

Who Will Pay for Public Access to Research Data?

Francine Berman¹ and Vint Cerf²

On 22 February, the U.S. Office of Science and Technology Policy (OSTP) released a memo calling for public access for publications and data resulting from federally sponsored research grants (1). The memo directed federal agencies with more than \$100 million R&D expenditures to “develop a plan to support increased public access to the results of research funded by the Federal Government.” Perhaps even more succinctly, a subsequent *New York Times* opinion page sported the headline “We Paid for the Research, So Let’s See It” (2). So who pays for data infrastructure?

The OSTP memo requested agencies to provide plans by September 2013 that describe their strategies for providing public access to both research publications and research data. Plans are expected to be implemented using “resources within the existing agency budget,” i.e., no new money should be expected. Currently, federal R&D agencies are working hard to foster approaches to public access, to assess needs for supporting partnerships and enabling infrastructure, and to develop timetables and approaches for implementation. We focus here on the research data portion of the OSTP memo, rather than on publications.

Digital data are ephemeral, and access to data involves infrastructure and economic support. In order to support the downloading of data from federally funded chemistry experiments, astronomy sky surveys, social science studies, biomedical analyses, and other research efforts, the data may need to be collected, documented, organized in a database, curated, and/or made available by a computer that needs maintenance, power, and administrative resources. Access to data requires that the data be hosted somewhere and managed by someone. Technological and human infrastructure supporting data stewardship is a precondition to meaningful access and reuse, as “homeless” data quickly become no data at all.

¹Rensselaer Polytechnic Institute, Troy, NY 12180, USA.

²Google, Reston, VA 20190, USA.

*Corresponding author. E-mail: bermaf@rpi.edu



Research data of community value are supported today in a variety of ways. Some of them, like those in the Protein Data Bank (PDB) (3)—a database of protein structure information used heavily by the life sciences community—are supported by the public sector. (In particular, U.S. funding from the National Science Foundation (NSF), the National Institutes of Health (NIH), and the U.S. Department of Energy for the Research Collaboratory for Structural Bioinformatics (RCSB) PDB is \$6.3 million annually.) Other data, as from the Longitudinal Study of American Youth (LSAY) (4)—a longitudinal study of student attitudes about science and careers—are available through subscription from the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. (ICPSR membership ranges from \$15,750 for doctoral research—extensive academic institutions to \$1680 for community colleges and provides access to 7500 data collections.) Some data live on researchers’ hard drives, some are stored by the commercial sector, and some are hosted in academic libraries, private or public repositories, or archives. Much of our federally funded research data are “at risk,” with no long-term viable economic model in place to ensure continuing access and preservation for the community. An in-depth study of the economics of digital preservation (5, 6) explored the complex issues of supporting valued data for the public good, but ultimately there is no economic “magic bullet” that does not require someone, somewhere, to pay.

When economic models and infrastructure are not in place to ensure access and preservation, federally funded research data are “at risk.”

What happens to valuable data when project funding ends? Consider, for example, a 3-year research project in which valuable sensor data are collected from an environmentally sensitive area. Those data may be useful not just for the duration of the project but for the next decade or more to collaborators and a broader community of researchers. For the first 3 years, the costs of stewardship (including development of a database that supports analysis, access to the data for the community through

a portal, adequate storage and management of the data collection, and so on) may be paid for by the grant. But who pays for subsequent support? In such cases, research data may become more valuable just as the economics of stewardship become less viable.

Up to this point, no one sector has stepped up to take on the problem alone, and it is unrealistic to expect as much. In the public sector, federal R&D agencies are unlikely to allocate enough resources to support all federally funded research data. The costs of infrastructure would absorb too great a portion of a budget that must support both innovation and the infrastructure needed to drive innovation. The private sector, especially in information technology, has tremendous capacity and expertise to support the stewardship of public-access research data; however, there are few explicit incentives to take this on. In early 2008, Google announced that it would begin to support open-source scientific data sets. By the end of the year, the project was shut down for business reasons (7). Without explicit incentives and credits, it is challenging for companies to step forward and partner productively to support the common good. In the academic sector, university libraries are natural foci for the stewardship of digital research data. But they need financial support to evolve in this direction at a time when many budgets are being cut.

The key is not to look to a particular sector alone but to develop much stronger

partnerships among sectors. Such a division of labor can provide a framework of options that distribute the burdens and benefits of stewardship and economic support. This is a growing trend internationally, where initiatives such as the Vector Ecology and Control Network (VEcNet) (8) are combining private funding (The Gates Foundation and others), public funding (the Australian National Data Service and others), and academic sponsorship (Oxford, the University of Pittsburgh, James Cook University, and others) with the goal of building tools to analyze malaria transmission and reduce its spread by vector control interventions.

... research data may become more valuable just as the economics of stewardship become less viable.

The recent emergence of the community-driven Research Data Alliance (RDA) (9) is also capitalizing on cross-sector and transnational opportunities to build the social, organizational, and technical infrastructure needed to accelerate open-access research data sharing and exchange.

Within the United States, the public access for research data stewardship problem could begin to be addressed through four coordinated approaches:

Facilitate private-sector stewardship of public access to research data. With sufficient public incentives, including tax credits and other means, federal and state governments could make it attractive for the private sector to host, preserve, and serve up public-access research data. Support for the public good is not new to the private sector. Private companies frequently sponsor the arts and social causes, often leveraging advantageous economic models that promote support for the public good. Why not utilize the same approaches to provide stewardship for public-access research data, itself a quasi-public good? Note that it will be important for federal and state incentives to incorporate a commitment to smooth the transition for research data collections when companies choose to move their investments elsewhere. Moreover, adequate safeguards need to be in place to avoid private-sector control of access and use (10, 11).

Use public-sector investment to jump-start sustainable stewardship solutions in other sectors. The success of the Alzheimer's Disease Neuroimaging Initiative (ADNI) (12) demonstrates that public-private partnerships can accelerate research

discovery through data sharing and collaboration. Public partnership with the academic sector can help provide access and stewardship options as well. At present, many progressive university libraries, such as the Johns Hopkins Sheridan Libraries (13), are proactively seeking to address community needs for digital research data stewardship. With an initial public or private short-term "ramp-up" investment in library capacity and workforce (perhaps through something analogous to the federal government's Small Business Innovation Research and Small Business Technology Transfer programs), as well as with the expectation that univer-

sities will work with their libraries to create sustainable economic models when such initial funding has ended, libraries can begin to curate and provide access to some of the data that researchers are generating.

Create and clarify public-sector stewardship commitments for public access to research data. There is ample precedent for the government to support certain data collections of great community value, such as the PDB. However, with thousands of grants per year and data sets ranging from megabytes or less to terabytes or more, not all federally funded data can be realistically hosted within a public repository. In light of this, clarification about which collections would be supported by the public sector, for how long, and under what circumstances would be tremendously helpful to the research community. Knowing what will not be hosted within public repositories could help drive new stewardship efforts in other sectors as well.

Encourage research culture change to take advantage of what works in the private sector. Finally, researchers, like the general public, subscribe to digital versions of newspapers, donate to Wikipedia, pay for and download iTunes, buy data services online, and so forth. In other words, researchers pay for many kinds of digital data. Yet there is widespread expectation that access to research data should be supported by the government or academic institutions and be free to the research community. As community infrastructure becomes increasingly fundamental for data-driven research, the research community could begin to use economic models that are effective else-

where. Imagine supporting the National Virtual Observatory (astronomy data) from telescope advertisements or paying a small download fee for data from digital marine collections in the same way we download music from the Internet. Such economic models will not solve the whole problem but could help provide some infrastructure support necessary for access and preservation of research data. [The possibility of charging for the commercial use of databases as data tools under the take-and-pay rule has been suggested (10).]

Public access presupposes that the research data supported by public funding will be available when they are sought. Such availability is dependent on the existence of effective data infrastructure, i.e., to access data, they must be hosted somewhere, and someone must fund the human and technological infrastructure that hosts the data. Without viable economic models for this infrastructure, valuable research data may disappear, making it accessible to no one and deterring us from making the most of our research investments.

References and Notes

1. J. P. Holdren, Public access memorandum from the Office of Science and Technology Policy; www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
2. Editors, "We Paid for the Research, So Let's See It," *New York Times*, 25 February 2013, p. A24; www.nytimes.com/2013/02/26/opinion/we-paid-for-the-scientific-research-so-lets-see-it.html?_r=0.
3. Biological Macromolecular Resource, RCSB PDB. www.rcsb.org/pdb/home/home.do.
4. Longitudinal Study of American Youth, www.isr.umich.edu/cps/project_lsay.html.
5. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation: Interim Report* (BRTF-SDPA, La Jolla, CA, 2008); <http://brtf.sdsc.edu/publications.html>.
6. BRTF-SDPA, *Sustainable Economics for a Digital Planet: Final Report* (BRTF-SDPA, La Jolla, CA, 2010); <http://brtf.sdsc.edu/publications.html>.
7. P. Lenssen, "Google stops research datasets program," Google Blogoscooped; <http://blogoscooped.com/archive/2008-12-23-n33.html>.
8. The Vector Ecology and Control Network, www.vecnet.org/.
9. The Research Data Alliance, <http://rd-alliance.org/node>.
10. J. H. Reichman, R. L. Okediji, *U. Minn. Law Rev.* **96**, 1362 (2012).
11. J. H. Reichman, P. F. Uhlir, *Law Contemp. Probl.* **66**, 315 (2003).
12. Alzheimer's Disease Neuroimaging Initiative, ADNI Overview; <http://www.adni-info.org/Scientists/ADNIOverview.aspx>.
13. R. Monastersky, *Nature* **495**, 430–432 (2013).

Acknowledgments: We thank G. Alter, H. Berman, P. Bourne, B. Plale, S. Choudhury, J. Bicarregui, R. Wilkinson, and the reviewers for their help with specific information for, and comments on, this piece.

10.1126/science.1241625