Correspondence

Understanding the open access data movement

Editor's note:
The internet provides unprecedented opportunities to share summary data and complete datasets of original research online, in order that readers and researchers alike can review such data in detail. The implications of this 'open data movement' are considerable, though not entirely resolved. In this expanded Letter to the Editor, Dr. John Doyle comments on several ramifications of the open data movement.

Donald R. Miller MD FRCPC
Editor-in-Chief

To the Editor:
The open data movement has as its goal that useful data, especially scientific data, be freely available to all interested parties without restriction, in the same ethos as open source software and open access scientific publications.[1] This concept applies both to data in raw and processed form, including data as varied as genetic sequences, maps, data from medical experiments, mathematical formulae, and so on. As a specific example, the authors of a study of acupuncture treatment for chronic headache[2] have made their raw data available online for those individuals who may wish to study it in detail.[3]

The open data movement presents a number of interesting practical and ethical issues. For instance, if a researcher reanalyzes posted experimental data from another individual and if he/she is able to produce some valuable new insight worth publishing, should the original researcher be given authorship in the resulting publication, or should the original researcher merely be provided a note of appreciation in the acknowledgement section of the publication? Complicating this issue are various editorial policies pertaining to byline authorship, such as those of the International Committee of Medical Journal Editors:[4]

> "Authorship credit should be based on 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published. Authors should meet conditions 1, 2, and 3."

Another issue concerns the use of data by an investigator who believes that the collection of the original data may have involved a minor breach of ethics or involved a research protocol that was approved by the Ethics Committee / Institutional Review Board at the original institution but likely would not be approved at the institution of the second investigator.

Finally, authors may have a number of reasons to be reluctant to embrace the notion of open access data. First, preparing the data in a standardized format (e.g., Excel spread sheet) and uploading the data to a repository may involve extra steps, money and effort that some scientists may not wish to take. Second, some may have a concern that other researchers may discover errors in that data or may analyze the data using alternate statistical techniques that lead to different conclusions. Third, scientists may feel that they "worked hard" to collect the study data and see not any reason why the data should be made available to "just anyone", especially to people who did not contribute to collecting the study data and could conceivably act in a means not helpful to the original scientists involved. Fourth, the researchers themselves may have specific plans to examine the data using new or alternate methods at some future time, but if they make this data publicly available, other researchers may do exactly this, thus precluding the publication of further papers by the original team.

These are not entirely theoretical concerns. For example, in the field of genomics, one of the early adopters of the open data movement, controversies over academic credit and priority have already occurred.[5] To a limited extent, these issues have been addressed through a public statement by Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical and Medical Publishers (STM):[6]

> "There is considerable controversy in the scholarly community about 'ownership' of and access to data, some of which arises because of the difficulty in distinguishing between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course

of research or other efforts ('data sets'). Another point of difficulty is that in many cases data sets or even smaller sub-sets of data are also provided as an electronic adjunct to a paper submitted to a scholarly journal, either for online publication or simply to allow the referees to verify the conclusions.

We believe that, as a general principle, data sets, the raw data outputs of research, and sets or sub-sets of that data which are submitted with a paper to a journal, should wherever possible be made freely accessible to other scholars. We believe that the best practice for scholarly journal publishers is to separate supporting data from the article itself, and not to require any transfer of or ownership in such data or data sets as a condition of publication of the article in question. Further, we believe that when articles are published that have associated data files, it would be highly desirable, whenever feasible, to provide free access to that data, immediately or shortly after publication, whether the data is hosted on the publisher's own site or elsewhere (even when the article itself is published under a business model which does not make it immediately free to all). We recognize, however, that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost which, in certain circumstances, the host site may need to recover. We also recognize that on occasion the generation of data has been privately funded, and the funding entity may have a particular reason for restricting access to the data (either temporarily or even permanently) but we believe these should be limited exceptions, and that journal publishers themselves should claim no ownership interest in such data. The academic and publishing communities should discuss further (in the context of the debate on the public funding of research) whether more reliable and more permanent sites should be established to host research data."

One proposed approach is to make uploading of raw data a condition of publication. This data might be uploaded to a confidential location while the manuscript is under review (allowing reviewers to look at the work in a manner not usually possible), followed by the data being located at the journal's data repository upon final acceptance.[7] Another approach might be for granting agencies to require that raw data be available online as a condition of funding a project. For instance, the National Institutes of Health (NIH) requires that individuals receiving grants of $500,000 annually or more provide a "data-sharing plan" in the grant application.[8] While this is admittedly a modest requirement, and few NIH grants are in this league, it is a start. Yet another solution, motivated by such tragedies as the Vioxx-related deaths,[9] might be for government legislators and regulators to require that appropriately de-identified raw data for clinical trial patients be published to allow public scrutiny by the scientific community.

Regardless of what approaches, if any, are ultimately implemented, it is likely that issues related to requiring that raw scientific data to be openly available will be the topic of debate among scientists and journal editors for some years to come.

D. John Doyle MD PhD FRCPC
Cleveland Clinic, Cleveland, USA
E-mail: doylej@ccf.org
Support/conflict statement: Support was provided solely from institutional and/or departmental sources. The author has no conflicts of interest to declare.
Accepted for publication July 20, 2007.

References

1 *Vickers AJ.* Whose data set is it anyway? Sharing raw data from randomized trials. Trials 2006; 7: 15.
2 *Vickers AJ, Rees RW, Zollman CE, et al.* Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. BMJ 2004; 328: 744.
3 Available from URL; http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1489946&blobname=1745-6215-7-15-S1.xls (accessed July 9, 2007).
4 *International Committee of Medical Journal Editors.* Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. Available from URL; http://www.icmje.org (accessed July 9, 2007).
5 *Anonymous.* Debates over credit for the annotation of genomes. Nature 2000; 405: 719.
6 *International Association of Scientific, Technical and Medical Publishers.* Databases, data sets, and data accessibility – views and practices of scholarly publishers. Available from URL; http://www.alpsp.org/ForceDownload.asp?id=129 (accessed July 9, 2007).
7 *Altman DG, Cates C.* Authors should make their data available. BMJ 2001; 323: 1069–70.
8 *National Institutes of Health.* Frequently asked questions on data sharing. Available from URL; http://www.grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm (accessed July 9, 2007).
9 *Topol EJ.* Failing the public health—rofecoxib, Merck, and the FDA. N Engl J Med 2004; 351: 1707–9.