COMMENTARY

A parasite’s perspective on data sharing

YoSon Park1 and Casey S. Greene1,2,*

1Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, 3400 Civic Blvd SCTR 10-131, Philadelphia, PA 19104, United States and 2Childhood Cancer Data Lab, Alex’s Lemonade Stand Foundation, Philadelphia PA 19103, United States

∗, Correspondence address. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, PA 19104, USA, Casey S. Greene, E-mail: csgreene@upenn.edu http://orcid.org/0000-0001-8713-9213

Abstract

Data generation is expensive in terms of both time and money. Sharing data enables the rapid replication, validation, and application of discoveries, increasing the pace and accuracy of research. As research parasites, or users of other people’s data, we recognize that a strong science ecosystem requires that those who share best be recognized. We find that widely accessible benchmark datasets have provided outsized benefits, and we hope that the benefits of sharing will also begin to accrue to individual investigators who share well. Funders can enhance progress by adjusting incentives to better support data sharers, which will make their programmatic investments more effective. We note some funders who are making such efforts.

Keywords: data sharing; open data; research parasite

The primary goal of scientific publishing is to disseminate findings to the community and disclose supporting observations that allow the community to infer their validity and robustness. Widely used forms of data release, such as making data available only upon reasonable requests, can delay sharing indefinitely. Though these practices reduce the scientists’ ability to critically evaluate or build upon previous work, some journals still allow these practices. Some editors have questioned whether data sharing may bring harm to certain fields, going so far as to call someone who has re-analyzed shared data without providing co-authorship to sharers “research parasites” [1]. This has given rise to “Research Parasite Awards” focused on data reuse [2] and “Research Symbiont Awards” focused on data sharing [3] that seek to recognize exemplar participants in this scientific ecosystem. As research parasites, we provide our perspective on some of the best examples of data sharing. We note how these practices provide an efficient path to accurate, impactful findings. We also discuss how funders can encourage innovation and improvements in science by preferring to support researchers who are sharing.

During the outbreak of the deadly Zika virus (ZIKV) in South and Central America in 2013 and 2014, the virus spread so rapidly that researchers studying it felt the need to share their data before completing the publication process. Acknowledging this, several researchers announced a data release, including sequence data of candidate strains, to the World Health Organization to enable other researchers to develop vaccines [4]. The Zika experimental-science team, who were studying rhesus macaques infected with Zika virus, made their data public rather than publishing them in a journal and continued to update their results daily for others to benefit from their progress [5]. These scientists committed to sharing information rapidly during public health emergencies, but there were no guidelines for formal recognition and rewards for such efforts at the time.

There are also sharing-first projects outside of infectious disease research. As part of a $62 million initiative, the UK Biobank recruited 500,000 individuals from the United Kingdom aged between 40–69 years in 2006–2010 and publicly released all generated data for bona fide research use [6]. The consortium behind this effort collected and continues to collect extensive phenotypic data including various physical measurements, accelerometry, questionnaires, biochemical assays, and genotyping data from all of its participants. The conventional model for novel, multisite data collection is to restrict access to only a
Encouraging data sharing

Figure 1: Citations for publicly shared data between 2008 and the present. For four main public data repositories, the Cancer Genome Atlas (TCGA), Genotype-Tissue Expression Project (GTEx), Encyclopedia of DNA Elements (ENCODE), and UK Biobank, we estimated the number of publications citing each dataset on PubMed [8]. For ENCODE and ENCODE_Community publications, the ENCODE official website [9] was used in place of PubMed for searches up to its finalization in 2016. For the past decade, we estimated that approximately 6,516 publications benefited from these public resources.

small set of participants who produce publications and other accomplishments linked to the resource. However, the UK Biobank is transforming the way public genetic and health data should be shared. This consortium, while providing strategically de-identified electronic health records, genotypes, and other research data to benefit public health, explicitly stated that no resource users will be expected to add the UK Biobank as co-authors on their publications or share income generated from research using these resources.

Certain data resources undergo continuous updates. In an example from ecology, researchers are actively building platforms to embrace evolving or dynamic data [7]. Yenni and colleagues describe an automated workflow where researchers can automatically integrate new data with versioning and archiving and update associated manuscripts and expand relevant supplementary tables. This framework not only allows faster integration and automated analysis of continuously updated data, it also ensures systematic review of the workflow and faster dissemination of results and provides transparency for others to validate the study and its findings. Figure 1 depicts citations for publicly shared data between 2008 and the present. For the past decade, we estimated that approximately 6,516 publications benefited from these public resources.

If data sharing plays such a key role in more rapid and accurate scientific progress, how can this behavior be better encouraged? Funders have a unique role in the scientific ecosystem. Those working at universities and other research institutions often face incentive structures that are challenging for those who seek to make change, such as promotion and tenure committees that are not accustomed to crediting data sharing. Alberts et al. [10], in an article describing flaws in the biomedical research ecosystem, note that at many institutions, a faculty salary is almost entirely covered by research grants. This increases universities’ reliance on funder dollars and provides funders with an expanded role in driving evaluation metrics throughout the ecosystem.

Funders benefit when scientists find errors earlier and make discoveries sooner, and some are strengthening their resource-sharing expectations. The Chan-Zuckerberg Initiative expects grantees to share protocols and preprints for projects that they fund. Alex’s Lemonade Stand Foundation (CCDL), and the Chan Zuckerberg Initiative Donor-Advised Fund (2018-182718), an advised fund of the Silicon Valley Community Foundation to C.S.G. C.S.G. and Y.P. are funded by the National Institutes of Health (R01 HG010067).

Note from the Editors

The 2019 Research Parasite Awards will be awarded at the 2019 Pacific Symposium on Biocomputing in early January, and GigaScience, for the third year running, is proud to support the Junior Parasite prize with a travel grant.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is funded in part by grants from the Gordon and Betty Moore Foundation (GBMF 4552), Alex’s Lemonade Stand Foundation (CCDL), and the Chan Zuckerberg Initiative Donor-Advised Fund (2018-182718), an advised fund of the Silicon Valley Community Foundation to C.S.G. C.S.G. and Y.P. are funded by the National Institutes of Health (R01 HG010067).

Author contribution

Original draft preparation: C.S.G. and Y.P.; review and editing: C.S.G. and Y.P.; investigation: Y.P.

References

1. Longo DL, Drazen JM. Data sharing. N Engl J Med 2016;374(3):276–7.
2. Greene CS, Garmire LX, Gilbert JA, et al. Celebrating parasites. Nat Genet 2017;49(4):483–4.
3. Byrd JB, Greene CS. Data-sharing models. N Engl J Med 2017;376:2305–6.
4. Zika virus strain BeH818995 polyprotein gene, complete cds. https://www.ncbi.nlm.nih.gov/nuccore/KU365777.1. Accessed 25 Sept. 2018.
5. Zika OpenResearch Portal. https://zika.labkey.com/project/OConnor/ZIKV-001/begin.view. Accessed 25 Sept. 2018.
6. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779.
7. Yenni GM, Christensen EM, Bledsoe EK, et al. Developing a modern data workflow for evolving data. bioRxiv 2018; doi:10.1101/344804.
8. PubMed.gov. https://www.ncbi.nlm.nih.gov/pubmed/, Accessed 30 Sept. 2018.
9. ENCODE: Encyclopedia of DNA Elements. https://www.encodeproject.org/. Accessed 30 Sept. 2018.
10. Alberts B, Kirschner MW, Tilghman S, et al. Rescuing US biomedical research from its systemic flaws. PNAS 2014;111(16):5773–7.