Emma Ganley

PLOS data policy

Catalyst for a better research process

The Public Library of Science (PLOS) implemented a new data policy in March 2014.[1] The policy requires that authors who publish in any PLOS journal provide an explicit statement of where the underlying data that were used to arrive at the conclusions in the manuscript can be accessed.

These data are expected to be publicly accessible and available for reuse, with only a few specific exceptions for cases when sharing is not legal, ethical, or practical. We at PLOS provide suggestions for how and where authors can deposit their data, but we are also open to new solutions. Indeed for some data formats and large quantities of data, we acknowledge that existing solutions are not yet ideal (see the related FAQ)[2].

The PLOS Data Policy will help the scientific community to better understand the different kinds of data that researchers have, and, more importantly, what resources they need to archive them. Some have commented[3] that the policy has been adopted before we have all of the solutions, but we hope it will serve as a catalyst for change and invigorate the development of new resources and infrastructure for research and access. For many researchers, this may be the spur needed to trigger more of a thought process—how can we better look after the data that we produce?

PLOS is seeking to ensure the ongoing utility of research, as making a paper openly accessible is enhanced enormously if that paper is linked seamlessly to the data from which it was constructed. In a time when post-publication peer review is more prevalent and data frequently come under intense public scrutiny, with whistle-blowers, blogs, and websites dedicated to investigating the validity and veracity of scientific publications, requiring access to the relevant data leads to a more rigorous scientific record.

Reception of the PLOS Data Policy by the scientific community was initially polarized. Many researchers welcomed the announcement, sharing PLOS's view that making data open fits with the overarching goals of open access publishing. Granted, researchers in genomics and structural biology have been sharing research data for decades.[4,5]

In recent years, an initiative from the ecology and evolutionary biology community to adopt a joint data archiving policy for their publications spurred the development of the Dryad Digital Repository.[6] However, some have voiced concerns about making their data available upon publication for fear of being "scooped" by others using "their data" before they can themselves, especially if the dataset took a long time to collect. Others

Emma Ganley is acting deputy editor of PLOS Biology, e-mail: eganley@plos.org

Contact series editors Zach Coble, digital scholarship specialist at New York University, and Adrian Ho, director of digital scholarship at the University of Kentucky Libraries, at crlnscholcomm@gmail.com with article ideas

have expressed concern that the policy will place additional burdens of time, effort, and cost on scientists, which they say would be better spent on research. There has also been confusion about which data are being requested and, depending on what sort of data they are, how to best meet the terms of the policy. PLOS is not oblivious to the disquiet, but we see some of these points as easier to address than others. Here, I will discuss the policy and also some of the points of contention that have been raised.

Genuinely difficult cases

Clinical patient data and individual genomic data: The need for data access committees

For clinical patient data or individual genomic data—research data that might reveal personal and confidential medical information about individuals—there are valid reasons why they should not be openly available. Even when anonymized, such data might still be traceable back to individuals. One example of how this might happen would be if the data include post-code information or other location-related information. Although these data are anonymized, there is always a risk in subsequently making them available.

There is also a concern about how the data will be put to use by others, and whether it will conform to the original consent. The former concern about revealing identity is valid; the latter concern about what others might do with these data is less clear, unless identification is compromised.

Similar concerns apply to other data types, including the locations of endangered species or otherwise at-risk populations and certain non-clinical but sensitive datasets (e.g., personal genomic data). We agree that if valid ethical concerns preclude making data available, there is a case for not releasing them without some form of moderation.

If scientists or clinicians submit studies based on data that cannot be made freely accessible, we understand that this is a case where meeting the PLOS Data Policy requirements presents a challenge. The policy explicitly makes an exception for such cases, and asks that researchers be explicit about where their data were sourced from, and what ethical proposals had to be met in order to access them. This way, subsequent researchers are informed of the process for obtaining access to the sensitive data.

Some institutions have in place Data Access Committees (DAC), and PLOS sees this as a potential route to successful data sharing. A DAC is a governing body where members are elected to serve a specific term by the convening organization, and are tasked with reviewing, assessing, and either granting or denying data access requests for any ethically or otherwise access-restricted data by mediating whether the proposed uses are appropriate.

We anticipate that more DACs will form in the coming years as access to data becomes more of a priority for scientists, institutions, and funders. If an author's institution does not have a DAC in place, PLOS will work with the author to find an acceptable alternative.

Examples of established institutes and projects that have DACs include the European Molecular Biology Laboratory's European Bioinformatics Institute and the International Cancer Genome Consortium. In the United States, the National Institutes of Health has created DACs to oversee genomic data sharing for genome-wide association studies. Such DACs comprise one or more senior employees selected owing to their human subject research experiences as well as their scientific and bioethical expertise.

For access to patient data in the United Kingdom, the DAC takes the form of specifically appointed National Health Service Caldicott Guardians, a system established after a report in 1997 investigated how patient data were being used.[7] Secure biorepositories that hold clinical tissue and other biological samples that can be requested for the purpose of medical research are also subject to ethical approval by a DAC that assesses all requests for access.

In short, many forms of DACs already exist, and there is no right or wrong format.

Researchers and institutes are encouraged to explore this option.

Large datasets and ill-supported data formats
It is becoming clear that our ability to generate immense quantities of data has outpaced our capacity to archive them. Or, perhaps this is not about outpacing, but rather it has not been seen as a priority issue. Now that more funders are requiring data management plans, the lack of sufficient places that provide access to and archiving of research data reveals an area where the research process is currently severely lacking. To be fair, creating a data repository—whether at the institutional or disciplinary level—requires significant investments in staffing and technology, and it often takes time for such resources to materialize. Libraries are working to rectify this problem, and hopefully the PLOS Data Policy will speed up the establishment of the necessary infrastructure for managing, preserving, and providing access to data.

Another difficult scenario consists of datasets obtained in proprietary formats that can only be easily viewed on the software with which they were captured, such as high-throughput imaging and microscopy data. There are, however, some facilities where these data can be uploaded and shared, such as ASCB's The Cell: An Image Library.[8] This repository was built around open source software produced by the Open Microscopy Environment (OME),[9] itself available to be installed as a server for image data storage. Furthermore, the OME and its partnering Bio-Formats library[10] permit conversion of many proprietary file formats to a more usable format (OME-TIFF) that enables visualization of those files and retention of otherwise impenetrable yet relevant metadata. If the image files in question are not supported by these facilities, they are willing to investigate and add new formats into the Bio-Formats library when sent some examples.

We do agree that there is not much sense in costly storage if the reality is that it would be cheaper to reproduce the data than to archive them. But this does mean that sufficient information must be provided alongside the study for reproducibility to be possible. If you have performed calculations on terabytes of image data, generating files with all of your measurements and calculations, these files should be included as supplementary information with your study.

Non-edge cases: A need for better scientific process

For those concerned about being "scooped," the short answer is that we feel they do not understand the basis of an open access approach, which is specifically designed to allow others to use published research. This might mean they do something the authors have thought of, but the much greater possibility is that they will do something more and different than if the data are kept to one lab or research group. We do understand that in some disciplines, datasets take years to generate, and that the researchers who generate them might feel strongly that they should have primacy over the resulting data. But unless they personally paid for the research, it is very hard to accept this reasoning.

On the surface, asking researchers to provide access to their data seems to be a simple request. However, having regularly made requests of authors to provide a specific piece of original data—while a manuscript is still under consideration—it is unacceptable that the requested data are already lost due to, for instance, a hard drive failure or a post-doc having left the lab.

Despite our increasing reliance on technology, we have yet to set up adequate data management practices. Thus, while some researchers argue that preserving data is an unnecessary burden with respect to time, effort, and money, I simply could not disagree more. The PLOS Data Policy aims to encourage researchers and other stakeholders to define the processes necessary to ensure optimal potential for research data.

Conclusion

For some data types we already have wonderful international databases that provide

unique accession numbers and identifiers and have the facility to provide reviewer access after the paper is published. Dryad and Figshare[11] are both excellent examples of data repository services, and I expect we will see many more endeavors emerging soon. Dryad is already seamlessly linked with PLOS Biology and PLOS Genetics.[12] This service is being extended to other journals so that the data can be uploaded simultaneously to the consideration of the research article and a digital object identifier can be linked to in both directions upon publication. Similarly, some labs subscribe to services like labarchive and iPython notebooks, and others may create their own simple file systems.

It would be good to see more institutional repositories and instances of server installations that handle specific data formats, such as the OME for microscopy imaging data. We hope that DACs will be convened where required to oversee data that require access control. But in general, we need better lab practices and a scholarly infrastructure that makes it simple for researchers to store and share their data.

To adapt a well-known quote from Theodosius Dobzhansky[13] to PLOS's stance on open data—nothing in science makes sense except in light of data.

Notes

1. T. Bloom, E. Ganley, and M. Winker, "Data Access for the Open Access Literature: PLOS's Data Policy," *PLoS Biology* 12(2): e1001797. doi:10.1371/journal.pbio.1001797.

2. Available at http://www.plosbiology.org/static/policies#faqs (accessed April 24, 2014).

3. Share Alike, Nature, 507, 140 (13 March 2014) doi:10.1038/507140a.

4. G. G. Kneale and M. J. Bishop, "Nucleic acid and protein sequence databases," *Computer Applications in the Biosciences,* 1985; 1(1):11-7.

5. F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: a computer-based archival file for macromolecular structures" (1977) *Journal of Molecular Biology,* 112, 535–542

6. Available at: http://datadryad.org/ [Accessed 24th April, 2014]

7. The Caldicott Committee (December 1997), "The Caldicott Report," Department of Health is available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_406840 (accessed April 24, 2014).

8. Available at https://www.cellimagelibrary.org/ (accessed April 10, 2014).

9. Chris Allan, Jean-Marie Burel, Josh Moore, Colin Blackburn, Melissa Linkert, Scott Loynton, Donald MacDonald, William J. Moore, Carlos Neves, Andrew Patterson, Michael Porter, Aleksandra Tarkowska, Brian Loranger, Jerome Avondo, Ingvar Lagerstedt, Luca Lianas, Simone Leo, Katherine Hands, Ron T. Hay, Ardan Patwardhan, Christoph Best, Gerard J Kleywegt, Gianluigi Zanetti and Jason R. Swedlow, "OMERO: Flexible, model-driven data management for experimental biology," *Nature Methods* 9, 245–253. doi:10.1038/nmeth.1896.

10. Melissa Linkert, Curtis T. Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, Josh Moore, Carlos Neves, Donald MacDonald, Aleksandra Tarkowska, Caitlin Sticco, Emma Hill, Mike Rossner, Kevin W. Eliceiri, and Jason R. Swedlow "Metadata matters: access to image data in the real world," *The Journal of Cell Biology,* Vol. 189 no. 5777-782 doi: 10.1083/jcb.201004104.

11. Available at http://figshare.com (accessed April, 24, 2014).

12. Blog post available is available at http://blogs.plos.org/biologue/2013/09/18/plos-genetics-partners-with-dryad/ (accessed April, 24, 2014).

13. Theodosius Dobzhansky, *The American Biology Teacher,* vol. 35, no. 3 (March 1973), 125–129, see http://biologie-lernprogramme.de/daten/programme/js/homologer/daten/lit/Dobzhansky.pdf.