Data Sharing I: Principles of Data Sharing

The Human Brain Project 2002 PI Group: D. Gardner, A.W. Toga, G.A. Ascoli, J. Beatty, J.F. Brinkley, A.M. Dale, P.T. Fox, E.P. Gardner, J.S. George, N. Goddard, K.M. Harris, E.H. Herskovits, M. Hines, G.A. Jacobs, R.E. Jacobs, E.G. Jones, D.N. Kennedy, D.Y. Kimberg, J.C. Mazziotta, P. Miller, S. Mori, D.C. Mountain, A.L. Reiss, G.D. Rosen, D.A. Rottenberg, G.M. Shepherd, N.R. Smalheiser, K.P. Smith, T. Strachan, D.C. Van Essen, R.W. Williams, and S.T.C. Wong

Data Sharing: A Neuroinformatics Perspective

The NIH has recently issued a policy statement and implementation guidelines to 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 (> $ 500,000/yr) grantees to share research data, starting 1 October 2003.

The policy and other NIH data sharing documents and resources may be accessed at:
`http://grants.nih.gov/grants/policy/data_sharing/`

As investigators funded by the 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 we offer it as a joint effort on our part, not as an NIH sponsored or initiated activity. Our goal is to ensure that data sharing is, and is recognized to be, effective and rewarding.

A version of this perspective is scheduled to appear as a commentary in the journal *Neuroinformatics*.

Data Sharing: The Goal

Data sharing is central to science, and we agree that data should be made available.

We encourage sharing both to enhance the utility of data and to promote competition in the marketplace of scientific ideas. Data sharing 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; current hypotheses can be re-tested on new data. Archived data can be used as well to develop or validate new analytic methods or technology.

The Breadth of Data Sharing

Data sharing is a complex issue with multiple technical, social, financial, and legal facets. 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 presented in *Science* in 2001 by the Governing Council of the Organization for Human Brain Mapping. 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.

In this poster, we offer comments, followed by guidelines to promote effective and rewarding policies and methods and to reduce technological and sociological barriers to data sharing. Although this perspective is based on our experience advancing neuroinformatics, these proposals are designed to serve a broader spectrum of biomedical investigations.

Reanalysis of Shared Data

The NIH policy recognizes, but perhaps 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.

Intellectual Property Issues Indicate a Need for Safeguards

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 is the relative ease of re-use 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.

Technical Issues as Well Require Safeguards

There are technical barriers as well to re-analysis, 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.

There Are Both Intellectual Property and Technical Barriers to Re-use of Sharable Data

NIH implementation guidelines address concerns of investigators requested to release their own data. We propose easing barriers to 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.

Data Sharing, Whether Central or Distributed, Requires Standards

Separate technical challenges are presented by such differing models for data sharing as peer-to-peer exchange and central database resources. 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.

Data Are Unique

In draft versions of the NIH policy, selected classes of data were characterized as "unique." We take exception to this, and our comments and recommendations reflect our belief that all data are unique, and their uniqueness derives from the focus, techniques, protocols, selection, and expertise inherent in each investigation.

Authors' Acknowledgments:

Each of the authors acknowledges Neuroinformatics research funding via the Human Brain Project. Our individual support comes from a broad range of contributing Institutes and Agencies.

We thank Stephen H. Koslow for supporting the dissemination of these Principles of Data Sharing. We also thank the editors of *Neuroinformatics* for inviting the publication on which this document is based.

We Offer these Principles for Data Sharing . . .

We propose a series of emendations that we believe would support data sharing, strengthen the policy and its goals, and promote the acceptability and success of data sharing throughout neuroscience.

1. Data Sharing Should Be Universal

Current NIH policy mandates data sharing for high-direct-cost grantees. We urge extension of this policy to all grantees, following evaluation of the experience of this initial group. Even in the absence of such an extended mandate, what we see as the benefits to science, and to society, lead us to urge as well broader, voluntary, sharing of data, and adoptions of norms for the proper presentation and use of sharable data.

2. Data Sharing Should Reflect a Publication Model

Such norms should properly extend those of publication, universally recognized as an essential component of research:

- Just as results are published freely and openly, without restrictions, so most data should be made available for sharing, consonant with appropriate privacy or proprietary restrictions.
- Just as the rewards of publication are universally recognized to outweigh the risks, with active competition for placement in high-citation journals, so data sharing should be encouraged and sought—and rewarded.
- Just as publications are citable archives, so shared data and its locators should be maintained.
- Just as citation of publications is essential, so citation of shared data should be required.
- Just as publication is an appropriate direct cost of research, so data sharing expenses should be supported.

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. We propose a publication model for data sharing, with methods 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 as well the scope of data sharing for some fields; we note that 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 pre-emptive re-use 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.

3. Recognize, But Don't Exploit, Privacy

Appropriate de-identification should allow sharing of human data while maintaining privacy required by both HIPAA and the Common Rule; privacy issues should not be exploited to avoid sharing.

4. A Citation and Credit Paradigm for Data Sharing Must Be Developed and Encouraged

A data sharing policy should include safeguards against re-use 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. Mere acknowledgement may well not be adequate credit for some types of data. Safeguards should require that re-analysis 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.

5. Data Sharing Should Encourage and Enhance Collaborations

We urge encouragement and support for active collaborations as well as mechanisms for passive re-use 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.

6. Multiple Models for Data Sharing Should Be Recognized, Developed, Promoted and Supported

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 NIH's Institutes and Centers. Development and support of specific data sharing methods should thus reflect 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).

7. Continue and Expand Support for Data Sharing

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 supporting informatics 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.