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

Sally Murray, Stephen Choi, John Hoey, Claire Kendall, James Maskalyk, Anita Palepu

"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 to pay for individual articles. But is access to the final polished version of research enough? Could we do more to encourage collaborative re-use and re-analysis of existing data, or verification of analyses? Could we move from open access to open science?

Open Science is emerging as a new way of approaching collaborative and transparent research. It is the idea that all data (both published and unpublished) should be freely available and that private interests shouldn't stymie its use in the form of copyright, intellectual property and patents. It also embraces open access publishing and open source software (rather than proprietary software limiting 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 impacts of these related movements are starting to emerge: there is an explosion in the use of free software such as Linux and Open Windows, more than 2600 journals have been converted to open access with studies finding that articles published in open access journals are cited more widely [4] while a recent study found that making data openly accessible also increases citation advantage [5].

Open Medicine itself is using all open source software to underpin its journal management, blog and electronic publishing platform as an exemplar for what is technically feasible for all journals (rather than just those with big budgets) in scholarly publishing. Open Journal Systems (the open source software we use for journal management) has also recently developed Lemon8 - a program for automating conversion of HTML to XML – ensuring that text labeled in a way that allows meaningful computer searching of text (see http://pkp.sfu.ca/?q=ojs). For example, it allows us to 'tag' 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 as XML conversion is currently done manually at many journals or through the use of 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] and the recent CIHR Draft Policy on Access to CIHR-funded Research Outputs also requires researchers to state how they intend to make their research outputs 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 the peer-reviewed publication, and any other data sets supporting the research publication [6].

Data sharing has also garnered international support. In 2004 the Organization for Economic Cooperation 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 facilitating 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 offers the opportunity for creative re-analysis of data. Most of us have worked single-mindedly before with neither inspiration nor time to explore alternative ways to look at our data. Sharing our data with other researchers with different research expertise may allow insight to a given health problem other than that which we were aware. Other researchers are also able to validate our findings, supporting and strengthening their conclusions. A changing attitude to transparency in research also supports data sharing; encouraging openness in science promotes integrity, challenges potential problems such as fraud, and encourages public faith in the scientific endeavor.

A recent example 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 reviewing their editorial procedures [11]. The Committee recommended that more extensive information be put in the published supporting material and that primary data are essential and should be available to reviewers and readers (see http://www.sciencemag.org/cgi/content/full/314/5804/1353/DC1) In a climate where publication and prestige are closely linked and publication gains 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 tax-payers) involved in scientific research to maximize the benefit of their study subject participation and personal risk, and dollars spent on research and its output. These are of course human volunteers that might not have volunteered, and dollars that could have been spent elsewhere - either on

other research or service provision etc - with this opportunity cost in mind the argument for data sharing gains currency.

Of course some researchers find the idea of sharing data difficult. Sharing original data means that others may find flaws in our analysis or gain benefit from data that were difficult or time-consuming to obtain. There may also be problems with proprietary or classified data, confidentiality of patient data, or concerns that others won't attribute their source of data, 'steal' ideas, or publish them before we can.

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

The practice of open data sharing isn't as unlikely as one might initially think. Recent agreements for data sharing in the genetics field 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 (see http://usefulchem.wikispaces.com/) under the banner of Open Notebook Science. Using a freely accessible URL, anyone can access their laboratory findings and validate, confirm or repudiate their results. The team also ensure 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].

These sorts of initiatives 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 allowing novel uses of existing research. The development of the semantic web (searching the web by linking ideas rather than just words or phrases) offers a critical step forward in generating these kinds of research hypotheses [13].

At Open Medicine we are following the lead of our colleagues at PLoS Medicine (see http://journals.plos.org/plosmedicine/policies.php#sharing) and the reproducible research policy of the Annals of Internal Medicine [14]. While 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 encourage authors who publish secondary analyses to use the same Creative Commons license that we use. Open Medicine will not handle datasets

etc directly; 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: ideas like these need time to grow and develop in the scientific community and we welcome debate and dialogue in the growth of our policy around data sharing. We also need to find ways to deal with some of the problems of data sharing, for example how to notify other researchers (or computers that are data mining) about problems with the data (e.g., in its collection, biases, potential confounders etc) and ways to manage original datasets in large databases. Data security, managing data requests and monitoring its appropriate use also 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 also recently started helping researchers exchange very large datasets (up to 120 terabytes) at no charge provided that the data have no copyright or licensing restriction (see http://www.earlham.edu/~peters/fos/newsletter/01-02-08.htm#2007). These sorts of options could be more efficient than multiple journals with less necessary infrastructure taking on this task.

However it evolves, an inexorable drive to make science truly open is clear. Indeed, we believe the debate isn't 'if' we will share data in the future but '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? (see http://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: encouraging knowledge production and its dissemination to improve 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 your data 'open' you are choosing to build a stronger research base, stimulate debate and dialogue and promote public confidence in our published research.

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

*AMT- I am having Endnote problems: can you please ensure that the author in refs 3 and 4 is Bill H**oo**ker.*