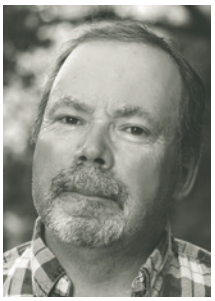
PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne**, **Jon R. Lorsch** *and* **Eric D. Green**.



Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries (including the United States), and data are being generated at unprecedented rates, so the research community must find more efficient models for storing, organizing and accessing biomedical data. Simply putting more and more money into the current systems is unlikely to work in the long term.

To better understand this situation, we are examining the current and projected costs of managing biomedical data at the US National Institutes of Health (NIH). Our initial analyses indicate that even if we leave out the National Center for Biotechnology Information, which is a special case, the 50 largest NIH-funded data resources have a collective annual budget of US$110 million. And this figure represents just the tip of the iceberg for future needs.

UNDERSTANDING USAGE

Today's biomedical data resources typically treat all items in their collections equally. This does not always make sense, given that the usage patterns of the data vary. But how do we decide which data get more attention? As larger and larger data sets are generated more easily, and the cost of maintaining and annotating these data continues to rise, this question is becoming increasingly important.

Answering it requires a better understanding of how research data are used. This has rarely been thoroughly explored. Historically, funders have been interested primarily in knowing how the data resources that they support are used and by whom. They tended not to look closely at the details of how and why individual items and types of data within a collection are used.

Analyses of these details can be revealing. Preliminary studies suggest that typically a small subset of the data is used frequently, whereas most of the data are rarely accessed. However, the exact subset of data that is used heavily may change over time, and most of the data access may be performed after the data are downloaded, so this is not recorded. All of this means that absolute numbers are hard to interpret.

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

When we have a better understanding of data usage, we can develop business models that consider supply and demand, and develop sustainable practices. In addition, finding economies of scale and harnessing market forces will be essential.

For a typical biomedical data resource, the cost of simply keeping the data is only a small fraction of the total cost of data management. The remainder is largely the cost needed to support the finding, accessing, interoperating and reusing (the FAIR principles; see go.nature.com/axkjiv) of the data — a cost that is widely underappreciated.

Is the FAIR fraction of the cost justified? Are services from different data resources redundant? Are resources subject to 'feature creep' — the addition of costly 'bells and whistles' that are of limited value? Do our funding mechanisms contribute to these problems? And most importantly, is the way we currently maintain biomedical data optimal for the science that needs to be done both today and in the future?

Current practices typically use many disparate sources of data to conduct a study. These data are located in a variety of repositories, often with different modes of access. This lack of centralization and commonality may hinder their ease of use and reduce productivity. We need a better understanding of usage patterns across multiple data resources to use as a basis for redesigning such resources to preserve valuable expertise and curation, and for improving how the data are found, accessed, integrated and reused.

The nature of curation and the quality assurance for biomedical data must also change. Complete and accurate automated or semi-automated extraction of literature is needed to provide metadata and annotation. We should consider crowdsourcing curation, with appropriate validation and incentives. Additionally, the role of professional curators must be better appreciated by data users, by the institutions where the curators work, and by the funders.

THE RESEARCH COMMUNITY MUST FIND MORE **EFFICIENT** MODELS FOR **STORING, ORGANIZING AND ACCESSING** BIOMEDICAL DATA.

In the longer term, we need models that are better aligned with the research life cycle. There is an unnecessary cost in a researcher interpreting data and putting that interpretation into a research paper, only to have a biocurator extract that information from the paper and associate it back with the data. We need tools and rewards that incentivize researchers to submit their data to data resources in ways that maximize both quality and ease of access.

BUSINESS MODELS

One business model worth exploring is the 'freemium model'. Here, the primary data are available free of charge, but services that add value to these data have an associated charge that generates funds that are used to maintain the primary data. This approach is used in other disciplines, notably chemistry. But there are two knotty questions. Should for-profit institutions be charged the same as non-profits? And who should own the intellectual property associated with value-added content?

Another potential business model is the 'subscription model', which is used to access the genetic and molecular databases that are provided by The Arabidopsis Information Resource (TAIR), for example. This option delivers support for a data resource from its active users, but it restricts access, which may be problematic for public-access data policies.

Taking the business-model idea further, what happens if data resources are merged, acquired or go out of business? Would existing resources be more useful and cost-effective if they were merged in some way? Should some services be dropped owing to lack of demand to make way for new services? Would reducing funding for particular data resources over time promote increased efficiency? To answer such questions, we would benefit from advice and help from the private sector and from other scientific communities.

COMMON GROUND

Cloud computing creates an element of data virtualization, takes computing to the data, and may help to solve some of the problems facing biomedical big data. At the NIH, we propose to exploit these opportunities by creating a 'commons' as one possible sustainable model.

Physically speaking, the commons will be collections of public and private resources (including cloud resources) for storing data and computing with those data. To be commons-compliant, such resources must abide by two simple rules. First, each research object in the commons — for example, data, software, narratives or papers — must be uniquely identified, sharable (taking into account privacy issues), and resolvable to its source by using a common identifier. Second, each research object must be defined by a minimal amount of metadata, as defined by the community.

The NIH Big Data to Knowledge (BD2K) programme (bd2k.nih.gov) aims to bring about the creation of the commons. The 12 new BD2K centres are encouraged to share research objects within the commons, and a BD2K consortium is prototyping an index that makes it easy to find commons content.

We also are studying the notion of computing credits, in which a grant recipient is given credits instead of funding to pay for computational time. A principal investigator would be able to spend those credits at any commons-compliant resource. Researchers whose work involves extensive computation on small amounts of data may spend their credits at a different commons-compliant resource to investigators who do minimal computing on large amounts of data.

This model is very different from the situation today. It shifts the initial burden of hardware, data and software maintenance from awardees and their institutions to third parties, notably cloud service providers. The funding model also has the effect of paying only for services used, and aims to create competition in the marketplace, so this approach could result in more data science per dollar.

If the pilot studies at the NIH are successful, it will be important to consider the longer-term implications of a commons model. One outcome is that data and software usage will be tracked both during an award period and after it has expired. Such tracking will yield important usage statistics that can inform future funding decisions.

LEN RUBENSTEIN/THE BROAD INSTITUTE

Research organizations such as the Broad Institute are rapidly evolving their practices for storing and accessing biomedical big data.

UNITING FUNDERS

The medical research community has too little money to start new data resources or to support the growth of more mature databases and services. Moreover, current funding schemes do little to foster the development of best practices; for example, each data resource is usually reviewed in isolation.

Changes to funding practices need to extend across both agency and international borders. Data generation and maintenance are typically funded nationally, but the data are used internationally. As a result, we need to develop more equitable funding models. The first step is for funding agencies to communicate more effectively about data science problems and to seek collaborative solutions. Working from the bottom up, scientists have been doing this for a long time.

Sustaining the biomedical big-data ecosystem is the responsibility of all stakeholders, and will require coordinated efforts among data generators, data maintainers, data users, funders, publishers and others in the private sector. The NIH BD2K programme, in collaboration with many stakeholders, is beginning to address these issues. ■

Philip E. Bourne *is associate director for data science at the US National Institutes of Health. He was previously associate vice-chancellor for innovation and industry alliances at the Office of Research Affairs at the University of California, San Diego.* **Jon R. Lorsch** *is director of the National Institute of General Medical Sciences. He was previously professor of biophysics and biophysical chemistry at Johns Hopkins University in Baltimore, Maryland.* **Eric D. Green** *is director of the National Human Genome Research Institute. He was previously its scientific director, chief of its genome technology branch and director of the NIH Intramural Sequencing Center. e-mail: philip.bourne@nih.gov.*