataset was the again split into three time periods covering new awards made in fiscal years 2001-2003 (resulting in 171,920 papers through 2009), 2004-2006 (resulting in 104,842 papers through 2009), and 2007-2009 (resulting in 27,415 papers through 2009). For each time period, we retrieved the number of papers produced per new project (i.e., new award), and plotted the results (see Figure 3).

Figure 3 clearly shows that grants made earlier had more time to generate papers. For the oldest projects (new awards made in 2001-2003), the median (25%, 75%) of publications generated was 5 (2, 10), with a minimum of one and maximum of 426 publications. For the next oldest projects (new awards made in 2004-2006), the median (25%, 75%) of publications produced was 3 (2, 7), with a minimum of one and maximum of 205 publications. The youngest projects (new awards made in 2007-2009), generated far fewer publications with a median (25%, 75%) of 2 (1, 3), a minimum of one and maximum of 101 publications.

A further examination of how many years it took new awards to generate their first publication revealed a remarkably consistent pattern. For each of fiscal years 2001 to 2003, approximately 11% of new grants made that year were cited by their first paper in first year of funding, about another 30% of new grants were cited by their first paper in the second year of funding (which was the peak or mode of the distribution) while still another 23% were cited by their first paper in the third year of funding, with the remainder of first citations made in later years. A similar pattern was observed for new grants made in later fiscal years, with the second year of funding always being the most frequent in which a first publication appeared in the literature (except for new 2009 awards which had only one year of follow-up). This pattern may further explain why the peak growth in publications occurred in 2004-2006, immediately after the doubling of the NIH budget, which ended in 2003.

Paper Distribution for Type 1 Projects

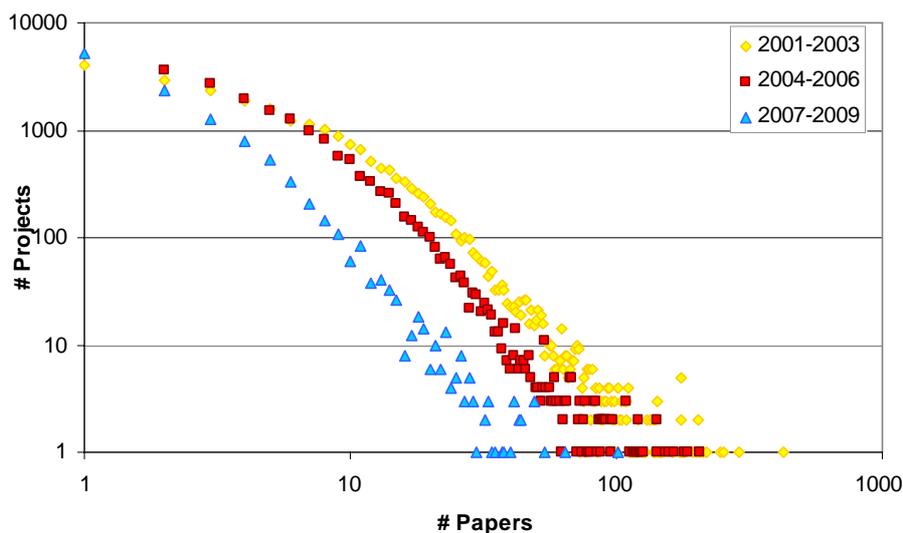


Figure 3: Number of papers per base project for new (Type 1) awards made in fiscal years 2001-2003 (yellow), 2004-2006 (red), and 2007-2009 (blue). Number of papers and projects shown on a log scale.

Discussion. These initial analyses provide insight into the research topics of the knowledge generated by NIH's extramural grant support over a nine year period. The areas of greatest topic coverage—Infectious Diseases, Medical Specialties, Health Professionals and Brain Research—coincide well with NIH's large investments in research grants focusing on infectious diseases and related topics (including HIV/AIDs and vaccine research), clinical research and clinical trials, and neurosciences, brain disorders and neurodegenerative diseases as described by NIH's Research, Condition and Disease Categories system, which is based on a sophisticated text mining tool (see <http://report.nih.gov/rcdc/categories>).

Our findings also demonstrate that NIH's contribution to scientific knowledge, as measured by number of publications, has been increasing over the last decade. Our temporal analyses suggest that this growth is likely related to the substantial increase in grant awards associated with the doubling of the NIH budget over 1999-2003, which slowed down with the subsequent leveling of grant awards made in more recent years.

Our analyses also begin to shed light on the efficiency of knowledge production from NIH grant support. NIH-supported investigators quickly generate publications they link to their grant support. Amongst new grants which generated any publications and had sufficient years of follow-up to observe the majority of publication outputs (5 years), we found that approximately two-thirds of them were cited by papers published within the first three years of initial project funding. While it is possible that some of these early citations reflect work conducted prior to receiving support for the cited grant, these data suggest NIH grantees are highly productive.

Taken together, these initial results provide a baseline from which to conduct more detailed studies that may provide greater depth and insight into NIH's research enterprise and its impact on improving health. Future analyses can also be compared to similar research being conducted by others. As an example, our work complements the topic maps based on NIH-funded extramural grants already created and supported by NIH's National Institute of Neurological Diseases and Stroke and National Institute of General Medical Science (see <http://www.nihmaps.org>).

This study has some limitations. As noted above, the UCSD science map is based on journals in existence from 2001-2005, and thus may not include the most recent emerging fields of science. Also, the more recent grants in our study period have not had sufficient follow up time to generate all of the expected publications suggested by our temporal analyses of the entire nine year period. While we considered earlier start and end dates for our study period to allow all grants to have at least five years of post-funding follow-up time, we ultimately rejected this strategy because this would introduce another problem: it is generally accepted that publications from earlier years were less likely to cite their grant support. Furthermore, publications that could not be linked to their NIH grant support due to incomplete or incorrect grant number citations, were not ascertained for our analyses.

Suggestions for Introducing Science of Science Skills and Tools

We hope other federal agencies seeking to adopt similar tools may benefit from our experience. We learned that the successful transfer of these tools from academia to science policy practice via on-site training depends not only on how well the tools can address specific organizational needs, but also on the process used to impart this new knowledge.

From the perspective of the visiting scholar (Dr. Börner), twelve days were very little time to get to know new colleagues and their expertise; to obtain security clearance, to gain access, and to understand internal agency data (familiarity with the data is key to developing and testing meaningful tools and workflows, but data can be highly sensitive); to develop, test, and document new workflows that address the specific needs of the Reporting Branch; to specify, implement, and start using new algorithm plug-ins developed on a short timeline at the CNS

Center (new algorithms become available continuously, tool development is iterative); and to adapt to different work culture and practice.

From the perspective of a federal agency unit (the Reporting Branch), the intense tutorial schedule had significant advantages but also some drawbacks. The advantages included near daily exposure to the tools and hands-on practice during the tutorial sessions, and immediate, local access to the instructor/scholar while in residence at NIH. Feedback solicited from the tutorial participants was very effectively used by the instructor/scholar in real-time to improve the subsequent classes. However, the tutorials compressed a semester's worth of material into a month. Each tutorial presented completely new material and tools, unfamiliar to most participants, leaving little time to completely absorb what had just been learned. To master these new tools, significant time outside of class was needed to study the theories and their practical applications. In particular, more time was needed to learn how to "read" and recognize significant patterns and other information contained in these novel visualizations, such as tree maps or social network analyses, which are unknown to many researchers. Because the Reporting Branch has a heavy, and often unpredictable workload in responding to frequent, urgent requests, it proved challenging to spend as much time as desired outside of the classroom to practice using the tools during the month Dr. Börner was at NIH. Other agencies embarking on a similar training arrangement, might consider arranging for a semester sabbatical visit, if feasible.

Despite these limitations, the Reporting Branch staff has begun to apply these visualization tools in its work, thanks in part to the excellent archive of training materials provided, interest in the Branch and from another NIH IC in conducting joint projects, and the continued collaboration with Dr. Börner and her research team after she returned to Indiana University. Dr. Börner's team has implemented an "Ask an Expert" web site where Sci2 Tool information requests can be submitted and processed in a structured way. This way, a formal process is in place to request expert comments on a planned, in progress, or completed analysis; to inquire about specific algorithms or workflows; or to get estimates for new algorithm plug-in, e.g., a reader for a new data format.

While in-depth tutorials can facilitate knowledge transfer, we believe these tools can be highly useful to organizations that do not opt for intensive training. As noted earlier, other governmental agencies and private foundations have started to use the NWB and Sci2 Tools. As organizations vary with respect to access to data, missions, and cultures, each is applying these tools to suit its own needs and questions. Some agencies have simply registered and downloaded the tool(s), and have used them in their decision making, without additional assistance from the tools' developers. In other instances, agencies have requested feedback on or assistance with their analyses. Some agencies have awarded small contracts for developing new specific functionality in the tools, e.g., identifying the research topics covered by interdisciplinary funding proposals and grouping them by topical similarity. These contracts have resulted in new plug-ins (many of which can be freely shared with the larger user community), detailed documentation of new functionality and workflows, and dissemination of newly acquired insights through peer-reviewed publications. Independently, representatives from several agencies have successfully published on their tool use in peer reviewed papers (e.g., Bruer, 2010; and Kosecki et al., forthcoming).

We look forward to testing and improving existing, new, and planned extensions of the tools and analysis workflows, and to further improve the transfer of expertise and tools from academia to science policy practice.

Acknowledgements

We are grateful for the generous contributions of NIH and Indiana University which made this work possible. All Reporting Branch staff (see http://grants.nih.gov/grants/oer_offices/reporting_branch_brochure.pdf) participated in the tutorials and proposed many of the analyses described here. Staff from other NIH offices also participated in the tutorials and contributed research ideas, including Drs. Edmund Talley, James Onken, Israel Lederhendler, and Geetha Senthil. Ms. Leyla Desmond provided superb administrative support for the tutorials. For this paper, Reporting Branch members (Drs. Robin Wagner, Hong Jiang, and Carl McCabe and Ms. Katrina Pearson, Rediet Berhane, and Susan Ivey) collaborated with Dr. Börner to define the research questions, conduct analyses, interpret the data and results and write this manuscript. Joseph R. Biberstine implemented custom code. Michael J. Stamper helped with the graphic design. Nianli Ma did most of the data manipulation and analysis with input by Dr. Börner and the Reporting Branch authors.

We thank the NIH Division of Information Services and its Director, Dr. Richard Ikeda, for funding Dr. Börner's IPA to bring her to NIH and develop and deliver the tutorials. Dr. Börner's team's work on this project was also funded in part by the NSF under grant IIS-0715303 and the NIH under grants SBE-0738111 and CBET-0831636. 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 NSF.

References

- (2008). Big Data (Special Issue). *Nature*, 455 (7209):1-136.
- Börner, Katy, Huang, Weixia, Linnemeier, Micah, Duhon, Russell J., Phillips, Patrick, Ma, Nianli, Zoss, Angela M., Guo, Hanning Guo, and Price, Mark A. (2010). Rete-Netzwerk-Red: Analyzing and visualizing scholarly networks using the Network Workbench Tool. *Scientometrics*, 83:863–876.
- Börner, Katy, Angela Zoss, Hanning Guo, Scott Weingart, Ann McCranie, Mark A. Price. (2009). Network Workbench Tool: User Manual 1.0.0. <http://nwb.slis.indiana.edu/doc.html> (accessed on 09/30/2010).
- Bruer, John. (2010) Can we talk? How the cognitive neuroscience of attention emerged from neurobiology and psychology, 1980-2005. *Scientometrics*, 83(3):751-764.
- Jordon, P. (2007). Scientific Publication Information Retrieval & Evaluation System (SPIRES): Linking NLM PubMedData to NIH Extramural Grant Data. http://deais.nci.nih.gov/documents/NIH_SPIRES_overview_Nov2007.pdf. (accessed on 11/8/2010).
- Klavans, Richard, Kevin W. Boyack. (2007). Is There a Convergent Structure to Science? Paper presented at the Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Daniel Torres-Salinas & Henk F. Moed (Eds.), Madrid: CSIC, pp. 437-448.
- Kosecki, S., R. Shoemaker, R., and Kirk-Baer, C. (accepted) Scope, Characteristics, and Use of the U.S. Department of Agriculture's Intramural Research. *Scientometrics*.
- La Rowe, Gavin, Sumeet Adinath Ambre, John W. Burgoon, Weimao Ke, Katy Börner. (2009). The Scholarly Database and Its Utility for Scientometrics Research. *Scientometrics*, 79(2), 219-234.
- National Institutes of Health. (2010a). Office of Extramural Research: Division of Information Services (DIS). http://grants.nih.gov/grants/oer_offices/suboris_dis.htm (accessed on 09/30/2010).

- National Institutes of Health. (2010b). The Reporting Branch, Division of Information Services Brochure. http://grants.nih.gov/grants/oer_offices/reporting_branch_brochure.pdf (accessed on 09/30/2010).
- Orszag. Peter R, Director, Office of Management and Budget, and Holdren, John. P., Director, Office of Science Technology Policy. (July 21, 2010). Memorandum for the Heads of Executive Departments and Agencies, M-10-30, Science and Technology Priorities for the FY 2012 Budget, http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-30.pdf (accessed on 11/05/2010).
- Weingart, Scott, Hanning Guo, Katy Börner, Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick A. Phillips, Chintan Tank, Joseph Biberstine. (2010). Science of Science (Sci2) Tool User Manual. http://sci.slis.indiana.edu/registration/docs/Sci2_Tutorial.pdf (accessed on 10/06/2010).