Open Data Science Symposium 2016
How Open Data and Open Science Are Transforming Biomedical Research

Office of the Associate Director for Data Science
National Institutes of Health (NIH)

Bethesda North Marriott Hotel & Conference Center
North Bethesda, MD

December 1, 2016

The Open Data Science Symposium, the first public meeting of its kind, was both a celebration of progress and a public forum. Open data can be defined as accessibility to content that is part of the scientific process; open science extends that notion to all aspects of the research lifecycle. Currently, progress is at a balance point—science has come a long way but has much further to go. Democratization of the scientific process has begun, allowing those who traditionally may not have been able to contribute to do so, but the slowness of the publication process still delays life-saving treatments. Opening the traditional pathways through which data become available for use can remove some of those barriers.

Keynote: Vannevar Bush in the 21st Century

John Wilbanks, Chief Commons Officer, Sage Bionetworks

Dr. Vannevar Bush's *Science, The Endless Frontier,* written after World War II, proposed that government should foster the opening of new frontiers, an idea that harmonizes with recent open access efforts by the NIH. Dr. Bush asserted that science is both a frontier itself and the proper concern of the government; it should be made available to the American people to meet the particular challenges of a time period. However, Dr. Bush commented in 1945 that standard methods of communicating information were inadequate for the state of science at that time, and science has reached such a point again—PDFs and publications remain the standard for the dissemination of new science, but their capabilities no longer are sufficient to communicate the breadth and complexity of the field. The open access movement is a rejection of both the initial hub-and-spoke architecture of the internet and the recent control of data stacks by large organizations, moving the scientific data network toward a distributed model that favors access without negotiation or favor.

As the culture shifts toward openness, team science organically grows to support an open community. Not all science needs to be completely open at all times; networks that are closed in the early stages of development increase the diversity of opinions and the accuracy of the consensus model by providing more options. A larger volume of participant-centered data can illuminate subtler trends and provide a multidimensional picture of each individual participant, and stakeholders in centralized hubs need to design data collection processes that remove themselves from control to prevent restrictive data stacks. Critical to current progress is the idea that creating open networks and posting data are not enough—data must be consulted to be of use. Each of the Open Science Prize finalists has made it easier to consult data and ensure that the scientific frontier is available to everyone.

Open Science Prize Demonstration of Results

Robert Kiley, Head of Digital Services, the Wellcome Library, the Wellcome Trust
Philip Bourne, Associate Director for Data Science, NIH

The Open Science Prize is an innovative effort showing how funding agencies can collaborate internationally. Despite the global aspect of data and science, funding typically is tied to the country of origin, but open science projects tend to provide significant value for a limited financial investment. The Prize was designed to challenge innovators, demonstrate exemplars, and encourage international collaboration. Submissions were received, all of exceptionally high standards, from 45 countries on six continents. The six Phase I winners of the Prize provided demonstrations of their platforms.

OpenTrialsFDA, fda.opentrials.net

The OpenTrialsFDA app makes clinical trials data from the U.S. Food and Drug Administration (FDA) accessible and searchable. Reporting biases can dramatically skew the risk/benefit ratios critical to evidence-based medicine. OpenTrialsFDA allows users to see the raw results of a study, such as unpublished data or data that seem more significant than they really are, in a way that is much more user-friendly and easier to navigate than the current database.

Real-Time Evolutionary Tracking for Pathogen Surveillance and Epidemiological Investigation, nextstrain.org

Nextstrain is an app for tracking pathogen evolution in real time, critical in this era of high mobility. Contact tracing is the main way to fight a virus without a vaccine; sequencing the genomes of viruses such as Ebola can determine the shared mutations and phylogeny of each strain, allowing field epidemiologists a more nuanced way to trace contact. To facilitate treatment of active outbreaks of pathogens such as Zika, Nextstrain is able to show molecular epidemiology within days. It also is intended to be scalable and easy to interpret for teams on the ground. The Zika data collected reveal the sources of the U.S. isolates, but these paths only become visible when data from multiple sources are combined, providing an incentive for scientists to share their data.

Fruit Fly Brain Observatory, fruitflybrain.org

The Fruit Fly Brain Observatory allows data from fruit fly brain scans to be used as models for investigating human neurological and psychological disorders, and this open source platform will help accelerate model development. Features include a natural language portal so that researchers can query the database more easily and a graphic functional explorer to translate experimental data into code and visualization. These attributes allow biologists to pursue their own intuition when exploring neurological questions. The Fruit Fly Brain Observatory also has integrated healthy and diseased models of the human brain for study. The platform is modular, so it will be extendable to mice, zebrafish, and other experimental animals.

Open Neuroimaging Laboratory, openneu.ro/start

Users of the Open Neuroimaging Laboratory can search for neuroimaging files of interest and open them using the BrainBox system, which has a similar functionality to Google Docs. Collaborators can send information, make comments, and highlight particular locations on the images, and access can be restricted to allow collaborators to view the images without modifying them. The team demonstrated searching for images of nonhuman brains in the species catalog. Users also can search for images of human brains with specific neurological conditions.

MyGene2: Accelerating Gene Discovery with Radically Open Data Sharing, [*mygene2.org*](mygene2.org)
MyGene2 is a family-facing site for collecting and sharing data on the 350 million rare diseases known worldwide. Many of these are monogenetic, but the causal gene is known for only about 50 percent of known conditions. MyGene2 is in active use right now; one-half of the information on the site was submitted by clinicians and one-half by families. Families can write their own health story, and key health terms can be extracted from this and checked for overlap with other families' stories to identify possible matches. If multiple families enter the same gene, it becomes a candidate for a match. Even conditions that have not been published or named are available over the internet on MyGene2.

OpenAQ: A Global Community Building the First Open, Real-Time Air Quality Data Hub for the World, [*openaq.org*](openaq.org)
The OpenAQ team defined "air inequality" as unequal access to clean air around the world. Poor air quality causes 5 to 7 million deaths a year, but the most polluted places in the world are not well-researched, hindering scientific progress. The OpenAQ platform collects data every 10 minutes and allows users to view stored data and compare locations. Many organizations already have begun to use the OpenAQ platform. Open data and science are powerful, but they have no force without a community built around them.

Voting for the Phase II winner will be open until January 6, 2017, at www.opensciencepize.org. The audience, both in the meeting and watching online, was urged to vote and encourage friends and family to participate. The finalists demonstrate the tip of the iceberg of what can be done with open science, and further progress requires that the community remain engaged and committed.

Open Science: An NIH Perspective
Francis Collins, Director, NIH
Harold Varmus, Lewis Thomas University Professor, Weill Cornell Medicine; former Director, NIH; former Director, National Cancer Institute (NCI)
Moderated by Chris Wiggins, Chief Data Scientist, New York Times, and Associate Professor of Applied Mathematics, Columbia University

Although the concept of open data is one aspect of open science, the broader aspects encompass all the components that must work together to succeed: people, processes, and technology. Communities have the power to shape science—scientists are people, after all. The task before the open science movement is to determine how to engage the broader community and still maintain the quality of the science. Citizens should have access to public goods, including data that have been provided through public funds. The current promotion criteria for scientists are not consistent with providing public goods, so the scientific community must work to improve the ways its members are evaluated. Evaluations should not be impersonal checklists of metrics such as publication counts but should be tools to enhance the community as faculty learn more about each other's capabilities. The gatekeeper role of publication is changing, and pre-print servers have become much more common, but there is more work to be done. Institutions must figure out how to reward openness, such as considering it part of the citizenship evaluation criteria.

At the first meeting of the Human Genome Project, the collaborators decided to place all data into a public database. This was a radical idea at the time, but it has become a signal moment for open science. It is a moral position to make data public rather than patented. Open science provides the maximum benefit to the maximum number of people, whereas patents should be reserved for discoveries that could benefit the public most as specific products.

The remaining challenges to open science should not be underestimated. Additional investment in infrastructure is required to maintain accessibility of the data, and one problem not yet addressed is the standardization of formats such as clinical nomenclature and criteria for disease stages. It is critical to remember that solutions to data problems are not just technical but involve language, politics, and getting people to agree. Such problems must be resolved at the human level.

The most important action that individuals can take to advance the progress of open science is to act as ambassadors, engaging colleagues and building relationships. If scientists espouse the principles of openness, they have a responsibility to make it a practical reality.

New Models for Open Science Emerging Around the Globe

Niklas Blomberg, Founding Director, Elixir
Peter Goodhand, Executive Director, Global Alliance for Genomics and Health
Robert Kiley, the Wellcome Trust
Tanja Davidsen, Project Manager, NCI
Moderated by Philip Bourne, NIH

National funding and differing research interests among members of international data infrastructures have been challenging for many of the organizations represented at the Open Science Data Symposium. Long-term commitments can help build relationships between disparate jurisdictions, and advance planning for systems expected to be shared can help negotiate gaps between countries or between disciplines. Open data in life sciences is extensively reused, and such transnational efforts can show stakeholders how open data feeds into the big picture of progress. Knowledge exchanges are dynamic and can be maintained only with active use.

Relationships among partners, funders, and the community, as well as support from all parties, are critical to furthering innovation and increasing the openness of research. The panelists shared challenges they have encountered, such as finding mutual recognition schemes, ways to reward researchers for sharing data, and approaches to recognizing the vital role of data curators. It also was noted that it is much easier to create alliances with English-speaking countries. For countries with higher barriers to alliance, the onus is on the community to reach out, listen to their needs, and make the cause of open science relevant to them.

Viewpoints on Open Science and Open Data in Biomedical Research

Heather Joseph, Executive Director, SPARC
Mike Huerta, Associate Director, National Library of Medicine, NIH
Jim Anderson, Director, Division of Program Coordination, Planning, and Strategic Initiatives, NIH
Open Science Prize Advisors:
Tim Clark, Harvard Medical School and Massachusetts General Hospital
Mark Hahnel, figshare
Ida Sim, University of California, San Francisco
Kaitlin Thaney, Mozilla Science Lab
Moderated by Jerry Sheehan, Assistant Director for Scientific Data and Information, White House Office of Science and Technology Policy

Specific goals are critical to furthering the cause of open science because they show what a culture values and can demonstrate relevant, relatable impacts that make open science desirable. Prizes such as the Open Science Prize have certain advantages over other methods of funding. For example, submissions must be inexpensive because competitors must be willing to lose the investment, and competitions can encourage submissions from nontraditional quarters and help fill unmet needs.

Panelists discussed whether current training is adequate, agreeing that scientists still are being trained to live in the last century. The systems and technology available to science have advanced, but educational practices have not kept up. One method to combat this is to reach potential scientists earlier in the educational pipeline and encourage young researchers such as the Phase I winners. The open access community does not yet understand much about what drives researchers and prize contestants. Many current participants are part of the early adopter community, but determination of incentives is critical to moving open science into the mainstream. On the clinical side, progress has been uneven, and there are not enough trained people and not enough funds to curate the data properly.

Open data and open science are powerful, but they are nothing without the power of community and without active consultation and use. Although hurdles and threats remain, openness is a gateway to additional progress, and advancing this cause provides real opportunities to accelerate the public health agenda worldwide.