BioMed Central's position statement on open data

Increasing transparency in scientific research has always been at the core of BioMed Central's strategy. Now, after more than a decade of open access research publishing, BioMed Central is extending its efforts in encouraging and facilitating transparency in scholarly communication beyond articles and onto data.

The data underlying published articles are an increasingly integral part of the scientific record. A principle benefit of sharing scientific data is that it facilitates the discovery of new knowledge, as the outputs of previous scientific research can be readily re-used, evaluated and incorporated into new research [1]. Sharing raw data reduces the potential for bias, improving the scientific evidence base and, with agreement on standard, machine-readable formats, we can help maximize the potential of the semantic web and facilitate automated knowledge discovery and drive further scientific innovation. We believe the future of scholarly communication depends on a commitment to data.

Why make data more open?

Sharing raw data is not a new idea [2] and fits fundamentally with the concept of reproducibility, a core principle of the scientific method. More and more reasons – and supporting evidence – to share data are emerging, however [3-10]. These include:

- Reproducing and checking analyses
- Secondary hypothesis testing
- Comparisons with previous studies
- Simplifying and enhancing subsequent systematic reviews and meta-analyses
- Teaching
- Reduction of error and fraud
- Integration with previous and future work
- Increasing academic credit (citations)
- Funder and journal requirements to share
- Reducing the potential for duplication of effort
- Interdisciplinary research

What data to make more open?

Genomic and microarray researchers have generally led the field in biomedical data sharing and standardization and BioMed Central already links to a number of databases in our journals' instructions for authors, such as GenBank [11] and ArrayExpress [12]. What comprises 'the data set' in some other fields is more open to discussion. Some BioMed Central editors , in working with clinical trialists to agree best practice for data sharing and publication, has previously defined the dataset to be, at the minimum, as "containing the minimum level of detail necessary to reproduce all numbers reported in the paper" [13].

BioMed Central is keen to enable data to be made open in a way that readily enables automated harvesting and re-use, and to this end we support the adoption of standard data formats (data standards). We believe the scientific community is best placed to agree on data standards and their adoption rather than publishers, but BioMed Central aims to maximize the opportunities to promote re-usable data formats. The publication of the journal *BMC Research Notes* was partly driven by this goal, and the journal recently launched a call for contributions to a major collection of commissioned articles on data standards [14].

Where to make data more open?

Authors of BioMed Central articles are able to publish almost unlimited numbers of additional data files (up to 20Mb each) and tabular data with their articles, enabling data publication. Publishing data as supplementary material has been a topic of recent debate [15, 16], but if no established domain- or institution-specific repository for supplementary material and data exists, then we believe publishing it in an online journal, which inherently assures the permanence of its content, is much more favourable to hosting on authors' personal websites or not having the material available at all. Furthermore, journals such as *BMC Research Notes* publishes Data Notes, which are a short description of a biomedical dataset, with the data being readily attributable to a source [17]. Such publications make the data the core component of the article. We recognize that many datasets are beyond what can be published in the context of journal article supplementary material, and therefore support the deposition of datasets in appropriate repositories, if one exists. To enable linking to data from published articles we recommend that data be deposited in archives and databases that ensure permanence by a system equivalent to a digital object identifier (DOI), as is used for scientific articles, or a permanent accession number system that can be reliably linked to. These identifiers can then be cited and incorporated in article reference lists, full texts and, potentially, metadata. We see a key role of publishers as being able to provide clear and permanent links to data hosted in repositories and are working with numerous general and domain- or institution-specific data repositories to establish processes and policies for linking data to articles [18]. As more repositories emerge, we hope to produce a resource for our authors that informs of the available repositories, and partner journals, for their work.

When to make data more open?

The decision to mandate data deposition as a condition of publication is another decision best made by the scientific community concerned rather than a single journal or publisher as, for example, has been established in the microarray and evolutionary biology communities [19]. We will, therefore, support data publication when it is mandated, but will also enable, encourage and recognize [20] data sharing and publication on a voluntary basis for scientists wishing to show leadership in their field.

When appropriate we encourage pre-publication data sharing and, through progressive publication policies [21], recommend data sharing and release does preclude publication in a peer-reviewed journal. Making data available during peer review is also encouraged; publication in a BioMed Central journal has always implied that readily reproducible materials and raw data will be available to any scientist on request. It is recognized, however, that not all data can be made fully open access, for example when it

would put privacy at risk or is otherwise sensitive. In such cases embargoes or restricted access are a logical solution [13].

How to make data more open?

We believe the concept of open data, analogous to our policy on open access, goes beyond making data freely accessible. Data should also be free to distribute, copy, re-format, and integrate into new research, without legal impediments.

All research articles published in BioMed Central journals are published under the Creative Commons attribution licence [22] (CC-BY), with which authors retain the copyright to their work. This licence allows unrestricted distribution and re-use provided that the original article is cited. We support the Panton Principles for open data in science [23] and open data should therefore mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. We encourage the use of fully open file formats wherever possible.

As raw data are not creative works, copyright cannot be asserted over raw data and facts, only the way in which they are presented. Indeed, publishers have previously been urged not to require transfer of copyright for data [24]. Moreover, licenses that require attribution, such as CC-BY, can be applied differently in different jurisdictions, and licences that restrict certain or all types of re-use of data, such as commercial use, could prevent the effective future use and integration of data by others, particularly when providing full attribution for all data in a large collection may not be practical.

Therefore, to eliminate potential legal impediments to integration and re-use of data, specifically, and to help enable long-term interoperability of data we believe an appropriate licence or waiver specific to data should be applied, and made explicit by the authors and publishers. There are a number of conformant licences [25] for open data, of which Creative Commons CC0 [26] is widely recognised. Under CC0, authors waive all of his or her rights to the work worldwide under copyright law and all related or neighboring legal rights he or she had in the work, to the extent allowable by law.

To clarify the difference between the legal right of attribution (copyright) and the cultural scientific norm of attribution (citation) a way forward would be to require that from a specific date, any author submitting to a BioMed Central journal agrees to dedicate the data elements of their article and supplementary material to the public domain and apply the CC0 licence.

Overcoming barriers to making data more open

The barriers to the open sharing of scientific data have been well documented and we do not intend to repeat them here – not least because there are counter arguments to them all. Proposing authors apply a new license to data that removes a legal requirement for citation however, warrants some further discussion of academic credit. Citation of scientific articles, in the typical context of writing a research article, is not a legal requirement but is a widely accepted cultural norm in science [27]. Therefore, by

waiving attribution rights to the data elements of published articles, especially if they are part of a published (citable) article as an additional file, we believe authors will not risk the loss of academic credit by applying CC0 to data within published articles.

An example of where attribution might not be feasible is where data from large numbers of different sources are combined in a derivative work. Indeed, large, domain-specific scientific databases for data types that are free to other researchers without requiring attribution to an individual or group have been in existence for many years and have been fundamental to key scientific advances [28]. The genomics community, again, has shown leadership in establishing such a framework for an "information commons", engrained in the Bermuda Principles, and have established built-in temporal latencies [29] to data for knowledge (when data are released), and rights (when rights restricting use removed).

Furthermore, the application of appropriate licences to different components of the products of research aims to ensure attribution, facilitate sharing of knowledge and ensure reproducibility [30]. By submitting a manuscript and all associated files to a BioMed Central journal authors have always confirmed that they agree to all terms of the BioMed Central Copyright and License Agreement [22]. This position statement, and suggestion that authors apply CC0 to data, therefore should rather provide clarity to the licensing of specific components of published articles and does not necessarily represent a substantial change to the overall licence agreement for authors' published work. And in instances where only the CC0 licensed aspects of published articles are used in derivative works we strongly encourage attribution (citation) wherever possible and appropriate, consistent with academic norms and initiatives such as DataCite [31]. This is also consistent with attribution policies of the growing number of archives for datasets [32].Encouraging the linking of articles to data in repositories, as described above, should also facilitate academic credit to be gained for data.

What do we mean by data?

Files types such as CSV, XML and RDF are strongly associated with raw data [33] but, in short, "data is difficult" [34]. Comprehensively defining the specific file types that will contain data, and therefore be CC0 licensed, is not currently feasible (although we plan to prepare guidance on the preparation of additional file types and data formats). We therefore define data at a less granular level – the raw, non-copyrightable facts provided in a BioMed Central article or its associated additional files, which are potentially available for harvesting and re-use.

Source code (software)

We do not consider open source code for software to be open data. Specific licences have been developed for source code for software and BioMed Central recommends that source code be made available via an Open Source Initiative [35]compliant licence before submission of the related manuscript.

References

1. Lynch C: **Jim Gray's Fourth Paradigm and the Construction of the Scientific Record**. In The Fourth Paradigm (2009)
2. Galton F: **Biometry.** *Biometrika* 1901 , **1:**7-10.
3. Vickers AJ: **Whose data set is it anyway? Sharing raw data from randomized trials**. *Trials* 2006 , 7:15
4. Kirwan JR: **Making** original **data from clinical studies available for alternative analyses.** *J Rheumatol* 1997 , **24:**822-825
5. Freese J: **Replication standards for quantitative social science: why not sociology?** *Sociol Methods Res* 2007 , **36:**153
6. Smith R, Roberts I: **Patient safety requires a new way to publish clinical trials.** *PLoS Clin Trials* 2006 , **1:**e6.
7. Piwowar HA, Day RS, Fridsma DB: **Sharing detailed research data is associated with increased citation rate.** *PLoS ONE* 2007 , **2:**e308.
8. Hersh WR: **Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance**. *Am J Manag Care*. 2007, 13(6 Part 1):277-8
9. Kolata, G: **Sharing of Data Leads to Progress on Alzheimer's**. New York Times [http://www.nytimes.com/2010/08/13/health/research/13alzheimer.html?_r=2&hp]
10. RIN: **To share or not to share: publication and quality assurance of research data outputs.** *Research Information Network* 2008 , 1-56
11. GenBank [http://www.ncbi.nlm.nih.gov/genbank/]
12. ArrayExpress [http://www.ebi.ac.uk/microarray-as/ae/]
13. Hrynaszkiewicz I, Norton MN, Vickers AJ, Altman DG: **Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers**. *Trials* 2010, 11:9
14. Hrynaszkiewicz I: **A call for *BMC Research Notes* contributions promoting best practice in data standardization, sharing and publication**. *BMC Research Notes* 2010, 3:235
15. Maunsell J:*The Journal of Neuroscience*, 2010, 30(32):10599-10600
16. Piwowar H: **Supplementary materials is a stopgap for data archiving** [http://researchremix.wordpress.com/2010/08/13/supplementary-materials-is-a-stopgap-for-data-archiving/]
17. Instructions for *BMC Research Notes* authors of Data Note articles [http://www.biomedcentral.com/bmcresnotes/ifora/?txt_jou_id=4005&txt_mst_id=104807]
18. The Yale Law School Roundtable on Data and Code Sharing: **Addressing the Need for Data and - Progress in computational science** [http://www.stanford.edu/~vcs/papers/RoundtableDeclaration2010.pdf]
19. Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ: **Data Archiving**. *Am Nat* 2010, 175: 145–146
20. Hrynaszkiewicz I: **BioMed Central Open Data Award: winner to be announced this week!** [http://blog.okfn.org/2010/06/08/biomed-central-open-data-award-winner-to-be-announced-this-week/]

21. Guidance on duplicate publication for BioMed Central journal Editors [http://www.biomedcentral.com/info/about/duplicatepublication]
22. BioMed Central copyright and license agreement [http://www.biomedcentral.com/info/authors/license/]
23. Panton Principles for Open Data in Science [http://pantonprinciples.org/]
24. Association of Learned and Professional Society Publishers (ALPSP) and International Association of Scientific, Technical, & Medical Publishers (STM): **ALPSP and STM issue joint statement clarifying publishers' views on access to raw data, data sets, and databases** (20 June 2006). [http://www.alpsp.org/ForceDownload.asp?id=128]
25. Open Definition Conformant Data Licenses [http://www.opendefinition.org/licenses/#Data]
26. Creative Commons CC0 1.0 Universal [http://creativecommons.org/publicdomain/zero/1.0/]
27. Wilbanks J: **Marking and Tagging the Public Domain** [http://scienceblogs.com/commonknowledge/2010/08/i_am_cribbing_significant_amou.php]
28. The Human Genome Project [http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml]
29. Contreras JL: **Prepublication Data Release, Latency, and Genome Commons**. *Science*, 2010 10.1126/science.1189253
30. Stodden, V: **Enabling Reproducible Research: Open Licensing for Scientific Innovation.** *International Journal of Communications Law and Policy*, Issue 13, 2009.
31. DataCite [http://www.tib-hannover.de/fileadmin/datacite/index.html]
32. Dryad Joint Data Archiving Policy (JDAP)[http://www.datadryad.org/jdap]
33. Wallis R: **Linked Open Data and Pavlova** [http://blogs.talis.com/nodalities/2010/08/the-linked-open-data-and-pavlova.php]
34. Murray-Rust P: **Open Data: why I need the Open Knowledge Foundation** [http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=2471]
35. Open Source Initiative [http://www.opensource.org/licenses/alphabetical]