THE CHRONICLE
of Higher Education

Commentary

Home Opinion & Ideas Commentary

August 3, 2010

Too Many Researchers Are Reluctant to Share Their Data

By Felicia LeClere

A new model of data sharing and openness is emerging in the scientific community that replaces traditional ways of thinking about research findings as the private property of the primary investigator. Large granting agencies, including the National Science Foundation and the National Institutes of Health, have embraced the new model of more-open access to research data. Later this year, the NSF will start requiring scientists seeking research grants to include a data-management plan in their applications, describing how and when their data will be shared.

The issue has also captured the attention of a U.S. House of Representatives subcommittee, which held hearings last week on an NIH data-sharing policy requiring that federally financed research data be freely available within 12 months of publication.

But the change has been slower to take hold among scientists themselves—resistance that is bogging down the pace of scientific progress. Policy changes come on the heels of a 2009 National Academies of Science report about data stewardship, which suggested that, with the help of technology, the scientific professions are moving, but only slowly and reluctantly, toward a paradigm shift.

It has become increasingly apparent that scientific data should be considered a product in much the same way journal articles or conference proceedings are, and they should therefore be shared as widely as articles and proceedings, while being credited to their producers. NIH embraced this perspective as early as 2003 by requiring data-sharing plans in certain types of grant applications. The momentum has clearly shifted toward more transparency, at least among those who finance science. But among those who do science, it remains much less clear how long the transition will take.

For the past five years, I have been on the front lines of this shift, and I have seen little consensus among scientists. As an archive director for the Inter-University Consortium for Social and Political Research at the University of Michigan, which is among the oldest and largest data archives in the world, it has been my job to persuade researchers financed by federal sources to share their research data with the broad academic community. I have attended countless scientific meetings, presented at workshops, shared coffee and meals, and been cornered in poster sessions by disgruntled scientists and worried graduate students. The topic is always access to secondary research data— the reuse of primary data for secondary analysis. Researchers' concerns include discomfort over possible misuse of their data and losing credit for their work.

Data sharing is a bit like going to the dentist. We can all agree that it is a good thing to do and intrinsic to good scientific practice. In reality, however, researchers tend to view data sharing with a mix of fear, contempt, and dread.

Of the reasons and excuses offered for not sharing research data, very few have substantial legitimacy from a scientific or even institutional perspective. Arguments related to protecting human subjects are valid. In the social and behavioral sciences, when researchers collect data, we promise to protect our subjects' identities. That is an important promise that, if broken, can substantially damage our research by making it more difficult to get cooperative subjects for future surveys, and by eroding our trustworthiness.

But the remainder of excuses for not sharing data are rooted in the nature of academic rewards, the financing of science, and misunderstandings about data. The misdirection occurs when we believe that we can protect our subjects by never sharing data with anyone beyond our work group. That would be true only if data still resided on pieces of paper.

The truth is, informal data sharing occurs every time a researcher puts a data file on a thumb drive and hands it to a graduate student, who then puts it on a shared network drive at a university, and then puts it on his or her laptop to take home and hook up to an unsecured wireless network. Formal data-sharing plans force us to think through data-protection and disclosure-control practices. Informal data sharing actually puts subjects at greater risk because we trust our colleagues, while never questioning the networks, computers, and places they use to store and analyze our data.

Moreover, many scientists extend the human-subjects argument from individuals to populations. They argue that when the data concern a vulnerable population, they could be misrepresented if widely released. The scientists are less concerned with protecting the identities of individuals than with controlling how the data are used to portray a particular population.

Again, that is a laudable goal, but misguided. Researchers often persuade a person to participate in a study to improve the condition of his or her community. Participants clearly believe in the value of the science brought to bear on their issues. The original research team, however, may not have all the answers. The value of science lies in the ability to exchange and test alternative solutions. By trying to protect a community from harm, the team may actually be hurting it by shutting out alternatives.

I've also heard and thought about many other arguments against data sharing, none of which ultimately hold water. For example:

- "I worked hard for this, and I want to exploit it as much as I can." It is true that academe is designed to reward publications and, thus, when we share data, we run the risk of being "scooped." That suggests, however, that the individuals who collected the data have no competitive advantage at all. Secondary-data analysis, by those who are not the primary researcher, is actually quite hard even when the data are well prepared and documented, because data collection has become so complicated it is difficult to navigate. Further, data producers gain the advantage of having completed their work first, forcing others to cite them in future publications.
- "People won't use the data properly." Can we dictate how other analysts use our written work? The scientific discourse is one of error and correction. The literature is filled with such exchanges. If we prevent people from entering the conversation because we are afraid they might say something stupid, we violate the basic principle of science that statements are considered valid when well supported by evidence or until proved wrong. Data are the raw materials of those conversations.
- "It's too expensive to clean it up." Collecting data is a bit like cooking a good meal. If you clean as you go, when you are full and sleepy you will have much less to do. Documenting and cleaning data are good scientific practice. It should be very little work to make data ready for someone else.
- "I won't share it because it's mine." That is the least credible and most objectionable reason for not sharing data. Call it the kindergarten gambit. In fact, data collection supported by the federal government belongs to the institution to which the grant was given. Contracts have different types of ownership embedded in their agreements. In the case of grants, a researcher's legal claims on the original data are minimal unless he or she has negotiated an alternative agreement with his or her institution. More important than the legal claim is the moral one. If we continue to ask American taxpayers to finance scientific research, we ought to be willing to share some of its products: data.

The most effective argument in favor of data sharing is simple: It is good science. The scientific community need only look to the field of astronomy for a well-documented example. As *The Chronicle* has reported, Alexander S. Szalay, a professor of physics and astronomy at the Johns Hopkins University, has helped change his field to one in which sharing is routine, building an archive that brings together millions of digital images of the universe. Genomics also serves as an example, with several large-scale projects engaged in broad collaboration on gene sequencing, like the Human Genome Project and the "genomewide association studies" that scan markers across many people to identify variations associated with particular diseases that it has made possible.

At our consortium for social and political research, which has nearly

700 members and is housed at the Institute for Social Research at the University of Michigan, data sharing has been our mission since our founding nearly 50 years ago. Demand for the data we disseminate is only growing. In 2010 we've seen record-high downloads from our Web sites. Our experience demonstrates that making data available for secondary analysis by the wider research community is an essential component of social-science inquiry.

Our hope is that the most recent effort by the NSF, along with practices already in place at NIH, will push more researchers to realize that being overly protective of one's data is counterproductive. When the shift finally occurs, perhaps I will spend less time on the road trying to persuade people of the value of data sharing, and more time facilitating its use.

Felicia LeClere is an associate research scientist and director of Data Sharing for Demographic Research and the National Addiction & HIV Data Archive Program at the Inter-University Consortium for Political and Social Research at the University of Michigan. She is also an associate research scientist at the university's Population Studies Center, Institute for Social Research and will be a principal scientist at the National Opinion Research Center at the University of Chicago in September.

The Chronicle of Higher Education 1255 Twenty-Third St, N.W. Washington, D.C. 20037