Preserving research data

James A. Jacobs, University of California, San Diego
Charles Humphrey, University of Alberta in Edmonton.

Communications of the ACM. Volume 47, Number 9 (2004), Pages 27-29.

Granting ownership rights to data, as if it were private property, only limits data access without ensuring the benefits of researcher precedence or the rewards for good data collection.

Forces opposing free and open access to scientific data are stronger than ever. As a result, researchers, as well as archivists, must deal with reduced access and increased risk of permanent loss stemming from claims of commercial ownership of scientific data [6], the failure of research projects to apply sound data documentation and management standards, and the dearth of institutions with mandates and funding to preserve research data.

As practicing data archivists, we routinely encounter these obstacles while striving to preserve and support access. The profession of data archivist, while small in number, is international in scope [7] and important to all sciences, though the social sciences have led its development [1]. This column reflects our more than two decades of managing data archives and assisting scientific data users.

Two ideas dominate the discussion among scientists about how to improve access to research data. One is tied to notions of data publishing, treating data as a publishable product the same way research findings are treated as publishable products; the other is tied to the emergence of digital repositories, emphasizing the deposit and storage of data. Both ideas stem from a desire to establish researchers' intellectual control over the data they collect or produce, reward good data collection, and improve the software tools needed to find, acquire, and cite data. Unfortunately, these ideas overlook established methods for preserving research data for long-term equitable access and use.

We see core weaknesses in the ideas of data publishing and digital repositories. First, treating data dissemination as a publishing activity automatically erects barriers to free and open access. And by emphasizing the deposit and storage of files rather than the preservation of data, repositories are an inadequate solution.

Data publishing conflates publication (peer review and distribution) with issues of intellectual property (controlling and limiting access) and therefore encourages the commodification of scientific data. Commodification has seriously negative consequences for access to data. In the context of today's data economy, it determines who gets what data, when, and how.

We've identified [5] three prevailing data markets. First is the exchange of research data based on the values of open and free access; we call this the commonwealth market, where data is shared openly as a resource vital to knowledge discovery and replication. Second is the barter economy operating through an underground market where the data is traded via an informal network of scientists seeking favors in exchange for data or selectively deciding who gets their data. Finally, data is sold or leased

in the commercial economy, depending on private or intellectual property claims. Science is hampered by the barter and commercial data economies. How one provides open access to research data in them is thus an important issue for research.

Granting ownership rights to data and encouraging scientists to claim copyright to it allows producers of data to control access. Treating data as property enables its "owners" to impose contractual restrictions on its use. This approach greatly limits equal access to the data and is not necessary for ensuring the benefits of researcher precedence or rewards for good data collection.

We support the rights of scientists to protect their intellectual contributions from plagiarism. We also support the rights of scientists to protect their research findings through the peer review and publication process. Scientists are entitled to protect their research results based on their own original interpretation and creativity, ensuring no one else claims authorship of their work. However, we see no need to treat research data as property in order to ensure these rights.

Research data should be part of the scientific commonwealth, in the same way ideas are shared. For the sake of open scientific discourse and scholarship, data should be treated like fresh air freely available to all. Treating research data as property will ultimately restrict the interpretation and development of ideas. Science that relies on the ownership of facts and data seeks only to generate commercial value -- and is a very narrow view of science.

The development of digital repositories does not go far enough to ensure data preservation or long-term access. Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a project and incorporate a schedule for depositing products over the course of a project's life cycle and the creation and preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method.

Archives generally exist to select, preserve, manage, and provide access to collections of materials deemed valuable in perpetuity. Data archives provide these services for research data, managing digital content by creating persistent identifiers and by providing accurate, verifiable, sharable descriptions of content through human-created, authority-controlled catalogues. A management concern in data archives is the survivability and usability of data across changes in computing technology. NASA's loss of data from its early Earth orbiters is an example of data being stored rather than being preserved [11].

Simply storing data does not constitute preservation. Simply storing minimal, descriptive metadata does not ensure usability. True data archives are equipped for storing, preserving, and providing access to the various kinds of metadata that accompany research data. Metadata includes data documentation, descriptions of methodology and instrumentation, formulas used for recoding and analyzing the data, and field notes, as well as administrative and preservation records that ensure the long-term access and usability of data files. An example of a design for complex metadata is the Data Documentation Initiative and schema [4]. Examples of preservation metadata include the Open Archival Information System [2], Cedars Metadata for Digital Preservation [3], and the Networked European Deposit Library's metadata for long-term preservation [9].

In May 2002, a Working Group affiliated with the Organisation for Economic Cooperation and

Development Committee for Scientific and Technological Policy reported: "In everyday research life, researchers may tend to act as owners of the data used in their projects. The responsibility for sustainable archiving of research data is not always assigned to the relevant parties. Lack of regulation on these aspects may hamper access to and sharing of research data" [10]. Repositories that rely on individual scientists to take on the additional, specialized role of curator of their data and that rely on voluntary data storage facilities with no long-term commitment or mission to preserve data and no funding for preservation invite disastrous loss of data. The solution must involve data archives created, mandated, funded, and staffed for data preservation. Some data archives exist today, but additional archives are needed, as advances in technology have vastly expanded the number and scope of research projects.

Technical and Human Infrastructure

We'd like to offer a few suggestions about how to create an effective technical and human infrastructure for ensuring scientific data preservation. First, trained professionals are critical. Data archiving cannot be achieved through automation alone. Professionals are needed to select and acquire data and documentation in multiple formats and create and preserve new metadata essential for long-term access to the data. Data archivists provide this kind of management for both the metadata and the scientific research data to ensure not just its availability but its usability as well. A key part of the overall scientific community's responsibility for ensuring the preservation of research data is the training of more data archivists to build capacity in the profession.

Second, the scientific community, including the U.S. National Science Foundation, needs to invest in the creation of data archives of national prominence (such as the the U.S.-based Inter-University Consortium of Political and Social Research) [8]. Whether they should be general or topical should be discussed by the academy and elsewhere. Moreover, the long-term funding and staffing commitment must extend way beyond the good intentions of a few revolving-door custodians. This investment will, however, improve the return on the research budget by enabling better and more permanent access to data.

Third, building new data archives is an opportunity to create partnerships among institutions to provide a safety net for capturing and retaining of the products of scientific research. Meanwhile, research libraries and their relationships with scientific publishing must continue. Data archives can work with these institutions to coordinate the capture and preservation of research data.

Though differences among the sciences tend to deter open discussion, the importance of research data to all sciences and their applications should be viewed as common motivation for preserving and making available data to today's, as well as tomorrow's, scientific community.

References

1. Clubb, J., Austin, E., and Geda, C. "Sharing research data in the social sciences." In *Sharing Research Data*, S. Fienberg, M. Martin, and M. Straf, Eds. National Academy Press, Washington, D.C., 1985, 39-88.

2. Consultative Committee for Space Data Systems. *Reference Model for an Open Archival*

Information System, Blue Book Issue 1. Washington, D.C., Jan. 2002; ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/Ccsds-650.0-B-1.pdf.

3. Curl Exemplars in Digital Archives (Cedars). *Metadata for Digital Preservation: The Cedars Project Outline Specification* (draft for public consultation). The Cedars Project Team and Ukoln, Leeds, England, Mar. 2000; www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html.

4. Data Documentation Initiative Alliance. Inter-university Consortium for Political and Social Research, Ann Arbor, MI; www.icpsr.umich.edu/DDI/.

5. Geraci, D., Humphrey, C., and Jacobs, J. *Data Basics*. Canadian Library Association, Ottawa, ON. [forthcoming]

6. Hotz, R. "Scientists sharing fewer discoveries." *Los Angeles Times* (Feb. 11, 2002), A12.

7. International Association for Social Science Information Service and Technology; iassistdata.org.

8. Inter-University Consortium for Political and Social Research. Ann Arbor, MI; www.icpsr.umich.edu/.

9. Lupovici, C. and Masanes, J. *Metadata for Long-term Preservation*, Tech. Rep. Networked European Deposit Library, The Hague, The Netherlands, July 2000; www.kb.nl/coop/nedlib/results/preservationmetadata.pdf.

10. Organisation for Economic Co-operation and Development Follow-up Group on Issues of Access to Publicly Funded Research Data. *The Public Domain of Digital Research Data, Tech. Rep.*; dataaccess.sdsc.edu/PlanningDoc.html.

11. Sniffen, M. "Some of America's history is lost in the computer revolution." Associated Press (Jan. 2, 1991).

Authors

James A. Jacobs (jajacobs@ucsd.edu) is the data services librarian at the University of California, San Diego.

Charles Humphrey (humphrey@datalib.library.ualberta.ca) is head of the Data Library and academic director of the Research Data Centre at the University of Alberta in Edmonton.
