Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices
L. Chan and F. Loizides (Eds.)

doi:10.3233/978-1-61499-769-6-1

Open Science and Accelerating Discovery in Rare and Neglected Diseases

Rachel J. HARDING[1]
Structural Genomics Consortium, University of Toronto

Abstract. New medicines for many diseases, in particular neurodegenerative disorders, are not forthcoming, despite patient demands and billions of dollars spent on biomedical research globally. Traditional publishing methods in biomedical sciences are generally slow and disseminate manuscripts, sometimes without the inclusion of primary data, to a privileged audience affiliated to institutions which can afford publication subscription costs. To overcome this barrier to progressive scientific endeavors, many researchers are championing the use of preprints, transparent subject-relevant data repositories, open access journals and open lab notebooks in an effort to more effectively and efficiently communicate their research to a wider audience. In this talk I shall discuss these options and the decisions I have made as an early career researcher, to share my research output on Huntington's disease in real-time through an open lab notebook. Included will be a discussion of the motivations, methods and assessment of open online publishing, including an evaluation of my own open notebook endeavors.[2]

Keywords. Open Science, Open Data, Open Access, Open Notebook, Repository, Preprint, Huntington's Disease.

New medicines for many diseases, in particular neurodegenerative disorders, are not forthcoming, despite patient demands and billions of dollars spent on biomedical research in laboratories throughout the world. Following the complete sequencing of the human genome, many researchers hoped that they could use this information as a manual to human biology and disease, expecting that an era of rapid drug discovery would follow. In fact, studies examining productivity of pharmaceutical and biotech companies, show a steady decline since the genome's publication [1]. The reasons why the research community is struggling to develop such therapies to meet patient demand are complex and multifaceted. The way in which research output is disseminated to the wider scientific community has been identified as a key problem area by many biomedical researchers in hindering the development of novel therapies in a timely manner.

Traditional publishing methods in the biomedical sciences are fraught with numerous problems. A well-documented issue is publication bias [2], which leads to the preferential publication of "positive" data and complete research stories, not reflecting the true breadth of academic output. Recent efforts to resolve this issue have seen the rise in short communication-style manuscript journals, such as PLoS Currents, and funding agency publishing platforms, such as Wellcome Open Research, both of which encourage the rapid publication of data, protocols and findings important to

[1] Corresponding author: Rachel Harding, Structural Genomic Consortium, University of Toronto. Email: rachel.harding@utoronto.ca
[2] This paper is the text of the keynote delivered at the ELPUB2017 conference.

various scientific fields, irrespective of the scale of the study. Nonetheless, the esteem in which extensive and comprehensive research publications are held, persists in the mainstream.

The conventional publishing process itself is generally slow with often long timeframes between manuscript submission, review and acceptance [3]. This is unsatisfactory for researchers in fast-paced, innovative and competitive fields where developments can take place rapidly and authors are nervous of being scooped by competing laboratories. The long time-frame from bench to publication can slow translation of key breakthroughs and discoveries into the clinic, meaning that major field advancements may not be felt by patients within their lifetimes or the time frame of their particular condition.

Manuscripts are often accepted by journals, reporting only polished results with no commitment for the researchers to share or deposit raw data or code for the readers of their paper. Many subject areas report a reproducibility crisis and it is common for researchers to not be able to reproduce data and outcomes reported in the published literature [4]. Clear, full and honest reporting of data, methods and analysis would create a more transparent model in which mistakes could be highlighted earlier, validation of results by independent groups would be more achievable and a more cooperative ethos would exist between groups researching similar areas.

In a bid to chase increasingly shrinking research grants and positions, researchers aim to publish in journals held in high esteem by their peers, funding agencies and departmental tenure panels, which, whilst appearing "glamourous", often have substantial fees for publication, subscription and double-dip pricing structures for open access [5]. This firstly places an onus on the research groups publishing in such journals to find additional funding by which to cover these publication costs and meet their funding stipulations, which increasingly demand short embargo times or open access publication. Secondly, where publication is not open access at the first point of publication, this hinders researchers at less well funded institutes around the world, which cannot afford large numbers of high cost subscription packages, as well as the general public, from being able to access up-to-date publications.

Beyond rigorous changes to the traditional publishing system, there are alternative communication strategies which may alleviate some of these issues to provide fast, inclusive, inexpensive, transparent and open dissemination of research and data.

Preprints have been a viable self-archiving arrangement in some disciplines for more than 20 years, working synergistically with traditional publication systems and other forms of scientific communication. In particular, the physics research community has created a discipline standard by which manuscripts are routinely published on arXiv, a preprint server which has now amassed more than a million manuscripts [6]. Released on specialist platforms prior to journal submission, preprints allow fast communication of research findings in an open access manner. ArXiv has many examples of highly cited manuscripts which have not been published in the traditional system, showing that in some cases, preprinting can be sufficient means of communicating findings as well as subsequent critique and citation by peers [7]. As articles are all presented in a similar format, readers are perhaps less biased in their assessment of the preprint manuscripts with relation to journal based factors such as impact factor, and instead can focus their attentions to the academic merit of the content within. Whilst allowing fast communication of research findings to peers, preprint servers still primarily publish traditional manuscript formats and deposition of associated data is not obligatory. None the less, preprint servers are increasingly

popular in a wide-range of subject areas and are rapidly growing in number as are the number of articles which are self-archived by their authors prior to, or post-printed following, formal publication [8].

Many researchers are in fact depositing both raw and analyzed data sets, associated with manuscripts or independently accumulated, a trend which has grown during a time when the reliability and credibility of scientific findings has been drawn into question [9]. Many researchers now publish the data generated within their investigations in a bid to improve transparency as well as allowing secondary analysis by interested parties for posterity. Repositories exist for both specialist as well as broad interests and generally allow the archiving of almost any digital output from researchers. This can prove particularly lucrative for early career researchers who can digitally archive early works such as project reports, literature reviews, conference posters and lab group presentations, allowing them to generate an online presence of their scientific output from an early stage, as well as preserving their works in the process. Both preprint servers and data repositories typically permit fast sharing of research outputs through social media platforms, promoting works among peers for review and assessment with online commentary of secondary analyses and opinions of the works. Repositories also represent a critical resource for data scientists and those working with big data for data mining and meta-analyses.

However, many experimental findings still elude publication of any kind for various reasons. These include failing to complete the research project or story within the time-frame of funding, having “negative” or contradictory data to the field dogma as well as struggling to resolve difficult methodological issues. The incentive to invest time publishing these types of projects within a traditional framework is low given the poor returns in the value added to a researcher’s profile and possible negative impact on a researcher’s reputation. None-the-less, these outputs include important findings which can be useful to other researchers in the field. Other work evades publication due to the filing of patents of the findings. However, a study of patents relating to genetic data showed that patents do not encourage innovation in their specific areas [10] so perhaps should not be considered to have a positive impact on academic advancement.

I currently work as a postdoctoral fellow at the Structural Genomics Consortium (SGC), a not-for-profit public-private partnership with a focus on accelerating science in understudied areas of human biology and disease. With funding from the CHDI Foundation, the SGC has a number of research projects focused on Huntington's disease (HD), a devastating inherited neurodegenerative disease with limited therapies no available cures. Both organizations have agreed not to file for patents on anything that results from this collaboration, as well as committing to make all their data and biochemical materials resulting generated during this relationship, freely available to the broader research community.

The aim of my particular project is to understand how the underlying genetic mutation of HD, the hallmark of the disease, gives rise to the disease phenotype seen in patients using structural biology methods. Very little evidence of similar research is available in the literature despite anecdotal evidence that similar projects have been pursued in both academic and industrial labs. In an effort to accelerate research and innovate within this field, I am writing up all of my findings in close to real-time in an open notebook using the data repository Zenodo in combination with my blog, labscribbles.com. I hope that by sharing all data generated for this project freely, honestly, effectively and efficiently within the field, to generate an online international team of scientists who can critique my research, offer suggestions for future

experiments as well as collaborate on certain aspects of the project to try and reach our common research goals as quickly as possible. To date this project has resulted in numerous open collaborations being established both within the HD field and beyond, resulting in faster progression towards research milestones. I am now working to establish a portal by which I can freely share reagents generated in the course of this project in addition to the data to the scientific community.

The SGC is now considering adopting the open notebook concept for other rare disease projects. These ventures would likely have common goals; to establish an openly shared tool-box of reagents and data as a starting platform for fellow and future researchers to use to continue research in specific disease areas. Creating a range of high quality research tools should also accelerate research in fields where few materials or starting reagents are commercially available. In under-studied fields where researchers lack the risk of being scooped, an open notebook would be an excellent resource to the fields future research base, with minimal career risk to the scientist generating the data and materials in the first instance.

In the last year, the SGC launched its target enabling package (TEP) scheme. This initiative is built upon the recognition that genetic data is a good starting point for understanding certain diseases, but is insufficient alone to propel a translational project in drug discovery or even deeper understanding of the drug target or disease. As such, TEPs generated by the SGC will provide a critical mass of reagents and knowledge on a given protein target, the aim of which is to allow rapid biochemical and chemical exploration as well as characterization of proteins with genetic linkage to key disease areas. The primary goal is to accelerate drug discovery for these new targets with a fast, open access approach. All data and reagents generated as part of a TEP are shared without restriction with interested parties, even prior to formal publication.

A crucial point for the successful implementation of innovative or novel communication strategies, is that they provide positive impact to the fields adopting them. This requires comprehensive assessment of the "success" of the different methodologies compared to traditional publication methods. For different fields and different strategies, success will likely be defined differently and I believe it unlikely that all disciplines will implement all approaches well. The TEP initiative and my own open notebook project are subject to internal assessment at the SGC for their effectiveness as well as providing insight as to what future initiatives might be developed. In particular, for labscribbles, I am keen to take risks and try different approaches to find the most efficient and effective way of running this project. Whilst striving to maintain key standards with respect to time from bench to online update, the depth of detail describing each notebook installment and so forth, I hope to continue to develop and evolve the project into a success, as well as promoting this approach to others.

In embarking on my open access endeavors, I have been fortunate to have the full support of the SGC, in whose labs I am based, as well as the CHDI Foundation, who have generously funded this work. In particular, I have appreciated the mentorship of Aled Edwards, Cheryl Arrowsmith and Leticia Toledo-Sherman.

References

[1] Bernstein Research Report, The Long View - R&D productivity, 2010
[2] Nissen S.B. et al., Research: Publication bias and the canonization of false facts, *eLife*, **5** (2016), e21451

[3] Powell K., Does it take too long to publish research? *Nature*, **530(7589)** (2016), 148-51
[4] Baker M., 1,500 scientists lift the lid on reproducibility, *Nature* **533(7604)** (2016), 452-4
[5] Pinfield S. et al., The "Total Cost of Publication" in a Hybrid Open-Access Environment: Institutional Approaches to Funding Journal Article-Processing Charges in Combination With Subscriptions, *ASIS&T* **67(7)** (2016), 1751-1766
[6] Vence T., One Million Preprints and Counting: A conversation with arXiv founder Paul Ginsparg, *The Scientists* (2014)
[7] Perelman G., The entropy formula for the Ricci flow and its geometric applications, *ArXiv* (2002)
[8] http://asapbio.org/preprint-info/biology-preprints-over-time
[9] How science goes wrong, *The Economist* (2013)
[10] Sampat B. et al, How do patents affect follow-on innovation? Evidence from the human genome, *NBER* (2015) 21666