

## PRINCIPLES OF DATA SHARING

We, as Human Brain Project principal investigators, support data sharing, and offer guidelines to promote effective and rewarding policies and methods. These proposals, based on our individual and collaborative efforts, are designed to reduce technological and sociological barriers to data sharing. We offer these guidelines as investigators, neither as representatives of our several institutions nor as NIH policy. Although this perspective is based on our experience advancing neuroinformatics—information technology for exchange and analysis of neuroscience data—these proposals are designed to serve the spectrum of biomedical investigations.

- Data sharing is a fundamental and vital component of science, and research data produced with public funding should in general be available to the community as a scientific resource.
- Sharing of data has great potential for advancing our understanding. Sharing will both enhance the utility of existing data and promote competition in the marketplace of scientific ideas. It will permit reanalyses and meta-analyses beyond the focus 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
- Current NIH policy mandates data sharing for high-direct-cost grantees. We urge broader, voluntary, sharing of data, and adoptions of norms for the proper presentation and use of shareable data.
- 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 publication is timely, so data should be made available without delay.
  - Just as publications are citable archives, so shared data and its locators should be maintained.
  - Just as citation of others' publications is essential to scientific communication, so citation and acknowledgment of shared data should be required.
  - Just as publication costs are recognized as appropriate direct costs of research, so expenses of data sharing should be supported.
- Sharing should not imply relinquishing. Re-use of data should require clear and prominent attribution of the data and acknowledgement of the originator; a standard citation and credit paradigm should be developed and adopted by the bioinformatics community. For extensive or targeted re-use of some classes of data, a collaboration between submitter and user, or an accompanying comment by the submitter, may be appropriate. Investigators and grantee institutions, not third-party repositories, should retain ownership of shareable data. Commercial exploitation of shared data without

appropriate recognition, including possible compensation, may represent misuse of intellectual property.

- The scope of sharable data should be acknowledged as variable and dependent upon the standards and practices of different fields or techniques. Consequently, a variety of models for data sharing may be adopted, including both central databases and peer-to-peer solutions.
- The magnitude of the efforts needed to convert local research data into a form that can be distributed and shared must be recognized. The volume of data produced by some techniques can be immense, and even mid-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.
- Detailed metadata descriptions including protocols and analytic specifications are required to annotate many classes of biomedical data. Submitters have the responsibility to supply, and recipients to interpret, such metadata. Without sharing of metadata as well as data, post-hoc analyses 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.
- Recognized technological and descriptive standards for data and metadata description, archiving, and exchange will enable the technological and sociological infrastructure needed to implement data sharing. Developing and adopting standards is desirable as well for *interoperability*: coordinating disparate data resources without the need for individual database-to-database negotiation to link types of data and descriptors.
- Appropriate de-identification techniques should allow sharing of human data while maintaining privacy required by both HIPAA and the Common Rule.
- Finally, we urge continued encouragement and support for the development of informatics methods enabling investigators to share data with accuracy, accountability, responsibility, and recognition.

Daniel Gardner  
Giorgio A. Ascoli  
Jackson 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 Hines

Gwen A. Jacobs  
Russell E. Jacobs  
Edward G. Jones  
David N. Kennedy  
Daniel Y. Kimberg  
John C. Mazziotta  
Perry 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  
Arthur W. Toga  
David C. Van Essen  
Robert W. Williams  
Stephen T.C. Wong