THE CHRONICLE
of Higher Education
The Digital Campus

think

April 29, 2012

Who's Paying the Data Bill?

By Francine Berman

Recent federal initiatives have highlighted the importance of managing and preserving digital data. They also raise key questions about the economics of digital preservation and, most specifically, about who will pay. Answering those questions is critical to current and future data-driven discovery and innovation.

In March 2012 the Obama administration announced a $200-million plan for "Big Data" research and development projects. The announcement follows a 2011 National Science Foundation requirement that grant applicants must submit a data-management plan with their grant proposals. The policy states: "Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled 'Data Management Plan.' This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results."

The Obama initiative and the NSF policy reflect a growing recognition of the role of research data in fueling innovation and discovery. What is not so clear is this: Who will pay the "price at the pump"? That is, who should pay for the human and cyberinfrastructure required to support stewardship of digital data? The question is critical because without reliable infrastructure and sustainable economic support, valuable research data may cease to exist. As research becomes increasingly data-driven, the economics of data access, stewardship, and preservation threaten to become our Achilles' heel.

The requirement for a data-management plan does not mean that all research data generated from NSF-sponsored research projects are valuable or should be retained. Investigators often discard data not considered useful to a broader community. At present, the

researchers themselves largely determine which data are valuable. This may change over time as feedback from proposal-review committees and requirements from agencies and publishers guide which data need to be retained, and under what circumstances.

Many data-management plans will be simple and inexpensive. For locally manageable research-data files that are not too big, are not subject to restrictive privacy policies (such as Hipaa, the Health Insurance Portability and Accountability Act), or do not require broad access, researchers may opt to retain data on a hard drive and provide access only to collaborators. Other types of data necessitate different approaches. For example, 200 terabytes of astrophysics data from supercomputer simulations (equal to 200 trillion bytes, or almost five times the amount of data produced in the first 20 years from the Hubble Space Telescope) cannot feasibly be kept on a hard drive. Longitudinal community-data collections such as the Panel Study of Income Dynamics, a 40-plus-year survey of families that tracks economic, social, and health factors over individual life spans and across generations, is now more valuable to social scientists than ever, and must be retained in an institutional-data repository with stable financial support.

Some data collections, such as the Protein Data Bank, a worldwide digital repository of information about the three-dimensional structures of large biological molecules, have become community resources and have garnered reasonably sustainable economic support. However, other community-data collections, such as the Arabidopsis Information Resource, a database of genetic and molecular-biology data for the model plant *Arabidopsis thaliana,* are struggling to identify viable sustainable economic models.

Moving the costs of research-data stewardship solely to the grantee, universities, or the government is not economically viable. Grantees can generally contribute only a small portion of their funds to data stewardship, and only during the project period. Retention of valued research collections past the project completion date poses a serious problem. Universities and their libraries need financial support for data stewardship and preservation. They cannot absorb the bulk of retention-worthy research data without additional money. The U.S. government is simply not in a position to retain all valuable research data for its grantees. With tight budgets for supporting research and innovation, the competing (and codependent) needs of financing for research—and financing for infrastructure that supports research—make it difficult for federal agencies to adequately deal

with both needs.

The good news is that we are not without economic models for data stewardship. Log in to Facebook (free for users), and the ads that appear help to pay the data bill for the stewardship of posts, pictures, and keystrokes of Facebook users. Log in to newspaper Web sites, and your digital subscription helps pay for online articles, video, and audio. Visit the Protein Data Bank, and a consortium of government agencies supports reliable access 24/7 to this invaluable community resource.

We can apply a broad set of economic models to academic research data as well, but this will require identification of appropriate repositories as well as additional stewardship options in multiple sectors. Here are a few such examples:

Universities, and in particular their libraries, can improve and expand their repository infrastructure for research data through institutional investment, user fees, federal investment, philanthropy, or some combination thereof.

Federal agencies can focus on investments in data stewardship by increasing the capacity of research-data repositories and investing in the development of data scientists and data-savvy professionals throughout the work force. Federal grant programs could seed additional repository infrastructure with the requirement that grantees incorporate a sustainable economic model that supports continuing data stewardship and infrastructure when federal financing ends.

The private sector can provide critical capacity and services to support research data. With "carrots" such as tax incentives for supporting research and other data in the public interest, and stringent requirements governing open access, transition plans, and other issues, government programs can incentivize private-sector data centers to host research data.

The best economic models may leverage the capabilities and capacity of multiple sectors and involve synergistic partnerships. Future research-data collections in the life sciences, for example, might come with a cost of 99 cents per access, advertisements from pharmaceutical companies, and partial institutional support from a host repository. Future curated astronomy data collections could require a subscription fee. Adoption of such models will require a cultural shift—most researchers are not used to paying to access or fully supporting research data—but these models are on the horizon.

Viable approaches to paying the data bill are needed to ensure the success of data-driven research. As we respond to Big Data opportunities and develop data-management plans in response to agency requests, we must also develop the requisite economic models and infrastructure required for appropriate stewardship of digital data. If we do not address all parts of the data-driven-research challenge, we are in danger of losing data on which future innovation and competitiveness relies.

Francine Berman is vice president for research at Rensselaer Polytechnic Institute and co-chair of the National Academies Board on Research Data and Information.