Published in final edited form as:
Nat Methods. 2011 June ; 8(6): 443–444. doi:10.1038/nmeth.1619.

Database of NIH grants using machine-learned categories and graphical clustering

Edmund M Talley[1], **David Newman**[2], **David Mimno**[3,6], **Bruce W Herr II**[4], **Hanna M Wallach**[3], **Gully A P C Burns**[5], **A G Miriam Leenders**[1], and **Andrew McCallum**[3]

[1]National Institute of Neurological Disorders and Stroke, Bethesda, Maryland, USA

[2]University of California, Irvine, Irvine, California, USA

[3]University of Massachusetts, Amherst, Amherst, Massachusetts, USA

[4]ChalkLabs, Bloomington, Indiana, USA

[5]Information Sciences Institute, University of Southern California, Marina del Rey, California, USA

To the Editor

Information on research funding is important to various groups, including investigators, policy analysts, advocacy organizations and, of course, the funding agencies themselves. But informatics resources devoted to research funding are currently limited. In particular, there is a need for information on grants from the US National Institutes of Health (NIH), the world's largest single source of biomedical research funding, because of its large number of awards (~80,000 each year) and its complex organizational structure. NIH's 25 grant-awarding Institutes and Centers have distinct but overlapping missions, and the relationship between these missions and the research they fund is multifaceted. Because there is no comprehensive scheme that characterizes NIH research, navigating the NIH funding landscape can be challenging.

At present, NIH offers information on awarded grants via the RePORTER website (http://projectreporter.nih.gov). For each award, RePORTER provides keyword tags, plus ~215 categorical designations assigned to grants via a partially automated system known as the NIH research, condition and disease categorization (RCDC) process (http://report.nih.gov/rcdc/categories). But keyword searches are not optimal for various information needs and analyses, and the RCDC categories are only intended to meet specific NIH reporting requirements, rather than to comprehensively characterize the entire NIH research portfolio.

To facilitate navigation and discovery of NIH-funded research, we created a database (https://app.nihmaps.org/) in which we use text mining to extract latent categories and clusters from NIH grant titles and abstracts. This categorical information is discovered using

[6]Present address: Princeton University, Princeton, New Jersey, USA.

Competing Financial Interests
B.W.H. II is employed by Chalklabs, LLC, which was contracted for a portion of the work described in the report. D.N. is owner of TopicSeek, LLC, which was also contracted for a portion of the described work.

two unsupervised machine-learning techniques. The first is topic modeling, a Bayesian statistical method that discerns meaningful categories from unstructured text (see Supplementary Methods for references). The second is a graph-based clustering method that produces a two-dimensional visualized output, in which grants are grouped based on their overall topic-and word-based similarity to one another. The database allows specific queries within a contextual framework that is based on scientific research rather than NIH administrative and categorical designations.

We found that topic-based categories are not strictly associated with the missions of individual Institutes but instead cut across the NIH, albeit in varying proportions consistent with each Institute’s distinct mission (Supplementary Table 1). The graphical map layout (Fig. 1) shows a global research structure that is logically coherent but only loosely related to Institute organization (Supplementary Table 1).

We describe four example use cases (Supplementary Data). First, we show a query using an algorithm-derived category relevant to angiogenesis (Supplementary Fig. 1). Unlike standard keyword-based searches, this type of query allows retrieval of grants that are truly focused on a particular research area. In addition, the resulting graphical clusters reveal clear patterns in the relationships between the retrieved grants and the multiple Institutes funding this research. Second, we examine an NIH peer review study section. The database categories and clusters clarify the complex relationship between the NIH Institutes and the centralized NIH peer review system, which is distinct and independent from the Institutes. Third, we show an analysis of the NIH RCDC category ‘sleep research’ in conjunction with the database topics, the latter providing salient categorical information in greater detail than the officially reported category. Finally, we show how the database can be used for unbiased discovery of research trends, and we document the remarkable increase in funding for research on microRNA biology from 2007 to 2009. Changes in topics associated with this burgeoning area demonstrate a transition in the nature of the research, from basic cellular and molecular biology to investigations of complex physiological processes and disease diagnoses.

In each case, the machine-learned topics are robustly correlated with funding by specific NIH Institutes, highlighting the importance of the underlying categories to the NIH. The patterns elucidated in this framework are consistent with Institute policies, but obtaining similar information in the absence of the current database would require extensive exploration of Institute websites, followed by time-consuming research on appropriate keywords for queries of specific categories. Our database offers an alternative approach that enables rapid and reproducible retrieval of meaningful categorical information.

To ensure transparent and accurate representations of the algorithm-derived topics, we provide extensive contextual information derived from the documents associated with each topic, in a format conducive to spot checks and to detailed examination for cases requiring precise categorical distinctions. Additionally, we implemented a new technique for automatically assessing topic quality using statistics of topic word co-occurrence (Supplementary Methods), which we used for curating the database to identify poor quality topics.

Our use of this graphing algorithm is somewhat different from previous gene expression analyses and scientometric studies based on journal citation linkages (see Supplementary Methods for references). We assessed the information-retrieval capabilities of the graphs and found that they performed well relative to the document similarity measures that served as inputs. Notably, rather than forming isolated clusters, in this case the algorithm produced a lattice-like structure, in which clusters are linked by strings of aligned documents whose topical content is jointly relevant to the clusters at either end of each string (Supplementary Fig. 1). In addition to providing extra 'subcluster' resolution of content that falls between clusters, this lattice-like framework formed a logical organizational structure, merging the local, intermediate and global levels of the graph.

The categories and clusters represented in this database are comprehensive and thus provide reference points from which various information requirements can be addressed by users with divergent interests and needs. Perhaps more importantly, they provide a basis for discovery of interrelationships among concepts and documents that otherwise would be obscure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge assistance and support from G. LaRowe and N. Skiba at ChalkLabs, and input and feedback from NIH staff during the project. We thank S. Silberberg, C. Cronin, K. Boyack and K. Borner for helpful advice and comments on the manuscript.

This project has been supported through small contracts from the NIH to:

University of Southern California (271200900426P and 271200900244P)

University of Massachusetts (271201000758P, 271200900640P, 271201000704P and 271200900639P)

ChalkLabs LLC (271200900695P and 271201000701P)

TopicSeek LLC (271201000620P and 271200900637P).

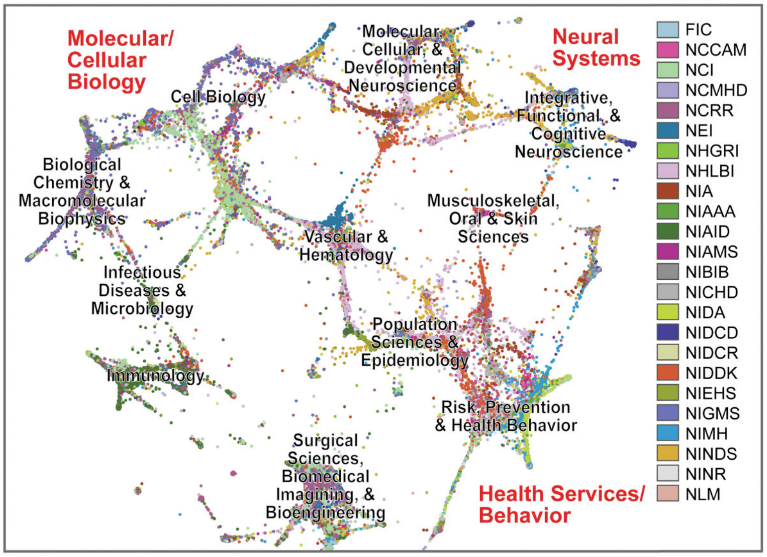

Figure 1. Graphically clustered NIH grants, as rendered from a screenshot of the NIHMaps user interface

NIH awards (here showing grants from 2010; ~80,000 documents) were scored for their overall topic and word similarity, and the resulting document distance calculations were used to seed a graphing algorithm. Grants are represented as dots, color-coded by NIH Institute and are clustered based on shared thematic content. For acronyms and separate views with each Institute highlighted, see the legend for Supplementary Table 1. Labels in black were automatically derived from review assignments of the underlying documents. Labels in red indicate a global structure that was reproducible using multiple different algorithm settings.