Mapping the money

Improving search tools for NIH grants will increase the transparency of US government–sponsored research and aid those seeking funding.

In 2010, 31 billion dollars of the US federal budget were allocated to the 25 Institutes constituting the US National Institutes of Health (NIH), an investment that the research community should exploit to its fullest, not just by applying for grants and learning from publications based on NIH funding but also by directly accessing information on funded research grants. The NIH are adding web tools that increase the accessibility of this information.

In 1972, the NIH launched the CRISP (computer retrieval of information on scientific projects) database, a manually annotated database with subject terms, similar to medical subject headings (MeSH) terms, assigned to each NIH award. It allowed searches of funded research projects, but a lot of staff were needed to maintain it, and results could not be ranked by significance.

In 2006, Congress passed the NIH Reform Act, which included a mandate that the NIH provide a better electronic system to search NIH research projects that applies a uniform process of accounting for NIH funding. CRISP was retired and replaced by the RePORTER database (research portfolio online reporting tools) in 2009.

In RePORTER, NIH replaced the annotated terms of CRISP with the research, condition and disease categorization (RCDC) system, a partially automated process that uses keywords chosen by NIH staff to determine the amount of funding for each of 229 categories at the end of each fiscal year. One-third of these 229 categories, which include research areas, diseases and conditions, are updated each year to ensure proper weightings of the terms.

Compared to CRISP, RePORTER provides better functionality, including links to various information associated with a grant, such as publications. It is a useful tool, but the limited set of RCDC categories used in searches make it still difficult to extract comprehensive and fine-grained information about funding trends and overlapping goals in the 25 different Institutes at NIH.

On May 2, 2011, RePORTER version 2.4 was released with a link to NIHMaps, a database that promises to provide just such detailed information. It uses two unsupervised machine-learning techniques: topic modeling to discover underlying categories in unstructured text, which are independent of RCDC categories, and a graph-based clustering method that groups grants based on similarity (see Correspondence on p 443; http://nihmaps.org/).

This database aims to address the needs of different user groups, from people involved in funding policy to individual investigators. Its goal is to retrieve grants related by topics, independent of preset categories. It shows, for example, how grants that have certain categories in common are distributed over various NIH Institutes, how closely they are related based on the frequency of category combined occurrence and how funding trends have changed between 2007 and today.

This database will allow researchers to more easily spot grants related to their interests and thus identify potential collaborators, or competitors, even though their grants may be funded by different Institutes. And a comprehensive understanding of which Institute is funding particular topics will help provide ideas of where to target a grant. This information also makes it easier to spot holes in a topic that could be filled with a new grant proposal.

Of course, these tools should only be helpful guidance to steer investigators to the appropriate place for their original ideas. We are not advocating a 'follow the money' mentality that induces researchers to gear their work toward a currently hot field and discourages unconventional thinking because no category exists for it.

The increased transparency could also be useful for people at the NIH who make funding decisions to ensure that there are no redundancies in grants and to identify areas in need of financial support.

The database is still in its experimental phase, and one can envision improvements, such as changing the currently static maps into dynamic ones that display the change of topics, and funding allocated to them, over time. It would also be intriguing to visualize the topics alongside demographic information, such as established versus new investigator, gender and geographic location.

A goal for the developers of NIHMaps is the incorporation of more funding agencies. A system that allows one to compare funding trends in different countries would be extremely useful to get a better sense of global expenditure of money and to identify topics that are underfunded.

NIHMaps uses only open-source published algorithms, and these could be applied to text-mining efforts by other organizations, such as journals or publication repositories to visualize publication trends over time or topic emphasis in a given journal. This would complement projects such as Map of Science (http://mapofscience.com/), which provides maps of research efforts in certain institutions or in a particular area of research but cannot be interactively queried.

NIHMaps will not have the details to satisfy everybody, but as a first of its kind it is a good start and community input from different users will improve its features. Try it yourself.