Open science, open access and open source software at *Open Medicine*

SALLY MURRAY, STEPHEN CHOI, JOHN HOEY, CLAIRE KENDALL, JAMES MASKALYK, ANITA PALEPU

The authors are on the editorial team of *Open Medicine*.

"Open access to and wide use of research data will enhance the quality and productivity of science systems worldwide."[1]

OPEN MEDICINE IS AN OPEN ACCESS JOURNAL because we believe that free and timely access to research results allows scientific knowledge to be used by all those who need it, not just those who can afford expensive journal subscriptions or user fees for individual articles. But is access to the final polished version of research enough? Could we do more to encourage the collaborative reuse and reanalysis of existing data, or the verification of analyses? Could we move from open access to open science?

Open science is emerging as a collaborative and transparent approach to research. It is the idea that all data (both published and unpublished) should be freely available, and that private interests should not stymie its use by means of copyright, intellectual property rights and patents. It also embraces open access publishing and open source software (rather than proprietary software, which limits others' use of source code and data analysis methods).[2*]

As the name seems to imply, there is no strict definition of open science, but it is inextricably linked to the parallel movements of open access publication and open source software.[3] The varied effects of these related movements are starting to emerge: there is an explosion in the use of free software such as GNU/Linux and open source software for other operating systems; more than 2600 journals have been converted to open access; and studies are finding that articles published in open access journals are cited more widely[4] and that making data openly accessible also increases citation advantage.[5]

Open Medicine itself is using open source software to underpin its journal management, blog, and electronic publishing platform to exemplify what is technically feasible for all journals (rather than just those with big budgets) in scholarly publishing. The Public Knowledge Project, developer of Open Journal Systems (the open source software we use for journal management) has also recently developed Lemon8-XML — a program to automate the conversion of text document formats to publishing layout forms such as XML — ensuring that text is labelled in a way that enables meaningful computer searching of text (see http://pkp.sfu.ca/?q=ojs). For example, it allows us to tag the date of publication and author names as distinct fields so that computers can search and find data that would usually appear as unrecognizable text. In addition to its potentially powerful contribution to data searching, Lemon8 has significant resource implications for the many journals where XML conversion is currently done manually or with proprietary software.

There is wide institutional support for "open" initiatives. Various funding agencies mandate researchers to make their findings available in an open access forum.[6,7] The recent Canadian Institutes of Health Research (CIHR) draft policy on access to CIHR-funded research outputs also requires researchers to state how they intend to make their research accessible to others, with specific reference to final research data ("factual information that is necessary to replicate and verify research results"), original data sets, data sets that are too large to be included in a peer-reviewed publication, and any other data sets supporting the research publication.[6]

Data-sharing has also garnered international support. In 2004 the Organisation for Economic Co-operation and Development (OECD) determined that "Coordinated efforts at national and international levels are needed to broaden access to data from publicly funded research and contribute to the advancement of scientific research and innovation."[1] They subsequently developed the Declaration on Access to Research Data from Public Funding (Annex 1)[1] and recently published a set of guidelines outlining principles that would facilitate cost-effective access to digital research data from public funding.[8]

What kinds of advantages would an initiative like data-sharing offer? For a start, data-sharing opens opportunities for the creative reanalysis of data. Most researchers have had the experience of working single-mindedly with neither the inspiration nor the time to

explore alternative ways to look at their data. Sharing data with other researchers with different research expertise may give rise to new insights, validated findings, or supported and strengthened conclusions. A changing attitude to transparency in research also supports data-sharing: encouraging openness in science promotes integrity, reduces the potential for scientific fraud, and fosters public faith in scientific endeavour.

A recent instance where problems might have been averted was the fraudulent publication of two high-profile papers on stem cell research.[9,10] The publishing journal, *Science*, subsequently convened a committee to review editorial procedures.[11] The committee recommended that more extensive information be included in the published supporting material and asserted that primary data are essential and should therefore be made available to reviewers and readers (http://www.sciencemag.org/cgi/content/full/314/5804/1353/DC1) In a climate where publication and prestige are closely linked and the gains of publication can be great, data-sharing offers a concrete way to monitor and ensure scientific veracity.

It could also be argued that there is an ethical obligation to patients and funding agencies (and to taxpayers) with a stake in scientific research to maximize the benefit to study subjects, who often participate at some personal risk, and to put to best use the money spent on research. These are also opportunity costs to consider: the human subjects who might have volunteered for a different trial, and the funding that could have been spent elsewhere, on other research or on health services. Thus, the limitations on resources provide another good reason for data-sharing.

Of course, some researchers find the idea of sharing data difficult. They may be concerned that others may find flaws in an analysis or gain benefit from data that were difficult or time-consuming to obtain. There may also be concerns about proprietary or classified data, the confidentiality of patient data, the failure to properly attribute data sources or ideas, or the possibility that one's research report may be "scooped."

For the most part these arguments can be countered quite easily: Surely we would want to know if we have made errors, or should be flattered if others think our ideas worthy of replication? With respect to attribution, various options are being considered. Open licensing — as with the Creative Commons license used by *Open Medicine* (see http://creativecommons.org/licenses/by-nc-sa/2.5/ca/) — is one way of dealing with issues such as intellectual property rights, allowing those who provide the original data to retain control over what others do with their work.[3] Creative Commons and the affiliated Science Commons Project are working hard to identify and simplify these kinds of barriers.[12]

The practice of open data-sharing isn't as unlikely as one might think. Recent agreements for data-sharing in genetic science allowed the development of the Human Genome Project, while Jean-Claude Bradley and his team of chemistry researchers post their results on the Internet every day under the banner of Open Notebook Science (http://usefulchem.wikispaces.com/). Using a freely accessible URL, anyone can access their laboratory findings and validate, confirm or repudiate their results. The team also ensure that their findings are indexed on common search engines. Importantly, posting their results like this means that information such as negative or inconclusive results or results that don't fit into published manuscripts are also posted.[2]

Initiatives such as these will become increasingly important as data mining technologies become more sophisticated. With automated computer searching it will be vital to have original data available so that data can be searched and linked in a manner that allows the novel uses of existing research. The development of the semantic Web (searching by linking ideas rather than just words or phrases) offers a critical step toward generating new research hypotheses.[13]

Open Medicine is following the lead of PLoS Medicine (http://journals.plos.org/plosmedicine/policies.php#sharing) and the reproducible research policy of the *Annals of Internal Medicine*.[14] Although the latter was initiated to support research integrity, it also supports a broader data-sharing agenda. We now ask authors to indicate their willingness to share their protocols, datasets, and the statistical codes used for their analysis with other authors, and we encourage authors who publish secondary analyses to use the same Creative Commons license that we use. *Open Medicine* will not handle datasets and other such material directly, but by publishing our authors' willingness to share their original data we hope to encourage fruitful collaboration.

Authors who do not choose to submit these data will not be penalized: we recognize that the acceptance of data-sharing needs time to grow and develop in the scientific community, and we welcome debate and dialogue as we develop our policy on data-sharing. We also need to find ways to deal with some of the problems of data-sharing, such as how to notify other researchers

(or computers that are data-mining) about problems with the data (*e.g.*, in its collection, biases, potential confounders.) and ways to manage original datasets in large databases. Data security, managing data requests and monitoring their appropriate use are other issues that need attention. Perhaps institutions will begin to archive original datasets in the same way that they are beginning to archive their researchers' publications? Google has recently started to help researchers exchange very large datasets (up to 120 terabytes) at no charge provided that the data have no copyright or licensing restrictions (www.earlham.edu/~peters/fos/newsletter/01-02-08.htm#2007). These sorts of options could be more efficient than multiple journals developing their own data repositories.

However its ways and means evolve, an inexorable drive to make science truly open is clear. Indeed, we believe the debate isn't about *whether* we will share data in the future but, rather, about *how* we will share it. Perhaps future researchers will be funded for collecting data with the understanding that all raw data will be deposited in public archives? Perhaps journal editors will require data deposition as a requirement of publication in the same way that they introduced clinical trial registration in 2004 (www.icmje.org/clin_trialup.htm)?

Choosing to share data published in *Open Medicine* gets to the heart of why we believe research is important: to encourage knowledge production and dissemination, with the ultimate aim of improving health. Allowing other researchers access to the data that you have collected considerably extends its value, and an open license encourages ongoing open access to data and the knowledge derived from it. By making their data "open," researchers choose to build a stronger research base, stimulate debate and dialogue and promote public confidence in our published research.

*Open source software ensures that source code is freely available and can be used, changed, improved or redistributed, encouraging code sharing and code integrity (http://en.wikipedia.org/wiki/Open_source_software).

REFERENCES

1. Organisation for Economic Co-operation and Development. Science, technology and innovation for the 21st century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 — Final Communique. 2004 [cited 2007 Oct 18]; Available from: http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.
2. Hooker B. The future of science is open, Part 3: An open science world. 2006 [cited 2007 Oct 18]; Available from: http://3quarksdaily.blogs.com/3quarksdaily/2007/01/the_future_of_s.html.
3. Hooker B. The Future of Science is Open, Part 2: Open Science. 2006 [cited 2007 18th October]. Available from: http://3quarksdaily.blogs.com/3quarksdaily/2006/11/the_future_of_s.html.
4. Eysenbach G. Citation advantage of open access articles *PLoS Biol* 2006:4(5):e157. [Full Text] [CrossRef] [PubMed]
5. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2007; 2(3):e308. [Full Text] [CrossRef] [PubMed]
6. Canadian Institutes of Health Research. *Draft policy on access to CIHR-funded research outputs*. 2007 Apr 3 [cited 2007 Oct 19]. Available from: http://www.cihr-irsc.gc.ca/e/32326.html.
7. National Institutes of Health. *Policy on enhancing public access to archived publications resulting from NIH-funded research*. 2005 [cited 2007 Oct 18]. Available: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-05-022.html.
8. Organisation for Economic Co-operation and Development. *OECD principles and guidelines for access to research data from public funding*. OECD, Paris; 2007.
9. Hwang WS, Young JR, Park JH, Park ES, Lee EG, Koo JM, et al. Evidence of a pluripotent embryonic stem cell line derived from a cloned blastocyst. *Science* 2004;303:1669-1674. [Full Text] [CrossRef]
10. Hwang WS, Roh SI, Lee BC, Kang SK, Kwon DK, Kim S, et al. Patient-specific embryonic stem cells derived from human SCNT blastocysts. *Science* 2005;308:1777-1783.
11. Kennedy D. Responding to fraud. *Science* 2006;314(5804): 1353. [Full Text] [CrossRef] [PubMed]
12. Wilbanks J, Boyle J. Introduction to science commons. 2006 [cited 2007 Oct 18]. Available from: http://sciencecommons.org/wp-content/uploads/ScienceCommons_Concept_Paper.pdf.
13. Machine readability. *Nature* 2006;440(7088):1090. [CrossRef] [PubMed]
14. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. *Ann Intern Med* 2007:146(6):450-553. [Full Text] [PubMed]

Citation: Murray S, Choi S, Hoey J, Kendall C, Maskalyk J, Palepu A. *Open Med* 2008;2(1):e1-3.

Published: 16 January 2007