Scientific Data Science and the Case for Open Access

Gopal P. Sarma*

School of Medicine, Emory University

Abstract

"Open access" has become a central theme of journal reform in academic publishing. In this article, I examine the consequences of an important technological loophole in which publishers can claim to be adhering to the principles of open access by releasing articles in proprietary or "locked" formats that cannot be processed by automated tools, whereby even simple copy and pasting of text is disabled. These restrictions will prevent the development of an important infrastructural element of a modern research enterprise, namely, *scientific data science*, or the use of data analytic techniques to conduct meta-analyses and investigations into the scientific corpus. I give a brief history of the open access movement, discuss novel journalistic practices, and an overview of data-driven investigation of the scientific corpus. I argue that particularly in an era where the veracity of many research studies has been called into question, scientific data science should be one of the key motivations for open access publishing. The enormous benefits of unrestricted access to the research literature should prompt scholars from all disciplines to reject publishing models whereby articles are released in proprietary formats or are otherwise restricted from being processed by automated tools as part of a data science pipeline.

I. Introduction

The growth of institutional science following the Second World War has resulted in a range of unanticipated infrastructural problems, ranging from the overproduction of PhDs relative to the number of faculty positions, protracted educational trajectories for many aspiring scientists, and most alarmingly, a "reproducibility crisis," whereby the veracity of large subsets of the research literature has been called into question [1, 2, 3, 4, 5, 6, 7, 8, 9].

A significant target of institutional reform to address the larger set of issues created by dramatic scientific growth has been the academic publishing model. Access to scholarly output has traditionally been restricted to wealthy universities, whose library systems are charged exorbitant fees to maintain annual subscriptions.

In contrast, an "open access" model is one in which research articles are made freely available to all, largely taking advantage of the infrastructural efficiencies provided by the Internet. Indeed, many new journals are online only and do not distribute printed copies of their collections.

Traditionally, there have been two primary arguments for open access publishing. The first is maximizing the accessibility of research output. Subscriptions to the top journals can be prohibitively expensive and only the wealthiest universities and industrial research labs can afford them. Removing the barrier to access opens the possibility for novel results to gain significantly greater exposure and scrutiny, particularly in countries with developing scientific infrastructures that

*Email: gopal.sarma@emory.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2566v1 | CC BY 4.0 Open Access | rec: 1 Nov 2016, publ:

have to be thrifty with resource allocation. The same is true for smaller universities or research-oriented companies for whom annual subscriptions for a full spectrum of journals cannot be reasonably budgeted.

The second argument for open access publishing is eliminating the "double-billing effect" of publicly funded research. Surely tax-paying citizens should not have to pay twice to read the output of research that they have already contributed to funding. Indeed, the Public Access of Policy of the National Institutes of Health now requires that publicly funded research be made freely available via BioMed Central within 12 months of publication [10].

However, an unfortunate consequence of a focus on these two issues is an escape mechanism for publishers whereby articles can be released in proprietary or "locked" formats which prevent them from being used in the context of a data science pipeline. I will use the phrase *scientific data science* to refer to data science efforts which treat the scientific corpus itself as a massive dataset to analyze and extract important, actionable insights.

In the remainder of the article, I give a brief discussion of scientific data science, its relationship to open access publishing and the "reproducibility crisis." I close with a call to all scholars to prioritize publication in journals that provide complete, unrestricted access to research articles, and to draw attention to those publishers who are only making a token nod to open access by releasing articles in restricted, proprietary formats.

II. Scientific Data Science

I use the phrase *scientific data science* to refer to data analysis of the scientific corpus, rather than the data sets that are produced by research studies. One can think of scientific data science as representing a full-fledged generalization of review articles, systematic reviews, and meta-analyses whereby sophisticated tools from the modern data science toolkit are utilized to extract novel insights from the scientific corpus itself.

The applications of data science to the scientific corpus is in its nascent stages. It has the potential to advance our understanding of global scientific trends, the relationship between fundamental research and technological development, and fraud detection, to name just a few possible applications (see for example, [11, 12, 13, 14, 15, 16, 17, 18, 18, 19, 20])[1].

Performing such analyses requires unrestricted access to the research literature so that articles can themselves can be treated as data sets to be used as part of a data science pipeline. Therefore, publishing companies which have released articles under proprietary formats, while complying with a narrow interpretation of open access, are preventing the development of a powerful set of tools and cultural practices for advancing science.

Scientific data science is particularly important in the context of the "reproducibility crisis." Indeed, in recent years, significant attention has been drawn to low rates of reproducibility of research studies across a number of disciplines. From reproducibility initiatives, to re-examining the incentive structures of academic research, to novel journalistic practices such as post-publication peer review, the reproducibility crisis has been a significant source of controversy, discussion, and institutional action [4, 5, 7, 8, 21, 22, 23, 24, 25, 26, 27].

However, we are barely beginning to understand the scope of the problem. For example, the studies which uncovered a large number of irreproducible results were conducted in

[1] There has been slow, but steady growth of research in recent years utilizing data science to investigate trends in the origin, growth, and dissemination of knowledge—the references given above are simply a sample of contemporary work. See, for example, research at The Santa Fe Institute and The Knowledge Lab at the University of Chicago for full-fledged efforts in scientific data science.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2566v1 | CC BY 4.0 Open Access | rec: 1 Nov 2016, publ:

a limited range of subjects, and we cannot generalize from these studies to understand what the "reproducibility distribution" looks like for the entirety of science. Scientific data science has the potential to play a key role in more accurately characterizing the status of different fields by identifying "linchpin results," which would be of particularly high-value to be the focus of targeted replication studies. For instance, by examining citation networks, it may be possible to identify a small subset of candidate papers that can be directly examined by specialists in a field to determine if adequate sample sizes or appropriate statistical tests were used.

The possibilities for exploratory data analysis are endless. We might imagine using natural language processing and textual analysis to characterize the transference of ideas between fields, the emergence of new concepts, and the transition from basic to applied research. Techniques such as these will also allow us to develop more refined methodologies for characterizing the importance of individual contributions and shift away from much abused metrics such as the impact factor. For example, by using predictive models or collaborative filtering, a fully digitized corpus should allow for articles to have citations automatically generated, in addition to the manually added citations by the authors themselves. These citations could be continually updated in real time as the models are refined or new research articles are published. This would allow the notion of citations to be extended to include future results as well as those that authors were aware of at the time of publication. However, none of these tantalizing possibilities can be realized if restrictions are placed on access to the research literature in bulk form.[2] It is simply inadequate for articles to be released in proprietary formats that cannot be processed by automated tools.

Therefore, I argue that one of the primary motivations for open access publishing is the *enormous benefit to science and society that scientific data science will allow for*. Ironically, this argument is not new, and indeed, was anticipated as part of the original definition of open access given at the Budapest Open Access Initiative:

> By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited [32].

One of the key phrases in this definition is allowing users to "pass the articles as data to software," that is, *to treat the scientific corpus itself as data.* However, when these words were

[2]The move towards a pre-print model for academic publications is a positive development for scientific data science. Long practiced in the physics community via the arXiv pre-print server, and steadily being adopted by other fields, the pre-print model allows for drafts of publications to be immediately available online prior to submission to a journal. arXiv, originally launched by Los Alamos National Laboratory and now overseen by the Cornell University Library System, has made the full source code and PDFs of all its pre-prints available for free via Amazon Web Services [28]. The availability of this corpus is a powerful resource for scientific data science. However, to date, it is the only pre-print server to do so and newer repositories should also follow in its footsteps.

As the pre-print model itself becomes more widely adopted, the simultaneous availability of both pre-prints and the final journal publication will allow for critical analysis of peer review, a facet of the modern scientific process [29, 30, 31] whose re-examination is essential in addressing the reproducibility crisis. For example, simple textual analysis of pre-prints and their published counterparts will allow us to characterize the extent to which peer review influences manuscripts, and how level of influence varies across different subjects.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2566v1 | CC BY 4.0 Open Access | rec: 1 Nov 2016, publ:

first written, the phrase "data science" had yet to be coined, and the enormous growth of the field, largely driven by social media, had yet to take place. The original framers of the definition of open access had the vision and foresight to anticipate that unrestricted access to the research literature should include far more than the ability for individuals to freely read scholarly articles. They should also be able to conduct sophisticated analyses of large bodies of literature using computational techniques that have only become possible in recent years.

III. Conclusion

Realizing the full vision of scientific data science requires unrestricted access to the scientific corpus. We should aspire to build a fully open infrastructure where there are APIs for every journal and pre-print repository, allowing anyone to access the data and metadata for every article and conduct exploratory or targeted data analyses. Taking advantage of a fully digitized and easily accessible corpus of knowledge, scientific data scientists will build information dashboards providing intuitive insights into an increasingly complex knowledge base. Most importantly, scientific data science may come to play a crucial role in addressing the "reproducibility crisis" by mining large corpuses of scientific papers to uncover "linchpin results" which would subsequently be the focus of targeted replication efforts.

Scholars from all disciplines should be aware of the enormous benefit to society at large of scientific data science and reject publishing models whereby articles are released in proprietary or locked formats which prevent them from being used as part of a data analytic pipeline.

Acknowledgments

I would like to thank Travis Rich and Daniel Weissman for insightful discussions and feedback on the manuscript.

References

[1] H. Bode, F. Mosteller, J. Tukey, and C. Winsor, "The Education of a Scientific Generalist," *Science*, vol. 109, no. 2840, pp. 553–558, 1949.

[2] V. Bush, "Science: The Endless Frontier," *Science Education*, vol. 29, no. 4, pp. 218–218, 1945.

[3] V. Narayanamurti, T. Odumosu, and L. Vinsel, "RIP: The Basic/Applied Research Dichotomy," *Issues in Science and Technology*, vol. 29, no. 2, p. 31, 2013.

[4] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Med*, vol. 2, p. e124, 08 2005.

[5] W. Gunn, "Reproducibility: fraud is not the big problem," *Nature*, vol. 505, no. 7484, pp. 483–483, 2014.

[6] D. Adam and J. Knight, "Journals under pressure: Publish, and be damned...," *Nature*, vol. 419, no. 6909, pp. 772–776, 2002.

[7] E. Check and D. Cyranoski, "Korean scandal will have global fallout," *Nature*, vol. 438, no. 7071, pp. 1056–1057, 2005.

[8] R. Horton, "What's medicine's 5 sigma?," *The Lancet*, vol. 385, no. 9976, 2015.

[9] B. Alberts, M. W. Kirschner, S. Tilghman, and H. Varmus, "Rescuing US biomedical research from its systemic flaws," *Proceedings of the National Academy of Sciences*, vol. 111, no. 16, pp. 5773–5777, 2014.

[10] "NIH Open Access Policy." https://publicaccess.nih.gov/policy.htm. Accessed: 2016-10-14.

[11] D. M. Markowitz and J. T. Hancock, "Linguistic Obfuscation in Fraudulent Science," *Journal of Language and Social Psychology*, p. 0261927X15614605, 2015.

[12] Y. Ding, "Applying weighted PageRank to author citation networks," *Journal of the American Society for Information Science*

and Technology, vol. 62, no. 2, pp. 236–245, 2011.

[13] Y. Ding, "Topic-based PageRank on author cocitation networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 449–466, 2011.

[14] Y. Ding, X. Liu, C. Guo, and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *Journal of Infometrics*, vol. 7, no. 3, pp. 583–592, 2013.

[15] W. Zhu and J. Guan, "A bibliometric study of service innovation research: based on complex network analysis," *Scientometrics*, vol. 94, no. 3, pp. 1195–1216, 2013.

[16] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 408–427, 2015.

[17] M. Song, S. Kim, G. Zhang, Y. Ding, and T. Chambers, "Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central," *Journal of the Association for Information Science and Technology*, vol. 65, no. 2, pp. 352–371, 2014.

[18] S. Valverde, R. V. Solé, M. A. Bedau, and N. Packard, "Topology and evolution of technology innovation networks," *Physical Review E*, vol. 76, no. 5, p. 056118, 2007.

[19] B. Gress, "Properties of the USPTO patent citation network: 1963–2002," *World Patent Information*, vol. 32, no. 1, pp. 3–21, 2010.

[20] R. V. Solée, S. Valverde, M. R. Casals, S. A. Kauffman, D. Farmer, and N. Eldredge, "The evolutionary ecology of technological innovations," *Complexity*, vol. 18, no. 4, pp. 15–27, 2013.

[21] P. Campbell, ed., *Challenges in Irreproducible Research*, vol. 526, Nature Publishing Group, 2015.

[22] F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?," *Nature Reviews Drug Discovery*, vol. 10, no. 712, 2011.

[23] Editors, "Trouble at the lab," *The Economist*, 10 2013.

[24] Neuroskeptic, "Reproducibility Crisis: The Plot Thickens," *Discover Magazine*, 10 2015.

[25] B. Carey, "Science, Now Under Scrutiny Itself," *The New York Times*, 7 2015.

[26] K. M. Palmer, "Psychology is in a Crisis Over Whether It's in a Crisis," *Wired Magazine*, 3 2016.

[27] J. S. Flier, "How to Keep Bad Science From Getting Into Print," *The Wall Street Journal*, 3 2016.

[28] "arXiv Bulk Data Access." https://arxiv.org/help/bulk_data_s3. Accessed: 2016-10-14.

[29] D. A. Kronick, "Peer review in 18th-century scientific journalism," *Jama*, vol. 263, no. 10, pp. 1321–1322, 1990.

[30] J. C. Burnham, "The evolution of editorial peer review," *Jama*, vol. 263, no. 10, pp. 1323–1329, 1990.

[31] R. Spier, "The history of the peer-review process," *Trends in Biotechnology*, vol. 20, no. 8, pp. 357–358, 2002.

[32] "Budapest Open Access Initiative." http://www.budapestopenaccessinitiative.org. Accessed: 2016-10-14.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2566v1 | CC BY 4.0 Open Access | rec: 1 Nov 2016, publ: