Neuroinformatics
Copyright ©Humana Press Inc.
All rights of any nature whatsoever are reserved.
ISSN 1539-2791/03/289–296/$25.00

Public Policy Forum

Towards Effective and Rewarding Data Sharing

Daniel Gardner,[*,1] *Arthur W. Toga, Giorgio A. Ascoli, Jackson T. Beatty, James F. Brinkley, Anders M. Dale, Peter T. Fox, Esther P. Gardner, John S. George, Nigel Goddard, Kristen M. Harris, Edward H. Herskovits, Michael L. Hines, Gwen A. Jacobs, Russell E. Jacobs, Edward G. Jones, David N. Kennedy, Daniel Y. Kimberg, John C. Mazziotta, Perry L. Miller, Susumu Mori, David C. Mountain, Allan L. Reiss, Glenn D. Rosen, David A. Rottenberg, Gordon M. Shepherd, Neil R. Smalheiser, Kenneth P. Smith, Tom Strachan, David C. Van Essen, Robert W. Williams, and Stephen T. C. Wong*[2]

[1]Laboratory of Neuroinformatics, Department of Physiology & Biophysics, Weill Medical College of Cornell University, New York, NY. [2]For the institutional affiliations of the other coauthors, please *see* Appendix.

Recently issued NIH policy statement and implementation guidelines (National Institutes of Health, 2003) promote the sharing of research data. While urging that "all data should be considered for data sharing" and "data should be made as widely and freely available as possible" the current policy requires only high-direct-cost (> US $500,000/yr) grantees to share research data, starting 1 October 2003. Data sharing is central to science, and we agree that data should be made available. As investigators funded by the NIH's Human Brain Project, we have promoted data sharing and thus applaud the initiation of a meaningful data-sharing policy. We have also explored related technical and sociological benefits and barriers, and our support is coupled to proposals for improvement and extension of the policy and guidelines. This perspective is based on our experience advancing the field of neuroinformatics and thus it is proper that we use the pages of *Neuroinformatics* to advance it. We offer this perspective as a private effort on our part, not an NIH-sponsored or initiated activity. Our goal is to ensure that data sharing is, and is recognized to be, effective and rewarding.

We encourage sharing both to enhance the utility of data and to promote competition in the marketplace of scientific ideas. Data shar-

*Address to which all correspondence and reprint requests should be sent. E-mail: dan@aplysia.med.cornell.edu ing permits reanalyses and meta-analyses beyond the expertise or time constraints of the original data collectors. Informed by shared data, new hypotheses can be advanced and current hypotheses can be retested on new data. Archived data can be used as well to develop or validate new analytic methods or technology. However, we take exception to the characterization of only some classes of data as "unique." All data are unique, and their uniqueness derives from the focus, techniques, protocols, selection, and expertise inherent in each investigation.

Data sharing is a complex issue with multiple technical, social, financial, and legal facets (Marshall, 2002a,b). The benefits, pitfalls, and techniques of sharing data depend upon both the type of data and the field within biomedical science; an example of issues related to neuroimaging databases was recently presented in *Science* (Governing Council of the Organization for Human Brain Mapping [OHBM], 2001). Policy development and implementation should reflect such complexities.

The NIH policy recognizes, yet incompletely addresses, the fundamental problems presented by the wide diversity and enormous scale of contemporary biomedical data. Data vary in type, size, storage requirements, and significance. Without standards for data formats, descriptive labels, and units of measurements, data may be "available" but not usable for sharing. Standards ease sharing of conforming data, but standards often require a huge amount of effort to establish.

Differing models for data sharing (such as peer-to-peer exchange and central database resources) present separate technical challenges. In addition to differences in scale, different modes of data sharing raise issues of privacy, technology, and standards, as well as responsibility of development and maintenance for each of these. At its simplest, data can be shared peer-to-peer; only two parties need negotiate constraints such as format, privacy requirements, and the meaning of data labels, and data security is easily maintained. Public sharing lies at the other extreme, in which data are placed on servers visible to a large community. This wider visibility dramatically increases the potential impact, but can require community-wide acceptance and raise data privacy issues.

Peer-to-peer data sharing asks individuals to establish and maintain data archives, but it can take considerable work to convert local research data into a form that can be distributed and shared. The volume of data produced by some techniques can be immense, and large-scale data storage imposes requirements for cataloging or indexing as well. Methods are needed to let potential users know that data are available for sharing, what the data represent, and how they may be selected, obtained, and used. Without standards for distributing data, the effort to develop peer-to-peer solutions can potentially require an ad-hoc solution for each pair of investigators.

Centralized data archives require standards as well. These standards must serve multiple users, including investigators recording or generating the data and investigators accessing the data, and must guide developers and maintainers of the databases. Such larger archives multiply data volume requirements by the number of submitters. However, their development effort is more efficient than for peer-to-peer models, as one resource serves many users. Developing and adopting standards is desirable as well for *interoperability*: coordinating disparate data resources. Here, widespread adoption of standards can avoid the need for individual database-to-database negotiation to link types of data and descriptors.

The NIH policy recognizes, but again minimizes the barriers—both technical and human—to analyses of data by those unaffiliated with the original investigators. In the absence of safeguards, data sharing potentially invites misappropriation, misuse, and misinterpretation.

Sharing should not imply relinquishing. Proper assessment and assignment of credit for data, recognition of the relative value of data acquisition versus data processing, and awareness of the potential for commercial exploitation of freely released data should inform any policy for data sharing. Biological data often require extended development work. Faint signals may require exceptionally difficult development and monitoring of methods for acquisition, filtering, transformation, or reconstruction. A single structural biology dataset may be the culmination of years of exploration of one macromolecule. Complex, massively parallel, high-throughput procedures may generate enormous extended data volumes requiring sophisticated search strategies. Studies in some human subjects require painstaking searching and selection to acquire a specific subject population, followed by extended data collection. Areas such as functional imaging may combine several of these aspects.

Particularly disturbing to many of us is the relative ease of reuse of data whose acquisition may represent extensive and as-yet-unrewarded effort, especially where performed by new or junior investigators who have not yet established a secure position or reputation. The problem of being scooped with one's own data may be particularly serious for data sets that are planned to yield multiple reports over time, or for studies where the design itself completely encapsulates a scientific insight. Since meaningful credit for research is largely tied to publication, sharing of experimental design, motivation, or data via extra-publication routes risks inappropriate allocation of scientific credit.

There are technical barriers to reanalysis as well, because datasets alone are rarely sufficient to extract and interpret the information provided by the experiment that generated them. Detailed *metadata*—descriptions of data including protocols and analytic specifications—are required to understand what the primary data meant in its original context. In the absence of such metadata, analyses of data by an outside investigator are open to misinterpretation. Such misreading could lead to the publication of unwarranted results that might improperly cast doubt upon the conclusions of the original work, or impugn unfairly the competence or scientific integrity of the original investigators.

The implementation guidelines address concerns of investigators requested to release their own data. We propose easing barriers presented by the reluctance of investigators to use others' data, arising from technical factors such as format differences as well as more fundamental questions including uncertainty about metadata, internal quality control, or clear traces of transformations or processing. An evolving data sharing policy should address these issues, to forestall collecting and archiving massive data sets that others are reluctant to use.

We therefore propose a series of emendations that we believe would strengthen the policy and promote its acceptability and success. Although the goals of advancing biomedical science through data sharing may be broadly accepted, the scope of sharable data may legitimately vary depending upon the standards and practices of different fields or techniques, and may thus include or exclude any or all of "raw," partially processed, processed, or selected datasets. Ideally, sharable data should be defined as the combined experimental data and descriptive metadata needed to evaluate and/or extend the results of a study. Policies should recognize that small amounts of adequately characterized, focused data are preferable to large amounts of inadequately defined and controlled data stored in a random repository. Further, data sharing will benefit from recognized, usable technological and descriptive standards for data and metadata. It is in part this diversity that leads us to recommend that a variety of models for data sharing should be acknowledged and even encouraged by the NIH's Institutes and Centers. Applicants should be guided to specify data-sharing plans by scope of data, type and format of data to be shared, metadata to be included, credit sought, and model to be used (such as peer-to-peer or database).

Active collaborations should be supported as well as mechanisms for passive reuse of data. Making data public through databases or other open resources should promote, not preclude, collaborations. For many types of data, and many designs of studies, the absence of universal or fixed standards means that viable data pooling requires explicit coordination between producers and users of data. We note that ongoing communication with collaborators aids mutual understanding of data and hypotheses, and avoids many potential pitfalls of analysis.

To promote data sharing, a citation and credit paradigm must be encouraged. A data sharing policy should include safeguards against reuse of data without recognition of the original investigator; such use is equivalent to appropriation and should not be tolerated. Where shared data are used, acknowledgment of the sources and collectors should therefore be mandatory, but mere acknowledgement may well not be adequate credit for some types of data. Safeguards should require that reanalysis of data be limited to that which can be meaningfully derived, given restrictions, parameters, or boundaries inherent to the original hypotheses, protocols, and techniques for acquisition and processing.

Publication provides an example of a familiar, open, near-universal methodology for sharing data as well as methods, concepts, conclusions, news, and reviews. It depends upon an established yet evolving infrastructure; there exist methods and recognized standards for manuscript content and preparation as well as publication of journals and books. Can a publication model serve for data sharing, and methods be established for archiving and retrieval of data comparable to those encompassed by the familiar terms manuscript, reviewer, editor, journal, subscription, library, reprint, photocopy, or PDF? Such a model might inform the scope of data sharing; papers present focused, relevant data rather than extended lab notebooks.

Publications are offered with the hope that they will be read and cited extensively. Just as information, once published, is open to any reader, so data once posted should be available to any viewer. Fear about rapid or preemptive reuse or post-hoc analysis of data might be lessened if data were equivalent to publication. Papers will be read, and as a consequence, hypotheses will be tested or advanced, and new suggestions, critiques, or analyses based on published data or ideas will arise. If a similar culture for data existed, including safeguards and a reward system, reluctance to make data available might diminish.

Finally, we urge Congress, the NIH, and other concerned Federal agencies to increase programs and funding for the development of informatics methods enabling investigators to share data with accuracy, accountability, responsibility, and recognition. Existing programs such as BISTI, the Human Brain Project, and others targeting informatics needs of specific communities or techniques should be expanded, and efforts at each of several levels instituted towards interoperability among current and future projects. In addition to standardized databases of selected domains that are sharable by particular research communities, these should include methods and pilot projects for technology development and application, standards for data description and exchange, and scalability to cover the large and expanding universe of biomedical data.

References

Governing Council of the Organization for Human Brain Mapping (OHBM). (2001) Science 292, 1673.

Marshall E. (2002a) Data sharing. DNA sequencer protests being scooped with his own data. Science 295, 1206.

Marshall E. (2002b) Clear-cut publication rules prove elusive. Science 295, 1625.

National Institutes of Health. (2003) Final NIH statement on sharing research data, available at http://grants.nih.gov/grants/policy/data_sharing/index.htm

Appendix: Coauthors' Affiliations:

Daniel Gardner
Laboratory of Neuroinformatics
Weill Medical College of Cornell University
New York, NY 10021
E-mail: dan@aplysia.med.cornell.edu

Arthur W. Toga
Laboratory of Neuro Imaging
UCLA School of Medicine
Los Angeles, CA 90025-1769
E-mail: toga@loni.ucla.edu

Giorgio A. Ascoli
Head, Computational Neuroanatomy Group
Krasnow Institute for Advanced Study
George Mason University
Fairfax, VA 22030-4444
E-mail: ascoli@gmu.edu

Jackson T. Beatty
Department of Psychology
University of California Los Angeles
Los Angeles, CA 90095
E-mail: beatty@psych.ucla.edu

James F. Brinkley
Department of Biological Structure
University of Washington
Seattle, WA 98195-7420
E-mail: brinkley@u.washington.edu

Anders M. Dale
Massachusetts General Hospital NMR Center
Charlestown, MA 02129
E-mail: dale@nmr.mgh.harvard.edu

Peter T. Fox
Research Imaging Center
UTHSCSA
San Antonio, TX 78229-3900
E-mail: fox@uthscsa.edu

Esther P. Gardner
Department of Physiology & Neuroscience
NYU School of Medicine
New York, NY 10016
E-mail: gardne01@endeavor.med.nyu.edu

John S. George
Biophysics Group
Los Alamos National Laboratory
Los Alamos, NM 87545
E-mail: jsg@lanl.gov

Nigel Goddard
Division of Informatics
University of Edinburgh
Edinburgh EH1 2QL UK
E-mail: Nigel.Goddard@ed.ac.uk

Kristen M. Harris
Institute of Molecular Medicine and Genetics
Medical College of Georgia
Augusta, GA 30912-2630
E-mail: kharris@mail.mcg.edu

Edward H. Herskovits
Johns Hopkins University
Baltimore, MD 21287
E-mail: ehh@braid.rad.jhu.edu

Michael L. Hines
Section of Neurobiology
Yale University School of Medicine
New Haven, CT 06520-8285
E-mail: michael.hines@yale.edu

Gwen A. Jacobs
Department of Cell Biology and Neuroscience
Montana State University
Bozeman, MT 59717
E-mail: gwen@cns.montana.edu

Russell E. Jacobs
California Institute of Technology
Pasadena, CA 91125-7400
E-mail: rjacobs@caltech.edu

Edward G. Jones
Center for Neuroscience
University of California Davis
Davis, CA 95616
E-mail: ejones@ucdavis.edu

David N. Kennedy
Department of Neurology
Massachusetts General Hospital
Charlestown, MA 02129
E-mail: dave@nmr.mgh.harvard.edu

Daniel Y. Kimberg
Neurology Department
University of Pennsylvania
Philadelphia, PA 19104
E-mail: kimberg@mail.med.upenn.edu

John C. Mazziotta
Director, Brain Mapping Center
UCLA School of Medicine
Los Angeles, CA 90024
E-mail: mazz@loni.ucla.edu

Perry L. Miller
Director, Center for Medical Informatics
Yale University School of Medicine
New Haven, CT 06520-8009
E-mail: perry.miller@yale.edu

Susumu Mori
Department of Radiology
Johns Hopkins University School of Medicine
Baltimore, MD 21287
E-mail: susumu@mri.jhu.edu

David C. Mountain
Department of Biomedical Engineering
Boston University
Boston, MA 02215
Email: dcm@bu.edu

Allan L. Reiss
Department of Child & Adolescent Psychiatry
Stanford University
Stanford, CA 94305-5719
E-mail: reiss@stanford.edu

Glenn D. Rosen
Department of Neurology
Beth Israel Deaconess Medical Center
Boston, MA 02215
E-mail: grosen@caregroup.harvard.edu

David A. Rottenberg
Department of Neurology and Radiology, University of Minnesota
VA Medical Center (127)
Minneapolis, MN 55417
E-mail: dar@neurovia.umn.edu

Gordon M. Shepherd
Department of Neurobiology
Yale University School of Medicine
New Haven, CT 06520-8009
E-mail gordon.shepherd@yale.edu

Neil R. Smalheiser
Psychiatric Institute
University of Illinois
Chicago, IL 60612
E-mail: smalheiser@psych.uic.edu

Kenneth P. Smith
The MITRE Corporation
McLean, VA 22102-7108
Email: kps@mitre.org

Tom Strachan
Institute of Human Genetics
International Centre for Life
Newcastle upon Tyne NE1 3BZ UK
E-mail: Tom.Strachan@newcastle.ac.uk

David C. Van Essen
Department of Anatomy & Neurobiology
Washington University School of Medicine
St. Louis, MO 63110
E-mail: vanessen@v1.wustl.edu

Robert W. Williams
Center of Genomics and Bioinformatics
Department of Anatomy and Neurobiology
University of Tennessee Health Science Center
Memphis, TN 38163
E-mail: rwilliam@nb.utmem.edu

Stephen T. C. Wong
Department of Radiology
University of California San Francisco
San Francisco, CA 94131
E-mail: swong@radiology.ucsf.edu