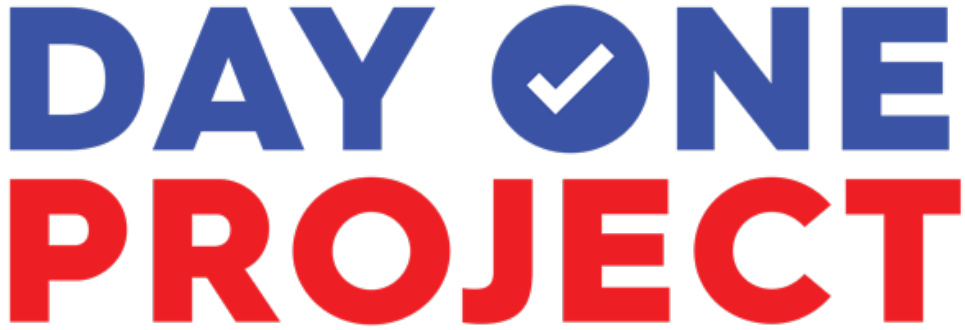

Open Access to Federally Funded Research Data

David Crotty
Ida Sim
Michael Stebbins

September 2020

Summary

The majority of scientific research data in the United States is not shared, meaning that our nation has vast untapped potential to fuel scientific advances. The next administration can dramatically accelerate scientific progress by (i) requiring scientists who receive federal funding to share their research data and (ii) directing federal research agencies to coordinate to build an International Research Data Commons that allows research data to be easily discovered and shared.

Challenge and Opportunity

Maximizing the value of taxpayer-funded research means making the results of—and the underlying data from—that research openly available, discoverable, and usable. Public access to scientific data and results fuels innovation and creates jobs.

The benefits of openly sharing research data have been proven repeatedly. For example, open data from the Human Genome Project catalyzed extraordinary breakthroughs in our understanding of biology while simultaneously creating new businesses, jobs, and economic growth.

Despite these benefits, publicly sharing research data remains an uncommon practice that is limited to a small number of fields and data types. Too often, the data produced in the course of scientific research supported by public dollars remain inaccessible to the public. Worse, when data are made available, they are often insufficiently documented and in forms that limit their discoverability, reuse, and repurposing, adding additional waste into the system. The lack of clear rules and norms for data stewardship in the sciences are preventing the full realization of public investment in science. Beyond the lost opportunities to drive innovation and spur economic growth, a lack of access to scientific data has exacerbated concerns over the reproducibility of scientific research.

Current research practices are largely haphazard, with research data ending up in various places ranging from commercial services to university servers to dusty hard drives stored in a laboratory's file cabinets. Calls for improved access to research results are intensifying from patient advocates, citizen and academic scientists, publishers, research sponsors, and politicians. Making the research data underlying publications available is not only essential for issues of scientific integrity and public trust, but for maximizing the returns on investment of federal tax dollars and maximizing the utility of research results from publicly funded science.

The first steps towards a federal policy requiring sharing of scientific data were taken in February 2013, when the White House Office of Science and Technology Policy (OSTP) announced the Obama Administration's policy on increasing access to the results of federally funded scientific

research.[1] In a memorandum to the heads of federal departments and agencies, OSTP Director John Holdren stated unequivocally that the Administration has a "commitment to increase access to federally funded published research and digital scientific data." The memorandum directed relevant departments and agencies to develop plans for increasing access to scientific publications and for improving management of and access to scientific data.

Although that memorandum expressed a policy position that "digitally formatted scientific data resulting from unclassified research supported wholly or in part by federal funding should be stored and publicly accessible to search, retrieve, and analyze," it did not direct agencies to take explicit action to require scientific data to be made available to the public.

Establishing a requirement for making the data underlying the conclusions of all federally funded research to be made freely available at the time of their publication would transform scientific research across the world to towards a more open enterprise. The next administration should also establish an International Research Data Commons to link databases and provide additional data-storage solutions for scientists. These complementary steps will facilitate easy storage, access, reuse, and repurposing of research data, strengthening science by improving reproducibility and accelerating scientific progress by increasing access to information.

Plan of Action

Open Science Data Policies**:** The most practical approach to addressing the challenges and opportunities laid out above is for the next President to issue an Executive Order (EO) requiring each federal agency that directly supports scientific research and development to develop and implement an Open Science Data Policy. These policies should collectively ensure the free, public availability of digitally formatted scientific data generated with public funds. Because science progresses primarily through the communication of results via peer-reviewed publications, sufficient data underlying published conclusions must be made available to enable expert verification or replication of those conclusions. Data should be made publicly available at or before the time of initial publication of the research results.

Specifically, each agency's Open Science Data Policy should:

1) Require covered data to be made freely available to the public on the original date of publication of any research articles.
2) Require that data be made available in machine-readable formats, unencumbered by restrictions that would impede use or reuse by the American public.

[1] Holdren, John P. (2013). Expanding public access to the results of federally funded research. White House Blog. Available at https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.

3) Require agencies to develop data-management plans that explicitly include provisions to ensure that data are shared for all research-grant and contract applications that are expected to result in peer-reviewed publications.

4) Require data-management plans to be scored for merit, consistent with OSTP's February 2013 memorandum.

5) Establish mechanisms and requirements enabling easy search, retrieval, and reuse of data.

6) Establish enforcement mechanisms, including agency authority to audit for compliance, adjudicate for noncompliance, recover funds in the event of noncompliance, and disqualify non-complying entities from receiving future funding until all data-access requirements are met from previous awards.

7) Establish formatting standards for long-term storage of data in digital, agency-maintained repositories or in any approved repository consistent with agency policy.

8) Ensure that data are stored in archives and repositories that:

    a) Allows long-term data preservation and access without charge.

    b) Uses widely available data standards and, to the extent practicable, nonproprietary archival formats.

    c) Provides access for persons with disabilities consistent with Section 508 of the Rehabilitation Act of 1973.

    d) Enables, as appropriate, integration and interoperability with other data-storage solutions and archives.

    e) Links publications stemming from federally funded research to corresponding datasets.

Open Science Data Policies would only apply to unclassified research supported wholly or in part by federal funding, including research conducted in-house by applicable federal agencies as well as research conducted by scientists receiving funding from federal agencies. Policy exemptions should be available for certain types of data, but exemptions should be limited and consistent across all federal agencies. For example, Open Science Data Policies should not apply to research reports presented at professional meetings, laboratory notes, author notes, preliminary data analyses, phone logs, or other information that is used to produce final manuscripts but is not needed to verify or replicate scientific conclusions. Exemptions should also be made for patentable discoveries to the extent necessary to protect a copyright or patent. Agencies will likely identify other cases in which exemptions are necessary to protect confidentiality, personal privacy, and certain business interests.

The EO should also direct agencies to work through the National Science and Technology Council to coordinate development and implementation of their Open Science Data Policies. Previous directives requiring coordination around open science and data have not been

effectively coordinated across federal agencies. Therefore, it will be paramount for the White House Office of Science and Technology Policy to require such coordination before agency plans are approved.

International Research Data Commons: Making large amounts of data open, machine readable, and accessible will require more commercial and public platforms for storing, sharing, finding, and using data. Excellent models of publicly available databases for the storage of highly structured, monotypic scientific data already exist (e.g., GenBank and Protein Data Bank for the biological sciences). The next administration can draw from these examples to build an International Research Data Commons (IRDC): a federated system linking databases together through common application programming interfaces (APIs). All databases included in the IRDC should comply with Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles. This requirement will create new norms in scientific data publishing and raise the floor on the overall quality of open research data.[2] In concert with launching the IRDC, the next administration should explore policy mechanisms for managing data access, requests, and credits. Different mechanisms will be needed for generalist repositories (those that accept data regardless of disciplinary focus) or specialist repositories (those that are disciplinary-specific).

Conclusion

Within the first 100 days, of the next administration, the President could dramatically accelerate scientific research progress by setting the U.S. (and therefore, the world) on a path to ensuring that the data produced in the course of publicly funded research is freely available for reuse, repurposing, and confirmatory analysis. Opening up scientific data is the most important step that can be taken to accelerate and strengthen the U.S. scientific enterprise without dramatically increasing the budgets of science-funding agencies. Taking this step would also be a significant milestone for addressing challenges in scientific reproducibility and replicability. Requiring that all federally supported research data be publicly accessible would unequivocally illustrate the next administration's commitment to spurring national innovation and would pave the way for unleashing a tsunami of creative and innovative impactful data-science insights from researchers around the world.

[2] Wilkinson, M.D.; et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3.

Frequently Asked Questions

Which data should be shared?

One of the most common arguments against research data sharing is that there will be little value in sharing all data created in the course of a research project. Reasonable people can disagree on this point, but it is hard to make a strong case that data directly supporting conclusions in research publications should not be shared. The research publication is the currency of the realm for most academic and governmental research endeavors. The research publication also gives scientists control over which results they report to the public and how they communicate their story. Publication authors decide which results are strong enough to support significant conclusions worth communicating to the larger scientific community. Making research data available enables external readers to explore for themselves whether they agree with the authors' conclusions.

How much data is shared already?

Many individual research journals already require that the data that contributes to the results described in a given paper be made openly available. For a handful of fields, sharing data is not just a norm, but also a requirement for publication. The federal government can help data sharing become a routine practice regardless of what field a researcher works in or what journal they target for publication. The federal government can also support development of standards for which data must be shared and in what formats. Because different scientific disciplines have different approaches to research and generate different types of data, it will be important for the federal government to engage the research community—potentially through scientific societies—in establishing and expanding discipline-specific standards for common data types.

Where should data be stored?

Infrastructure necessary for storing and providing access to research data is sorely lacking. Current research practices are largely haphazard, with research data ending up in various places ranging from commercial services to university servers to dusty hard drives stored in a laboratory's file cabinets. While repositories exist for some data types, many research data are forgotten and lost over time. Large-scale archival infrastructure is needed to fulfill the requirements of an open research-data enterprise and maximize returns on taxpayer investments in research. There is no single ideal solution for data storage and access. Rather, leveraging a variety of distributed but federated resources (commercial, non-profit, and governmental) that meet a common set of required standards will be the best and most feasible way to support open data practices.

How can federal agencies ensure compliance with a new open data policy?

Researchers are already over-burdened with reporting requirements. There is a real need to dramatically reduce compliance workloads, which take time and effort away from research itself. Data sharing, though, is so fundamental to reproducibility and replicability of experimental research—and to maximizing taxpayer investment in research—that it should be prioritized above most other reporting requirements. That said, it is incumbent upon federal agencies to institute data-sharing requirements in a manner that minimizes burden. Keys to success will be clear data sharing, storage, and archival standards, as well as coordination and cooperation between funding agencies and publishers around assigning metadata and persistent identifiers that enable easy data discovery and reuse. Attention and funding must also be dedicated to workforce needs with respect to data management and stewardship. Government can partner with academia to train, reward, and promote careers in data stewardship.

Requiring persistent identifiers for datasets associated with publications is an important way to drive compliance with data-sharing requirements. In other words, to receive credit for publications for the purpose of applying for or renewing federal funding, data associated with any publications must be readily available and easily findable in approved archival solutions. Failure to make data publicly available should limit a researcher's ability to apply for further funding.

Incentives for following best practices should also be offered. Datasets should be recognized by federal agencies as valid research outputs, and creation of data that are reused and that drive subsequent research should be recognized in grant applications. Similar incentives should be used for promotion and tenure decisions at universities. OSTP should therefore establish relationships with organizations representing universities and explore creating a consortium that will recognize open data, data reuse, and impact from open data in tenure and promotion decisions.

How long should agencies be given to implement recommended open-data policies?

A minority of scientists believe that infrastructure and standards for sharing data must be built out before any open-data directive can be issued. Scientists in this minority believe that open-data efforts should be driven by the scientific community and that policy should follow the actions of the scientific community, not the other way around. We, the authors of this memo, contend that this does not reflect the reality of how change happens in a large and diverse enterprise like science. At the same time, we recognize that a huge shift in scientific culture and practice cannot happen overnight. We believe that phasing in recommended open-data policies over a period of 3–5 years would give sufficient time to develop necessary archival solutions, common practices, and standards. The phase-in process should begin with common and important data types in specific fields. Requiring public availability of these data as an initial step will familiarize researchers with open-data practices, paving the way for expansion of such practices to other

data types and scientific disciplines without causing undue disruption to ongoing research projects.

What standards should be used for determining which data should be stored and in what format?

Data sharing is of little direct utility if the data does not comport with Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles. Making data findable is particularly difficult given the global spread of distributed data repositories representing an ever-expanding number of research subject areas. Although data would be more findable if housed in a centrally managed repository system, the continuous development of a wide variety of approaches to storing research data is likely to be more robust, adaptable, and realistic. That said, a common set of requirements and practices for data discoverability and interoperability are essential for open-data success.

Publicly available databases for the storage of highly structured, monotypic scientific data already exist (e.g., GenBank and the Protein Data Bank in the biological sciences). These databases are of enormous value to the scientific community. Establishing similar databases where appropriate for important common data types would do much to accelerate scientific progress. Because scientific research is so varied, and because new data types are continuously generated, it is unrealistic to think that monotypic databases will be sufficient. However, there are important lessons to be learned from these databases that can be applied more generally.

How would an International Research Data Commons be managed?

There are a number of ways that a research data commons can be established and managed. We favor establishing a nonprofit organization supported by public funds to certify that databases are complying with common data-storage and access standards, and to coordinate with publishers to ensure easy integration of data sharing into publisher workflows. These steps would establish a new paradigm for treating research data as an open asset. We expect that a U.S.-managed International Research Data Commons would act as a coordinating entity among nations, linking and establishing standards for common APIs that can be used by research databases worldwide. It is critical that the research commons be international in reach because so much of science today involves collaboration among labs across the globe, and because the scientific community is not structured by nations. Therefore, an inclusive approach reflects the reality of how science is conducted today.

About the Authors

David Crotty, Ph.D. is the Editorial Director of Journals Policy for Oxford University Press. He oversees journal policy and contributes to strategy across OUP's journals program, drives technological innovation, and serves as an information officer. Dr. Crotty previously managed a suite of research society-owned journals with OUP, and before that was the Executive Editor for Cold Spring Harbor Laboratory Press, creating and editing new science books and journals, along with serving as a journal Editor in Chief. Dr. Crotty received his Ph.D. in Genetics from Columbia University and did developmental neuroscience research at Caltech before moving from the bench to publishing. Dr. Crotty has been elected to the Boards of the Society for Scholarly Publishing, the STM Association, and CHOR Inc. As the Executive Editor of *The Scholarly Kitchen* blog, Dr. Crotty regularly writes about current issues in publishing.

Ida Sim, MD, Ph.D. is a primary-care physician, informatics researcher, and entrepreneur. She is a Professor of Medicine at the University of California, San Francisco, where she co-directs Informatics and Research Innovation at UCSF's Clinical and Translational Sciences Institute and is Director of Digital Health for the Division of General Internal Medicine. Dr. Sim is a global leader in the technology and policy of large-scale health-data sharing. She is a co-founder of Open mHealth, a non-profit organization that is breaking down barriers to mobile health app and data integration through an open software architecture. She co-developed CommonHealth, an open-source software suite bringing to the Android ecosystem the equivalent of Apple Health's ability to access and share EHR data. Dr. Sim is also co-founder of Vivli, the world's largest data-sharing platform for participant-level clinical trial data. In 2005, she was the founding Project Coordinator of the World Health Organization's International Clinical Trials Registry Platform, where she led the establishment of the first global policy on clinical trial registration.

Michael Stebbins, Ph.D. is a geneticist and public-policy expert who served as the Assistant Director for Biotechnology in the Obama White House Office of Science and Technology Policy. He is currently the President of Science Advisors, a science and health consulting firm he founded in 2018 to provide science, technology, and public-policy guidance to private companies, philanthropies, and non-profit organizations. While at the White House, Dr. Stebbins' work led to large initiatives across the Federal government to address antibiotic resistance, protect pollinators, improve veterans' mental health, increase access to federally funded scientific research publications and data, promote the preferential purchasing of antibiotic-free meats, reform the regulatory system for biotechnology products, drive federal purchasing of bio-based products, and improve the management of scientific collections. Dr. Stebbins previously served as the Vice President of Science and Technology for the Laura and John Arnold Foundation, science advisor to the Obama Presidential Campaign, and on the Obama White House Transition Team. He is the former

director of biology policy for the Federation of American Scientists and worked for U.S. Senator Harry Reid and at the National Human Genome Research Institute. Before coming to Washington, he was a senior editor at Nature Genetics. Dr. Stebbins is on the Board of the Value in Cancer Care Consortium and chair of the Board for Vivli. He serves on the scientific advisory boards for The Agenda Period, Amida Technology Solutions, Datavant, and KOKOMI.

About the Day One Project

The Day One Project is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans, and readying them for Day One of a future presidential term. For more about the Day One Project, visit dayoneproject.org.