Information Services & Use 34 (2014) 201–207
DOI 10.3233/ISU-140760
IOS Press

Science funding and science policy: Big data as a tool for supporting the research funding process

Christian Herzog
CEO ÜberResearch GmbH, Cologne, Germany
E-mail: christian@uberresearch.com

ÜberResearch pursues a comprehensive approach to big and small data to aid science funders in operations and analysis. We seek to consolidate global award data and funders' internal data, with a focus on data harmonization, natural language processing and disambiguation.

In this article, firstly, I give a definition and some examples of big data. Next, I look at the relevance of big data in science funding. Third, I look at important challenges in the science funding process. Finally, I turn to big data solutions and the ÜberResearch approach in implementing them.

1. Big data definition

The notion of 'big data' does not only signal large volume, but also that the data set or sets are so complex that it is not possible to process them using traditional database management tools and processing

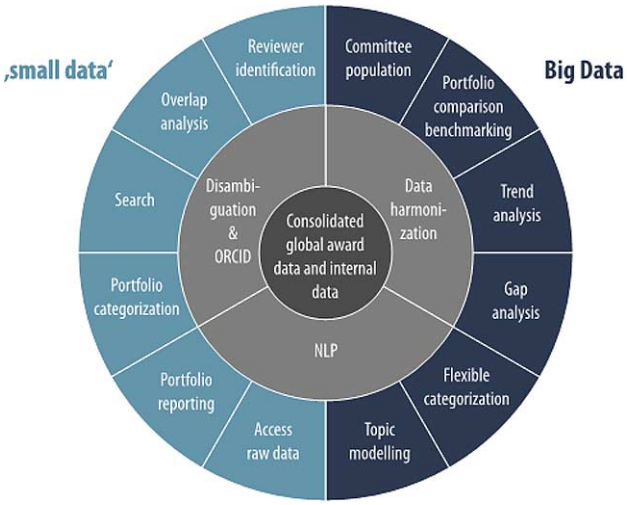

Fig. 1. The ÜberResearch approach. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140760.)

applications. New approaches are required. Moreover, the challenge is not only one of capture, curation and storage, but equally of search, sharing, analysis and visualization.

Interesting and important to me is the focus on creating value from data. Some examples how value is being created from big data:

- Hospitals: Analysis of fifty million records of the UK National Health Service show that patients admitted to hospital on Sundays have a higher mortality risk in the following thirty days. Concomitantly, patients admitted during the week, staying on over the weekend, and discharged the next week, have a lower mortality risk. This led to a discussion if and how expertise and skills must be equally high on all days of the week, including Sundays.
- Logistics: Analysis of delivery routes for UPS trucks revealed that it would be advantageous not to turn left, but to turn right twice. This reduces the risk of accidents and saves fuel: 1.5 m gallons less fuel are consumed annually.
- Retail store staffing: A large German drug store chain uses weather data and internal human resource data to optimize staff planning for the coming four to eight weeks. Employee satisfaction has gone up as unexpected calls to duty have gone down.

To my mind these examples show that the focus must be on exploring how big data approaches create value and deliver new insights.

2. Big data in science funding

Big data in science funding is underdeveloped because large enough pools of data for search, analysis and visualization are scarce. To date, relevant data is mainly available via distributed databases – one silo per science funder. Data sets are largely neither compatible nor harmonized, and funders seldom share them. Given the multitude of research funders, big data analysis needs to move beyond single data silos.

An enabling infrastructure is emerging. ORCID, the research identification and profile system, allows the tracking of researchers and individual research performance across databases and over time. FundRef, the system for reporting funding sources for published results, enables us to link funding with authors and outputs.[1]

More generally, progress is required on research classification schemes. Often these are developed and deployed in-house, possibly manually only. Nevertheless they may be utilized for data analysis of the institutional silo. If the silo is large enough, e.g. a national funder in a larger nation-state with a big budget, approaches and solutions may be interesting. However, this is not the norm, hence the area would benefit if research classifications were all electronic and public.

We have noticed some national initiatives at sharing data, but given the global nature of research, what the funder really needs, is a global database. The funder needs this already to determine that they are not funding a research proposal twice, i.e. somewhere else in the world this research is already underway.

Some examples of big data projects in science funding:[2]

- National Institute of Health: The NIH is the world's largest research funder, and its database may be searched. At NIH Research Portfolio Online Reporting Tools, search results for any keyword will be returned as a list and as map, aiding the identification of research hubs and their interconnection.

[1] http://orcid.org; http://www.crossref.org/fundref/.

[2] http://report.nih.gov; https://g-finder.policycures.org/gfinder_report/; http://gtr.rcuk.ac.uk.

- G-FINDER: This tool for the Global Funding of Innovation for Neglected Diseases is provided by Policy Cures and the Gates Foundation. It pools funding information and enables coordinated approaches in areas currently underfunded.
- Gateway2Research: The Research Councils UK collaborate by sharing their funding information and enabling further analysis via an open API. With due care taken to provide high-quality information in the right format, a new ecosystem of analytic tools for big data in science funding is emerging.

3. Data challenges in funding science

The core of science funding is to receive and evaluate research proposals. Hence, of utmost importance is the question how big data approaches would enhance and change the evaluation of proposals and the review process. When receiving a proposal, funders typically would like to know who funded similar research and what other funders' activities are. Initially important also: Have similar research proposals been rejected in the past, and why? As the proposal is submitted for review, the funder needs to identify the best reviewers, the right standing committee, and assess possible conflicts of interests.

Any research funder builds up a portfolio, and this portfolio is likely to have areas of strength, a focus on certain topics and the like. Hence, funders are interested to understand how their portfolio came about as a result of individual award decisions. What is the impact of the portfolio as well as individual research outcomes?

Based on the portfolio, funders want to compare with other funders: Where do we align? How do we diverge? This includes analysis of gaps and emerging trends. Which important areas of research are underfunded? What new research concepts are emerging? Which are the activities of other funders?

This overview of data challenges reinforces the point made above: funders critically are dependent on gaining insights from each other.

4. Solutions and the ÜberResearch approach

ÜberResearch is working with more than twenty research funders as development partners to drive forward the operational and analytical dimension of big data in science funding.[3] We began aggregating a global grant database, which is shared amongst our partners and customers. The database is built from publicly available award data as well as funders' internal data. Funders' internal data is kept separate and confidential, but added so that the same analytical tools may be utilized. Disambiguation and harmonization augment data quality, while natural language processing enables us to extract relevant information with one consistent tool across all funders.

The joint effort with our partners commenced mid-2013; focusing in the first few months on backend functionality, natural language processing and data curation. After having the basics in place we started with the fun part: to design and implement the analytical layer which transforms the data into insights, which were provided to the partners in September 2014 as 'Dimensions for Funders'.

The joint efforts with our development partners led to cloud based shared solutions, which can be provided at acceptable costs also to very small funders. Smaller funders have a large need for these

[3] http://www.uberresearch.com.

kinds of solutions but little access previously due to high costs. The joint development will continue in 2015 – focusing on the priorities defined by development partners and customers.

The ÜberResearch approach is to provide solutions for small and big data in modules from reviewer identification and overlap analysis through to portfolio benchmarking and committee population.

The functional modules in some detail (see Fig. 1):

- Reviewer identification: Automatic matching between a grant application and potential reviewer candidates to identify the most qualified ones without conflicts of interest (see Fig. 2).
- Overlap analysis: by analyzing the grant applications with natural language processing technology it is possible to show immediately the global funding landscape; which funder did fund similar research, research organization received it and who are the researchers who carried out the research.
- Portfolio categorization: Machine learning for automatic classification.

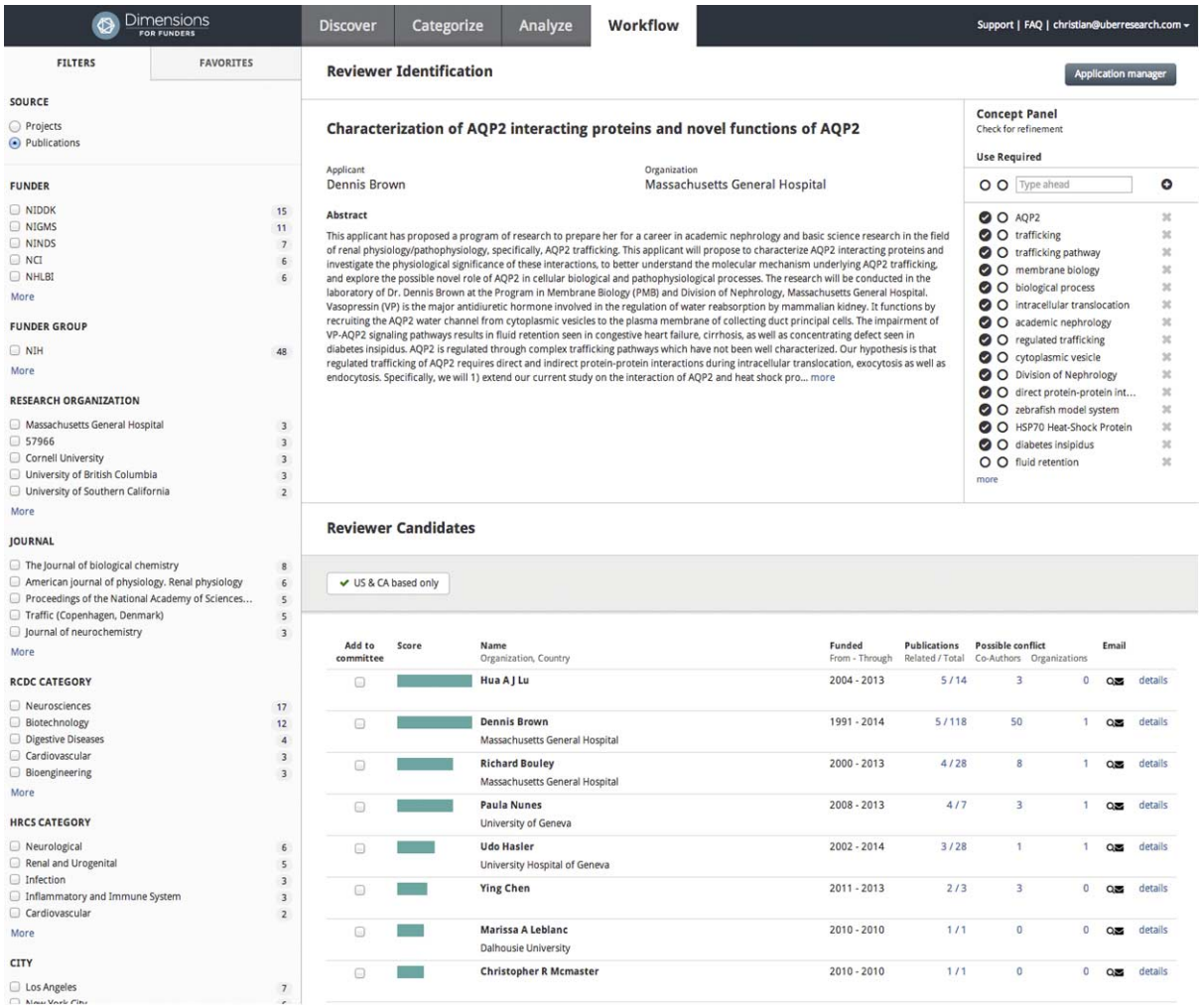

Fig. 2. Automatic reviewer matching with conflict of interest checking. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140760.)

- Portfolio reporting: Standardized and automatic reporting, allowing funders to use natural language processing for precise and automated reporting.
- Portfolio benchmarking: Comparison of selected funders' portfolios, based on the global award database it is possible to compare funders on a portfolio or very granular level.
- Trend analysis: Early understanding of impact of funded research.

Let me give you some visual examples how we implemented support for these use cases in our solution 'Dimensions for Funders':

(A) Reviewer identification: For interested funders we have built an automated workflow that automatically matches incoming proposals with suitable reviewers (see Fig. 2).
(B) Search: For funders in the United States we have built an integrated search interface that show who funded, who received the grant money, including where and how much (see Fig. 3).
(C) Trend analysis: Tracking and comparison of funding activities for selected research funders, including a visualization to highlight areas of high activity and strength (see Fig. 4).
(D) The figure below shows the portfolios of different funders in a visual comparison focused on one category – in this case Tuberculosis (see Fig. 5).

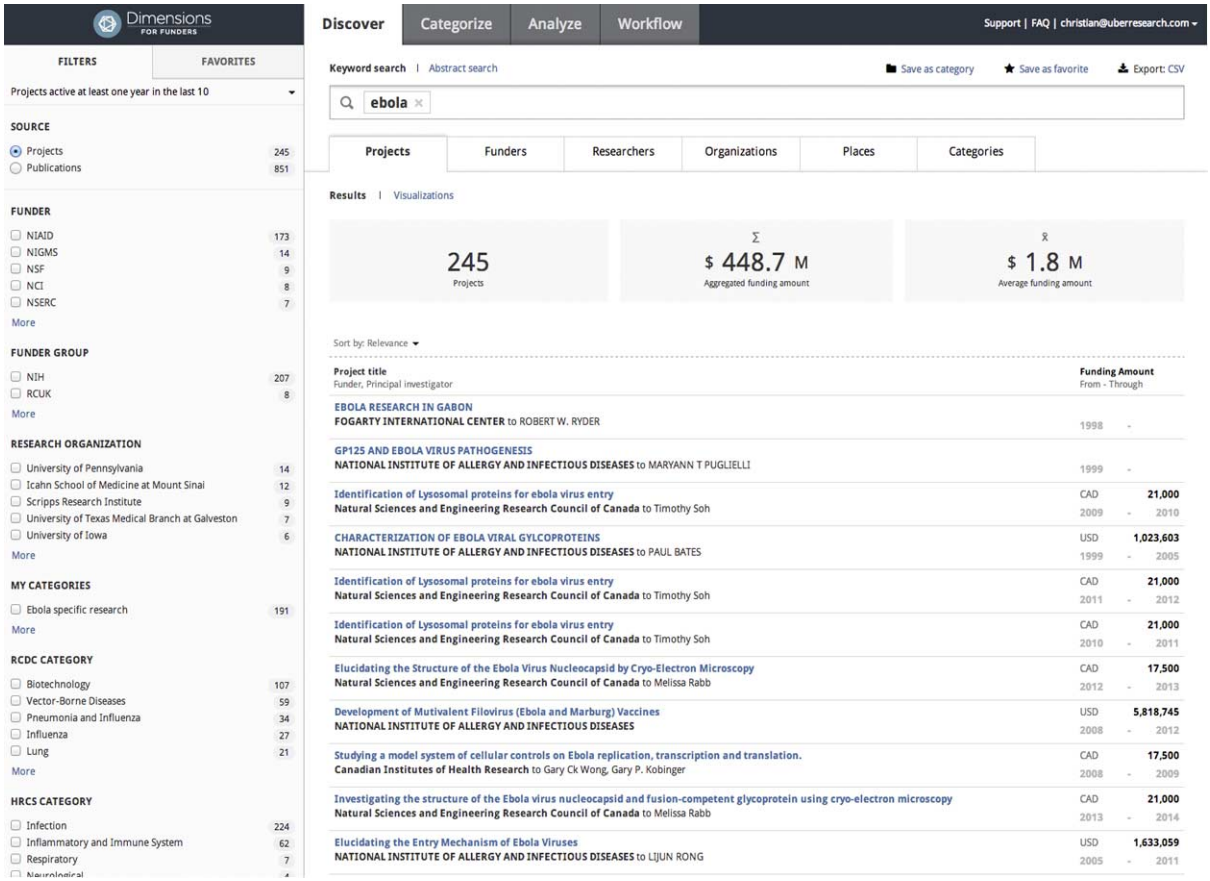

Fig. 3. Search and overview on spending in the area of the search. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140760.)

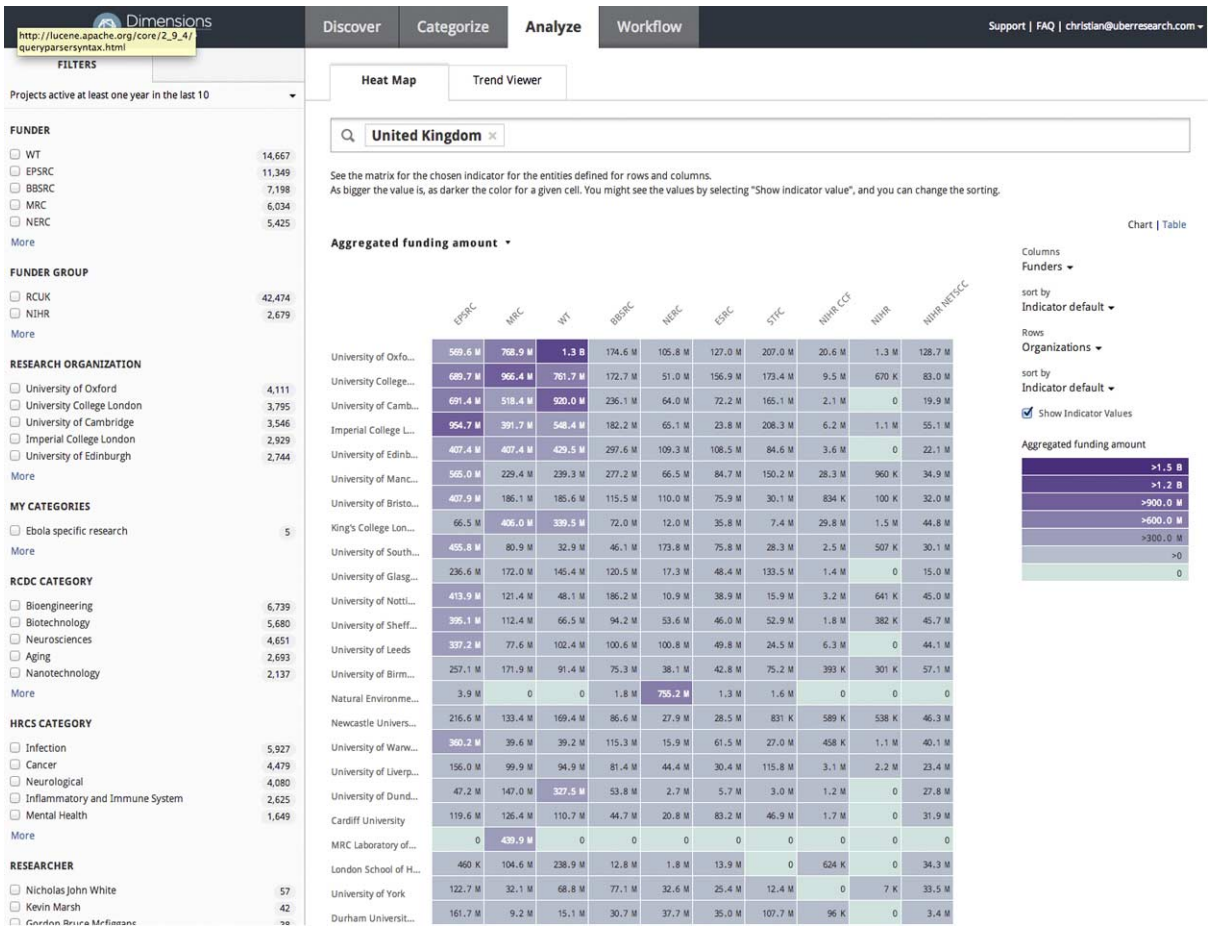

Fig. 4. Funders from the UK and funded research organizations. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140760.)

5. Summary and outlook

In sum, I believe with are at the beginning of a most interesting journey, in which research funding will be embedded in the Big Data ecosystem. To be sure, much of science already is. Yet, most research funding is public, and the public sector lags behind somewhat. The mission of ÜberResearch is to foster the adoption of big data as a support for expert decisions.

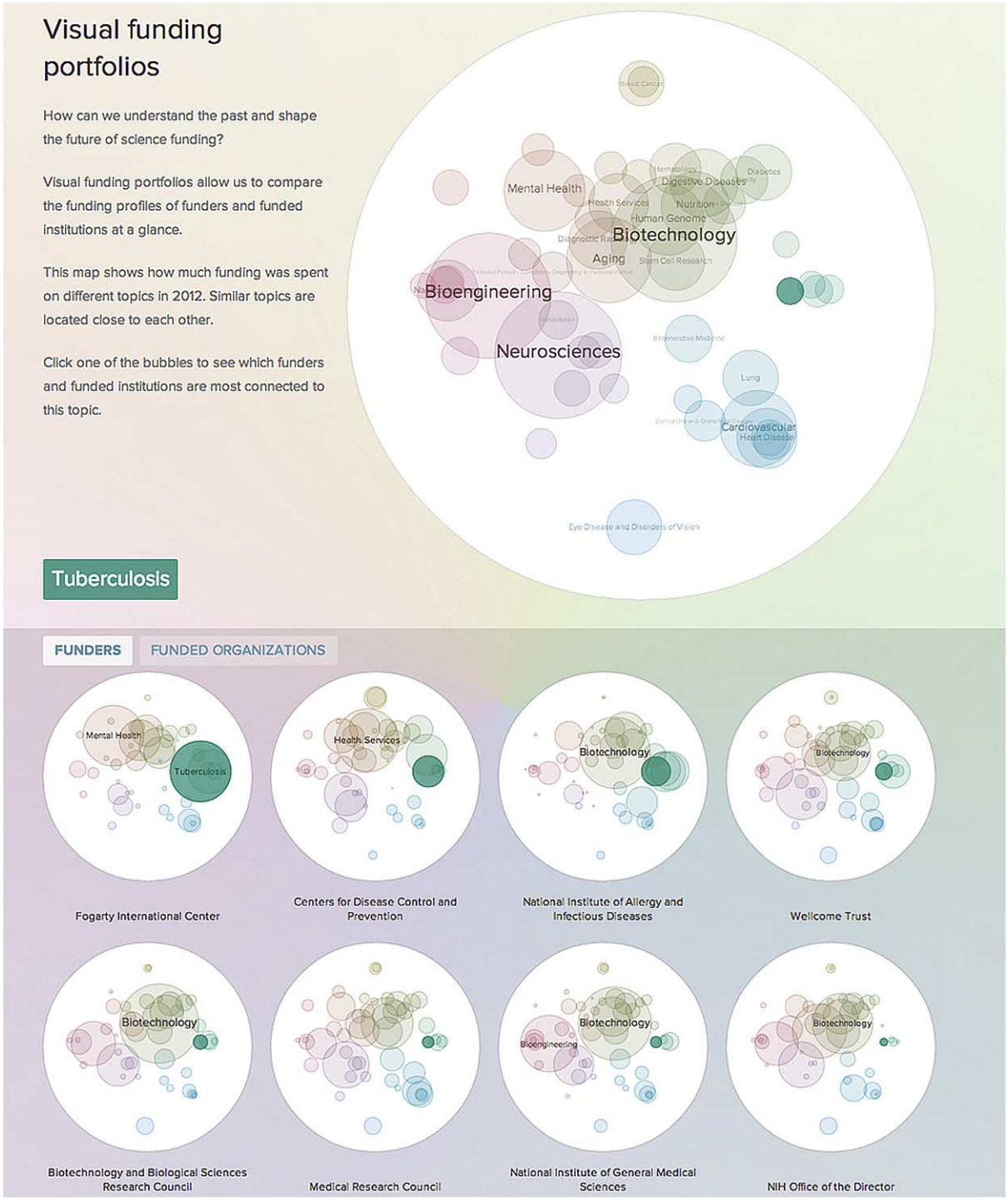

Fig. 5. Visual comparison of funder portfolios. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140760.)