Data Sharing IV: Ethics of Data Sharing

Joseph J. Fins[1] and Daniel Gardner[2]

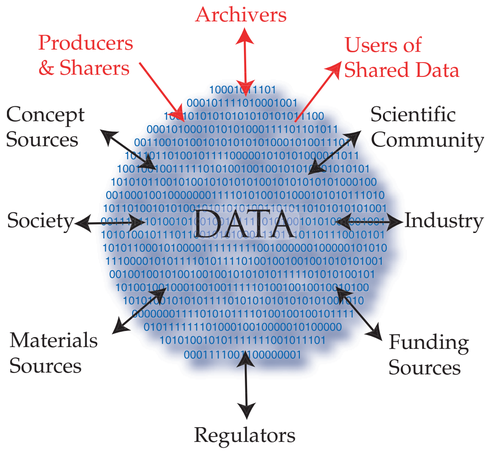
[1]Div. of Medical Ethics, Depts. of Medicine & Public Health, and [2]Lab. of Neuroinformatics, Weill Cornell Med. Coll., NY, NY

Ethical Perspectives on Data Sharing

In addition to generating technological and practical prescriptions, the new thrust for data sharing properly stimulates examination of ethical behavior for both suppliers and users of research data. While continuing to encourage neuroscience data sharing, we provide perspectives on how the culture of research might adapt to these new capabilities. Our goal is to stimulate informed collegial discussion both of ethical guidelines for data sharing and also of potential mechanisms for examination, evaluation, and resolution of conflicts between the right to know and other concerns including intellectual property and privacy. Adoption of such guidelines by experimental and clinical neuroscience communities will increase access to information and foster scientific progress.

Toward this goal, we offer linked sets of proposals for ethical presentation of data for sharing and for ethical use of shared data, coupled to discussions of ownership of, and interests in, shared data.

Shared Data: Multiple Interests

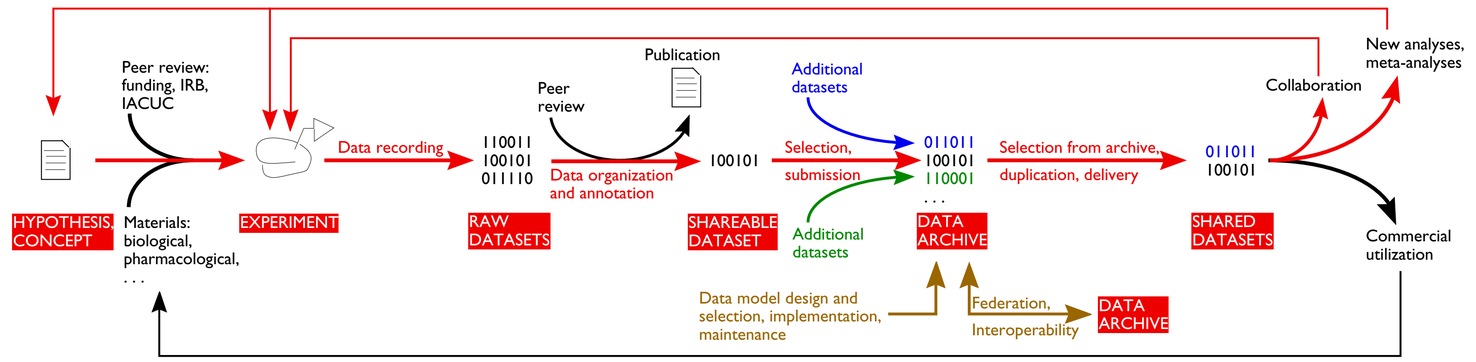


1. Sharing of Data Engages Many Individual and Collective Interests

In addition to the direct interests of producers, submitters, archivers, and users of shared data (shown in red), we identify additional instrumental, influential, and indirect interests. Data production follows the development or adoption of concepts, and often the acquisition of materials and funding. The scientific community, regulatory bodies, industry, and society as a whole each have related and contributory interests.

URL :

- Evolving versions of these and other perspectives may be found at: `datasharing.net`



2. The Data Sharing Pathway: A Biochemical Metaphor

Data sharing may be viewed as a form of recombinase that facilitates formation of hetero-oligomeric complexes of related yet distinct datasets. As for many biochemical pathways, a complex set of reactions, cofactors and modulators contribute. This familiar metaphor aids review of these contributions and provides a substrate for our discussion of interests and proposals for consensus formation and self-regulation.

Producers and Sharers of Data: Interests and Responsibilities

Producers of data should appreciate that they are members of a community that is informed by mutual and reciprocal obligations. Individual datasets, though produced by the effort of a small number of individuals, build upon concurrent and prior developments, transmitted by the (ideally) free and open culture of science. Such prior work includes, in addition to shared data, earlier hypothesis and concept formation, technique refinement and dissemination of results and conclusions, transmitted via publications and personal communications.

Because science is a collective enterprise, producers of data are obliged to share their products with their peers for review and utilization for the advancement of knowledge. As for publication, data sharing should be recognized as an opportunity rather than a burden. These expectations impose certain moral obligations upon both the producers of data and those that share data by providing open access or submitting it to centralized archives. First amongst these is the obligation to ensure access to data:

Sharers of data should ethically provide access to, and ensure accuracy of, data.

Offering data imposes a significant requirement for submitters to ensure accuracy of datasets, and to fully and accurately annotate the data, to prevent mistaken and therefore inappropriate re-use. As a parallel to the requirement that publications include methods sufficient to replicate a study's data, we propose as a guideline that data made available for sharing include sufficient metadata annotations to enable re-use of such data.

Access includes not only posting data, but doing so with timeliness and persistence. The recent NIH guidelines and the practices of several research communities provide several possible timepoints. These include acceptance or appearance of related publications, but several post-publication delay periods have been suggested, from 60 days (for proprietary data) to four years (for work funded by the SBIR mechanism). By analogy with publications, data persistence should be near-permanent, and we view this as more desirable than the 3-years post-grant suggested by the NIH guidelines and based on parallel record-keeping rules.

To Whom Do Data Belong?

To whom *do* data belong? Many classes of data represent, or rely upon, intellectual property of the investigators or institutions responsible for their recording, extraction, or generation. Archiving of data should not remove or modify any of these rights, and databases should recognize these rights rather than requesting or claiming transfer of ownership. In the databases at `neurodatabase.org`, for example, this ownership is emphasized by the following statement:

"Each dataset and metadata description archived in this database remains the intellectual property of the individuals, laboratories, or organizations responsible for the recording, processing, annotation, and submission of the attributed data."

Ownership of data can raise additional ethical questions, and re-use of some data will require recognition and extension of existing data use agreements. These will be especially significant for data obtained from special subject populations or extracted from indigenous or localized species of pharmacological significance.

In our view, recognition of these sets of rights should not conflict with the parallel goal that each dataset is part of a whole that can and should be utilized by members of the scientific community for the advancement of understanding and the betterment of society and human welfare.

If science is indeed a collective enterprise in the spirit of promoting the common good, then data must be shared as methods, results, and conclusions are shared. Proprietary interest in data should not outweigh the advancement of knowledge by the scientific community obtainable from reasonable access to the datasets themselves. As we propose in the next column, intellectual property interests may be respected, and presedrved, by acknowledging the source of the data and by citation of the methods used to obtain the data.

Users of Shared Data: Interests and Responsibilities

Users of shared data, as well as producers of data, are members of this *community of data*, and share in its mutual and reciprocal obligations.

Foremost among the obligations of users of shared data is recognition of the contributions, rights, and interests of producers and sharers. Requirements for acknowledgement and citation of data should be based on, but extend, those for citation of publications. Mere parenthetic or end-noted reference, or listing in an acknowledgement section, may not be sufficient, especially for re-use that is more than incidental. Such extensive re-use or re-interpretation may impose a requirement to notify the original submitter, or to offer the opportunity to comment or co-author. Again from `neurodatabase.org`, users are notified:

"Use of these data requires recognition of contributions of the above parties. For published datasets, this must include citation of literature references accompanying datasets. For unpublished datasets, this should include a citation of the form: (investigator(s) name(s), databased dataset(s)). Extensive re-use requires explicit permission of the submitter; in some cases, an agreed-upon collaboration may be appropriate."

Users of shared data should of course maintain the data's accuracy and integrity. Responsibilities of those who re-use shared data include the need to analyze such data using techniques consistent with the mechanisms, boundaries, experimental conditions, and limitations of the data acquisition protocols, in order to avoid incorrect interpretation of, or extrapolation from, others' data.

Although we favor a model in which archives such as databases do not claim ownership over data, we do urge recognition of the contributions of their developers and maintainers, as well as the scientific communities and funding agencies that support them. Again from `neurodatabase.org`:

"We also ask that re-use of any data from this site include as well an acknowledgment such as: 'Data used in this study were delivered via neurodatabase.org —a neuroinformatics resource funded by the Human Brain Project.' "

Balancing Privacy and Sharing

While respecting the strictures imposed by HIPAA and the Common Rule, we suggest that refusal to share human subject data, when such sharing is ethically justifiable, contradicts the open access model that sharing should promote. In such cases, the benefits of de-identified or data-sharing-agreement-enabled subject data should be weighed against the mandate to protect human subjects and maintain confidentiality.

Towards Community Development of Ethical Guidelines

Finally, we call for broad input to developing ethical guidelines, including workshops and open forums for the neuroscience research community, including producers, sharers, and users of data. Most important, we favor collegial and transparent mechanisms for the resolution of conflicts, structured to balance the interests of:

- producers and sharers of data
- users of shared data
- the community of science
- society

Elements of Ethical Data Sharing:

- Acceptance of the responsibilities of data sharing
- Accessibility of data without unreasonable delays or restrictions
- Assurance of academic freedom to share data and to utilized shareable data
- Maintenance of the integrity of data to be shared
- Responsible stewardship of shared data
- Recognition of the integrity of acquired data
- Assumption of individual responsibility for shared or shareable data, and recognition of individual obligations
- Development and adoption of guidelines incorporating individual, group, and societal rights and responsibilities.

. . . each of these in the service of

- Maximizing the scientific knowledge extractable from available data

Acknowledgments:

The authors acknowledge concept, if not data, from Gerhard Michal's canonical Biochemical Pathways chart, available at: `http://www.expasy.org/cgi-bin/search-biochem-index`

Human Brain Project Neuroinformatics research funded by NIMH and NINDS via MH/NS57153, and by NIMH SBIR MH60538.

LNI Laboratory of Neuroinformatics Weill Medical College of Cornell University