The NIH Big Data to Knowledge (BD2K) initiative

Philip E Bourne1,*, Vivien Bonazzi1, Michelle Dunn1, Eric D Green2, Mark Guyer1, George Komatsoulis3, Jennie Larkin1, and Beth Russell1

Understanding the human condition is a Big Data problem. This statement is nicely illustrated by the articles that follow from seven of the Centers for Data Excellence that have been funded by the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) initiative. BD2K is a trans-NIH program, funded by all Institutes and Centers at NIH as well as the NIH Common Fund; it is overseen by the NIH Office of Data Science within the NIH Office of the Director.

The beginnings of BD2K have been described previously,1 and the purpose here is to provide an overall context, from the perspective of NIH, for the emerging program, as exemplified by the work of the Data Centers of Excellence (7 of 12 described here), the Data Discovery Index Coordinating Consortium, the various training awards, and the various individual investigator awards that have been made.

What is aptly described by the seven Center articles is the research being performed across a rich array of data types and emergent infrastructure—metadata, analysis tools, frameworks, web resources, and more—that focus on problems inherent in extracting knowledge from large amounts of data with varying degrees of structure. Also detailed are plans to train researchers to make the most of the new opportunities presented by these data science advances.

Recognizing that what motivates researchers is the desire to understand the human condition, the BD2K initiative was designed so that driving biological problems are central to the effort, but with the solutions to those problems being deliverables in the form of new methods, tools, software, and training. The deliverables that emerge from the Centers should be FAIR2—that is, contributing to the ability to Find, Access, Interoperate, and Reuse the products of this research.

BD2K aims to have these digital objects exist, not in isolation, but rather as part of an emergent ecosystem that is shared with the biomedical research community at large. To this end, we have introduced the notion of the Commons, a shared virtual space that conforms to the FAIR principles. The Commons allows digital objects to be stored and computed upon by a broad community once found by the emergent data discovery index being developed by the BioCADDIE group.3 The Commons pilots that are under way are based on public cloud resources, but other compute and storage resources (such as high-performance computing facilities and institutional facilities) are expected to join the Commons as it develops. This assumes that the initial Commons pilots provide a cost-effective, sustainable, and usable environment. The primary aim of pilots now under way is to test this notion.

If the pilots are successful, the BD2K Centers described herein and individual investigator grants will seed the Commons with data and software, whereupon we can monitor usage, which is an important for determining the value of this research output.

We are at the beginnings of an exciting initiative, and the plans of the BD2K Centers are an excellent first step in advancing biomedicine in the era of Big Data.

REFERENCES

AUTHOR AFFILIATIONS
1 Office of the Director
2 National Human Genome Research Institute
3 National Center for Biotechnology Information, National Institutes of Health

*To whom correspondence should be addressed. E-mail: philip.bourne@nih.gov

© The Author 2015. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved. For Permissions, please email: journals.permissions@oup.com For numbered affiliations see end of article.