More damn lies about data access

More data than we can handle is no excuse to give up our efforts to promote data access, but it may make us think about new ways to make it sustainable.

[This draft was written by Myles Axton in the hope that participants of the Sage Congress will write an Nature Genetics Editorial in the manner of Tom Sawyer's white fence (Twain M. 1876). All contributions received by April 10th 2011 will be attributed.]

Sharing is the concept promoted by kindergarten teachers while the children under their instruction are learning from one another to make deals that are mutually beneficial and improving their ability to negotiate. Data sharing is a misnomer for what we are doing as adults. In science, data are either traded in a market for information or consigned to write-only databases. We should perhaps discuss incentives to an effective ideas market rather than restrict our discussion to ensuring open data deposition.

Data are often inaccessible or not usefully presented because some data producers do not want data to be used by everyone. There is a natural barter process among data holders that generates skeptical evaluation as well as discoveries. Whether this process among privileged labs is more productive than the scrutiny of many eyes on open data is not for discussion, we clearly need both. What has not been widely appreciated is that funders with their data access mandates risk short-circuiting the economy of knowledge production. It also costs to format and curate data and those costs are not fully borne by funders because they have no metrics with which to judge curation effort and they do not want endless commitment to resources that may not be used.

NCBI and EBI have been the trusted and accountable partners researchers have relied upon for sustainable data storage, without which we would be hard pressed to promote any data access model. However, recently, NCBI has closed two of its repositories for raw nucleotide sequence data http://www.ncbi.nlm.nih.gov/sra mainly because of an explosion of next generation sequencing runs that cannot be readily reduced to unambiguous calls. There are alternative places to store data. Until SRA returns, journals may have to trust the stability of the links to institutional databases that authors provide and handle the complaints of frustrated data users. To this end, it may help to have data producers publish citable data management plans explaining how to access and use the data. Cloud computing may eventually provide a solution to sequence data storage provided there is a suitable business model. Providers will only keep the data from which they can make money and reputation.

There is hope that we can arrive at solutions because we share common interests in promoting data access. The principle that your reputation is made in the labs of others means that good citizenship is good business. You simply cannot publish enough papers on your data yourself to equal the productivity of the researchers you inspire. The interests of the journals you publish in and the institutions and agencies that fund your work are likewise aligned to do everything they can to enable data sharing by their need to demonstrate that they are contributing to the impact of the datasets you produce.

The sustainability of data access is often discussed by publishers of journals of record who have in part ensured the stable accessibility of the (albeit smaller) datasets of the past. Journals could step up and charge depositors what it really costs to make a large dataset accessible in perpetuity. If they do charge users for access, the price should be transparently related to distribution costs and the need to sustain the archive. Maybe sequence data will not accumulate exponentially forever. Simple discounting suggests that it will be cheaper to resequence genomes than to store existing reads. It may be that many large datasets are not really useful for research but are consigned to public databases as merely the burn-ins for technology that moves on. Still, unless we develop suitable metrics for data citation and promote their adoption, the experiment to evaluate the utility of data has not been done. Maybe the funders' need for data to be useful coupled with the incentive for publishers to make open access sustainable would provide the motive to do this properly.

Other incentives that can help with data access are to link author and contributor roles to data accessions and to link data accessions semantically into a concept web. Attribution licenses for articles and data are a good concept but lack enforcement. Attribution is also currently insufficiently granular both at the data level and with respect to the author roles. There are no agreed data citation metrics and examples of resource reallocation or career decisions to point to. In discussions about ORCID (http://www.orcid.org/) and researcher disambiguation, it is essential that we discuss distributed as well as centralized ways for researchers to track and display their career achievements, connections and productivity. Popular sites like PubMed and Wikipedia provide places to start developing metrics, but it is important to give researchers a choice of individual, institutional, funder, journal and consortium sites to choose from and to agree on what we are counting. Everyone needs to be guaranteed minimum space and the ability to garden their own reputations. Institutions need to experiment with evaluating the totality of scholarly productivity, for example by participating in experiments like Vivo (http://vivoweb.org/).

Links are often unequal in not giving credit in both directions. Even asynchrony in establishing bidirectional attribution between databases or between databases and journals can lead to loss of timely attribution. The semantic web will work fine initially with small numbers of trusted early adopter databases with strong hierarchic organization. Reliance on linked data URIs may limit its adoption and eventual success. Why? Many small databases are happier with relational organization and only link to the outside world via generic high level container URIs that leave staff free to move items around at will. Unless these items carry indelible universal identifiers (UUIDs) there is no external global reference. It may therefore be better to label everything twice (belt and braces) as argued by Josh Knauer in his critique of Tim Berners-Lee's Concept Web http://rhiza.com/2010/05/27/berners-lee/.

The final barrier to data access is related to the way in which we currently protect human subjects and obtain their informed consent. Institutional review boards protect individual institutions as well as the research subjects they recruit. This concept has worked well but without a duty to put their rules and deliberations in the open, IRBs depart progressively from universally agreed ethical principles in ways that make it ever harder to combine data collected at multiple centers into well powered studies. Professional bioethicists dig ever deeper and address subsets of the issues in their field. Research subjects given multiple choices about protection and access to their data distribute pretty evenly over the options. A requirement for IRBs to deposit their rules and deliberations on the web might be the only way to encourage harmonized and transferrable consents because no central agency can ever persuade diverse institutions to adopt an "IRB in a box" franchise.
[Please rewrite this editorial to suit the data access aims of the Sage Congress]

http://www.nature.com/nmeth/journal/v7/n7/full/nmeth0710-495.html
http://www.sciencemag.org/content/331/6018/666.full
http://www.w3.org/DesignIssues/LinkedData.html

Twain M (1876) The Adventures of Tom Sawyer. Barnes and Noble Classics, New York, NY, USA, 2003, ISBN 1-59308-068-9