The need and drive for open data in biomedical publishing

The concept of open data goes beyond making data freely available. Data must also be free to reuse and build upon without legal or technical impediments. Funder and journal policies for data sharing and the growing open science movement are helping open data to spread across biomedical sub-disciplines. Editors should embrace open data to ensure that their decisions can stand up to close scrutiny; journals need open data to help them fulfil their stated goals, and publishers should utilize open data and data publication to serve the growing sector of the scientific community requiring it as a service, and to continue developing novel forms of scholarly communication in an increasingly data-intensive scholarly communication environment.

IAIN HRYNASZKIEWICZ
Journal Publisher
BioMed Central

What do we mean by open data?

There is a growing need for, and growth of, open data in biomedicine. But what do we actually mean by open data? The term can mean different things to different people. With increasing numbers of 'open data' initiatives across the world wide web, particularly in governmental data, we risk confusing the concept of public accessibility (free access to data) with that of interoperability and integration, and ensuring data are reusable and redistributable[1] .

This is analogous to the challenge faced – arguably, still faced – by the open access movement. The Berlin Declaration on Open Access to Scientific Knowledge[2] stipulated that by open access not only should articles be freely and permanently available, they should be free for others to reuse, redistribute and make derivative works. But many publishers continue to assert that content is open access when there are a variety of restrictions on reuse, particularly regarding commercial use [3].

The Open Knowledge Foundation[4] , Creative Commons[5], Panton Principles for Open Data in Science[6] , and the open access publisher BioMed Central[7], have all expressed that open data should mean more than data being freely accessible. Science depends on reproducibility of results, and being able to build on previous findings and reuse data to drive new discoveries without legal or technical impediments. This ideally requires data to be placed explicitly in the public domain by the application of an appropriate licence or waiver of rights specific to data, such as Creative Commons CC0[8] .

Why does public domain dedication of data matter? Licences, such as Creative Commons attribution licences, that legally require attribution for reuse, can lead to unmanageable attribution requirements for data gathered from multiple sources[9]. Imagine all of the thousands of contributors to the data in the Human Genome Project exerting legal rights of attribution or transfer agreements each time a post-doc researcher runs a query on the database, and you begin to see how this could become problematic.

Drivers for open data in biomedicine

A number of reports[10,11] and surveys[12] have identified benefits of data sharing in the life sciences for the public good, economy and for the advancement of knowledge. Sharing detailed research data has been associated with increased citations[13], but substantial empirical evidence of the benefits and rewards of data sharing and publication for the individual scientist is still being gathered. Meanwhile, policies and mandates from

funding agencies, institutions and journals are key drivers for change in researcher (author) behaviour.

A growing number of biomedical research funding agencies have data sharing policies (see Table 1). And in January 2011, 17 major international health research funding agencies, including the World Health Organization and the Bill and Melinda Gates Foundation, committed to working together to support data sharing[14]. This is logical, given data are the main product of the investment of research funding agency grants, often funded by public money and, therefore, preserving and archiving raw data in a reusable form maximizes its value[15] .

Creating at least one published, citable article about a scientific research project – preferably in a high-ranking journal – remains an essential part of the research lifecycle. Journals and their submission or publication requirements can influence author behaviour, as authors endeavour to meet the demands of the editors of their preferred publication. Of course, journal editorial policies are usually consensus driven, and motivated by meeting the needs of the scientific communities and audiences they serve. Journal policies have proven to be effective in changing author behaviour, such as requiring the prospective registration of clinical trials[16].

The leading life-science journal *Nature* requires that, as a condition of publication, its authors 'make materials, data and associated protocols promptly available to readers without undue qualifications in material transfer agreements' and that supporting data be available to editors and peer reviewers after submission. It also specifies how it deals with infringements of the policy, which includes publishing corrections or refusing publication[17]. But *Nature* is a high-impact journal with substantial resources, so how do less well-established publications treat this issue? BioMed Central, which publishes more than 200 journals across biology and medicine, requires that authors confirm on submission that they will provide data to other scientists on request[18], and the Public Library of Science author information, even more strongly, states that 'publication is conditional upon the agreement of authors to make freely available any materials and information associated with their publication'[19].

The high-ranking clinical medical journals *Annals of Internal Medicine*[20] and the *BMJ*[21] take an alternative approach. They require a statement as to the availability of supporting data rather than implying data sharing as a condition of submission or publication.

Journal policies have been associated with increased sharing of genetic sequence data (where a number of well-established repositories for the data exist)[22] but compliance with policies that rely on other data types have found low compliance rates (25% from 141 published articles in psychology[23] and one in ten from a sample of ten published clinical trials[24]).

A solution has been proposed in the form of the Joint Data Archiving Policy, which has been signed by a consortium of journals in ecology and evolutionary biology and requires that supporting data sets be archived in 'an appropriate public repository' and a link to the supporting data set(s) be included in the published article[25]. The Dryad repository is one such appropriate repository, which will host myriad data file types (unlike highly structured databases such as GenBank), promotes data citation by assigning digital object identifiers (DOIs) to data sets, and promotes reusability by requiring CC0 as its default waiver for published data sets[26].

Funding agency	**Policy available at**
Wellcome Trust	http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm
National Institutes of Health	http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
Medical Research Council	http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm
National Science Foundation	http://www.nsf.gov/bfa/dias/policy/dmp.jsp
Genome Council	http://www.genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf
European Research Council	http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf
Cancer Research UK	http://science.cancerresearchuk.org/funding/terms-conditions-and-policies/policy-data-sharing/
Biotechnology & Biological Sciences Research Council	http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf

Table 1. Major life science funding agencies with data-sharing policies

The need for open data in biomedical publishing

Reporting bias and distortion of the evidence base
The mission statements of medical journals often aspire to improving clinical decision-making or human health and wellbeing, but lack of access to data underlying publications, and data generated during clinical trials, can have the opposite outcome. Suppression of data potentially relevant to human health for monetary gain by pharmaceutical companies is indefensible, but – albeit inadvertently or subconsciously – incremental contributions of editors, journals and peer reviewers[27] during the publication process may also be distorting the clinical evidence base and, consequently, having deleterious effects on human health.

In autumn 2010, the widely-prescribed anti-depressant reboxetine was found to be ineffective or potentially harmful when previously unpublished data, from an alarming 74% (3033/4098) of patients from 13 clinical trials, were included in a systematic review and meta-analysis[28]. This is the latest in a number of high-profile cases (including celecoxib[29] and rosiglitazone[30]) of opacity in raw clinical data leading to reporting bias. This is the phenomenon whereby articles reporting results favouring the medical intervention being studied (i.e. positive rather than negative results) are more likely to be published – and more likely to be published quickly, in high-impact journals and published multiple times. (For a review, see McGauran et al[31]). Open data in medicine will enable journals and publishers to better fulfil their aims of advancing science and medicine by enabling more balanced and transparent reporting of research which will, ultimately, benefit human health.

In non-human biology the content of published articles may seem less immediately able to impact human health, but across science a sharing and publication of raw data would logically be predicted to reduce the potential for error and fraud[32].

The unique challenge of human subjects' research
Open medical data has much potential, but publishing data that have arisen from the doctor–patient relationship inherently carries risks to individual privacy, unless explicit consent for publication has been obtained. This is an issue for publishers, editors and journals, and indeed all those involved in the data acquisition and dissemination process, given the implications under privacy and data protection laws. In the age of open access and open data, de-identification of personal data for publication (where consent has not been obtained) is challenging. Published guidelines for authors, editors and peer reviewers[33] of clinical data sets recommend data sets including three or more indirect identifiers, such as gender or ethnicity, should be independently reviewed to assess the risk of patients being identified (see Table 2).

The same principle of privacy protection is true of publishing medical case reports, which now usually requires explicit consent for publication from living individuals described in cases[34].

Lessons in open data from (genome) biology

Inter-disciplinary and international research, and research conducted jointly by academia and industry, is growing, in part facilitated by the web and open access. The Human Genome Project was 'a watershed moment' for open sharing of scientific data across boundaries, as many pharmaceutical companies backed this collaborative effort instead of their propriety projects[35]. Despite human genome data driving new, commercially valuable, drug targets and discoveries (as well as discoveries in ecology, agriculture and beyond), participating commercial entities recognize that data are just the beginning of the drug-discovery process. They further recognize there is more to be gained from sharing without exerting intellectual property or patents in early stages of data collection. Indeed, scientific web services – machines – depend on immediate and unfettered access to data, exemplified for example by the GenBank database. Furthermore, the genomics community, via the Bermuda Principles, has agreed built-in temporal latencies that set out when data should be released, and when rights restricting use are

Ethnicity	Occupation	Place of treatment
White British	Doctor	London (England)
Black Caribbean	Judge	Paisley (Scotland)

Table 2. Is the second hypothetical patient anonymous with certainty? How just three indirect identifiers, which in isolation would be no cause of concern, when associated with an individual could potentially put privacy at risk

removed, allowing researchers defined periods (e.g. 12 months) for exclusive use of data for their projects – and papers[36].

The Sage Bionetworks initiative hopes to transfer the access principles ingrained in the Human Genome Project to human disease biology and biological networks (the study of changes at the molecular level linked to disease symptoms and traits). Sage Bionetworks aims to 'be the steward of the data and associated systems' and produce networked models of disease (from genomic, proteomic, metabolomic and clinical data), for several currently fragmented fields with no common repository for data. Importantly, this initiative is committed to ensuring all data are in the public domain by waiving all database and other rights to ensure reusability without restrictions[37].

Exploring the role of publishers in open data

Online publishers are service providers, who must respond to the needs of today's scientists to facilitate rapid dissemination and transfer of knowledge and, invariably, to stay in business. So changes in scientists' behaviour, such as those set out above, are important for publishers. For example, there are growing numbers of institutional, funder and scientific subject-specific repositories for data. Publishers such as BioMed Central are responding to this and developing links to data from published articles, integrating data viewing software with their content, and participating in initiatives to agree best practices for data publication and citation[38]. Some journal publishers are also, effectively, data publishers, by hosting online supplementary data files. Publishing supplementary material has been the source of much debate in recent months, as some journals have claimed it puts too many demands on peer reviewers, or that it moves important material from the article to non-printed supplements. However, I would argue it is unrealistic to expect every reviewer of an article to reanalyze supplementary or repository-held data, and for online open access journals space is virtually unlimited. Furthermore, many biomedical subdomains are yet to routinely post data in a repository (as is common in genomics) or indeed have a repository in which to deposit their data, making online supplementary files an important interim venue for data[39].

Gaining academic credit (in the form of citations) for data sharing remains a challenge. Publishers such as BioMed Central and the Ecological Society of America are addressing this issue by offering publication of 'data notes' (in *BMC Research Notes*) or 'data papers' (in *Ecological Archives*). These articles put a biomedical data set or database at the core of the publication, with the peer-reviewed journal article acting more as a wrap-around for the data, such that it is discoverable, indexable and citable via standard scholarly search engines and databases. *BMC Research Notes* has taken this concept further by offering to publish, as educational articles, biomedical domain-specific data standards (agreed ways of presenting and formatting biomedical data so that it is readily reusable and machine-readable)[40].

Via their interactions with different biomedical specialities, publishers are in a good position to share best practices across disciplinary boundaries, and identify – and work with – scientists with interests in open data. Novel open access journals such as *Trials* (http://www.trialsjournal.com), which puts a special emphasis on data sharing and publishing of all clinical trial results regardless of outcome (whether positive or negative), and *BioData Mining* (http://www.biodatamining.org/), which focuses on computational aspects of knowledge discovery from large-scale genetic, transcriptomic, genomic, proteomic and metabolomic data, are products of such a strategy.

The future of scholarly communication

With increasing availability of raw data, the scientific record itself – albeit slowly – is changing. New platforms for sharing, publishing and linking data to publications are being developed, making data more integral to the scientific record – traditionally a collection of documents (journal articles). A thought-provoking essay, in the open access book on data-intensive science, *The Fourth Paradigm*, envisages instantaneous translation of new medical discoveries into clinical practice – a 'healthcare singularity' – by around 2025[41]. Gillam and colleagues envisage doctors accessing, via their smartphones, data in real time that are generated from patients' electronic health records, linked to clinical evidence databases, genomic information, drug resistance and availability data, which could further be linked to records of ongoing clinical trials[42]. All of which, combined, will inform more effective and personalized

treatment. A fantastical concept? Perhaps, but the US Department of Health and Human Services Open Government strategy to expand health data access is already calling for patient data to be available in standardized, reusable formats[43]. Moreover, platforms such as Microsoft (Health Vault) and Google (Health), and platforms for sharing such as patientslikeme.com are already enabling patients to control sharing of their health information in secure 'clouds', potentially making a data-driven scholarly record a reality sooner than might, at first glance, seem plausible.

References

1. Chernoff, M, What "open data" means – and what it doesn't: http://blog.okfn.org/2010/12/10/what-%e2%80%9copen-data%e2%80%9d-means-%e2%80%93-and-what-it-doesn%e2%80%99t/ (accessed 10 January 2011).
2. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: http://oa.mpg.de/berlin-prozess/berliner-erklarung/ (accessed 10 January 2011).
3. Comparison of BioMed Central's Article Processing Charges with those of other publishers: http://www.biomedcentral.com/info/authors/apccomparison/ (accessed 10 January 2011).
4. Open Knowledge Definition - Defining the Open in Open Data, Open Content and Open Information: http://www.opendefinition.org/ (accessed 10 January 2011).
5. Science Commons Protocol for Implementing Open Access Data http://sciencecommons.org/projects/publishing/open-access-data-protocol/ (accessed 10 January 2011).
6. Panton Principles for Open Data in Science: http://pantonprinciples.org/ (accessed 10 January 2011).
7. BioMed Central's position statement on open data: http://blogs.openaccesscentral.com/blogs/bmcblog/resource/opendatastatementdraft.pdf (accessed 10 January 2011).
8. Schofield, P N, Bubela, T, Weaver ,T, Portilla, L, Brown, S D, Hancock, J M, Einhorn, D, Tocchini-Valentini , G, Hrabe de Angelis, M and Rosenthal, N, CASIMIR Rome Meeting participants: Post-publication sharing of data and tools, *Nature*, 2009, 461(7261), 171–3.
9. Thaney, K, Sharing data on the web: http://blogs.talis.com/nodalities/2010/02/sharing-data-on-the-web.php (accessed 10 January 2011).
10. Fry, J, Lockyer, S and Oppenheim, C, Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes (November 2008): http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf (accessed 10 January 2011).
11. Wood, J, Riding the wave: How Europe can gain from the rising tide of scientific data (October 2010): http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (accessed 10 January 2011).
12. Research Information Network: To share or not to share: research data outputs http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf (accessed 10 January 2011).
13. Piwowar, H A, Day, R S and Fridsma, D B (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate, *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308
14. Sharing research data to improve public health: full joint statement by funders of health research: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm (accessed 10 January 2011).
15. Walport, M and Brest, P, Sharing research data to improve public health. *The Lancet*, Early Online Publication, 10 January 2011, doi:10.1016/S0140-6736(10)62234-9
16. Laine, C, Horton, R, DeAngelis, C D, Drazen, J M, Frizelle, F A, Godlee, F, Haug, C, Hébert, P C, Kotzin, S, Marusic, A, Sahni, P, Schroeder, T V, Sox, H C, Van der Weyden, M B and Verheugt, F W, Clinical trial registration, *BMJ*, 2007, 334 : 1177 doi: 10.1136/bmj.39233.510810.80
17. *Nature* editorial policies: availability of data and materials: http://www.nature.com/authors/editorial_policies/availability.html (accessed 10 January 2011).
18. *BMC Medicine* instructions for authors: http://www.biomedcentral.com/bmcmed/ifora/ (accessed 10 January 2011).

19. *PLoS ONE* Editorial and Publishing Policies: http://www.plosone.org/static/policies.action (accessed 10 January 2011).
20. Peng, R D, Dominici, F and Zeger, S L, Reproducible epidemiologic research, *American Journal of Epidemiology*, 2006, 163(9):783–9. Epub 1 March 2006.
21. Groves, T, BMJ policy on data sharing, *BMJ*, 2010, 340:c564 doi: 10.1136/bmj.c564
22. Piwowar, H A and Chapman, W W, A review of journal policies for sharing research data. Proceedings of the ELPUB 2008 Conference on Electronic Publishing. Toronto, ON: 25–27 June 2008: http://elpub.scix.net/data/works/att/001_elpub 2008.content.pdf (accessed 10 January 2011).
23. Wicherts, J, Borsboom, D, Kats, J and Molenaar, D, The poor availability of psychological research data for reanalysis, *American Psychologist*, 2006, 61: 726–728.
24. Savage, C J and Vickers, A J, Empirical Study of Data Sharing by Authors Publishing in PLoS Journals, *PLoS ONE*, 2009, 4(9): e7078. doi:10.1371/journal.pone.0007078
25. Whitlock, M C, McPeek, M A, Rausher, M D, Rieseberg, L and Moore, A J, Data Archiving, *The American Naturalist*, 2010, 175(2), 145–146.
26. Todd, J, Vision: Open Data and the Social Contract of Scientific Publishing, *BioScience*, 2010, 60(5), 330.
27. Emerson, G B, Warme, W J, Wolf, F M, Heckman, J D, Brand, R A and Leopold, S S, Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial, *Archives of Internal Medicine*, 2010, 170(21), 1934–9.
28. Eyding, D, Lelgemann, M, Grouven, U, Härter, M, Kromp, M, Kaiser, T, Kerekes, M F, Gerken, M and Wieseler, B, Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials, *BMJ*, 2010; 341:c4737 doi: 10.1136/bmj.c4737
29. Eysenbach, G, Tackling publication bias and selective reporting in health informatics research: register your eHealth trials in the International eHealth Studies Registry, *Journal of Medical Internet Research*, 2004, 6(3):e35.
30. Moynihan, R, Rosiglitazone, marketing, and medical science *BMJ*, 2010, 340:c1848 doi: 10.1136/bmj.c1848
31. McGauran, N, Wieseler, B, Kreis, J, Schüler Y B, Kölsch, H and Kaiser, T, Reporting bias in medical research – a narrative review, *Trials*, 2010, 11:37.
32. Vickers, A J, Whose data set is it anyway? Sharing raw data from randomized trials, *Trials*, 2006, 7:15.
33. Hrynaszkiewicz, I, Norton, M L, Vickers, A J and Altman, D G, Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers, *Trials*, 2010, 11:9.
34. Kidd , M R and Hrynaszkiewicz, I, Journal of Medical Case Reports' policy on consent for publication, *Journal of Medical Case Reports*, 2010, **4**:173.
35. Tapscott, D and Williams, A, The New Alexandrians. In: *Wikinomics*, 2007, Atlantic Books, London.
36. Contreras, J L, Prepublication Data Release, Latency, and Genome Commons, *Science*, 2010 10.1126/science.1189253
37. Sage Bionetworks: Background on the Commons: http://sagebase.org/commons/background.php (accessed 10 January 2011).
38. BioMed Central's position statement on open data, ref. 7.
39. Hrynaszkiewicz, I and Cockerill, M, In defence of supplemental data files: don't throw the baby out with the bathwater: http://blogs.openaccesscentral.com/blogs/bmcblog /entry/in_defence_of_supplemental_data (accessed 10 January 2011).
40. Hrynaszkiewicz, I, A call for *BMC Research Notes* contributions promoting best practice in data standardization, sharing and publication. *BMC Research Notes*, 2010, 3:235
41. Lynch, C, Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In: *The Fourth Paradigm: Data-intensive Scientific Discovery*, Ed. Hey, T, Tansley ,S and Tolle, K, 2009.
42. Vickers, A J and Scardino, P T, The clinically-integrated randomized trial: proposed novel method for conducting large trials at low cost, *Trials*, 2009, 10:14.
43. Conway, P H and VanLare, J M, Improving access to health care data: the Open Government strategy, *JAMA*, 2010, 304(9), 1007–8.

■ **Iain Hrynaszkiewicz**
Journal Publisher
BioMed Central
236 Gray's Inn Road
London WC1X 8HB, UK
E-mail: iain.hrynaszkiewicz@biomedcentral.com

To view the original copy of this article, published in *Serials*, click here:
http://serials.uksg.org/openurl.asp?genre=article&issn=0953-0460&volume=24&issue=1&spage=31
The DOI for this article is 10.1629/2431. Click here to access via DOI:
http://dx.doi.org/10.1629/2431
For a link to the full table of contents for the issue of *Serials* in which this article first appeared, click here:
http://serials.uksg.org/openurl.asp?genre=issue&issn=0953-0460&volume=24&issue=1