The Frontier of Data Discoverability: Why We Need to Share Our Data

Author: Theresa M. Culley

Source: Applications in Plant Sciences, 5(10)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1700111

Applications in Plant Sciences

Applications in Plant Sciences 2017 5(10): 1700111

EDITORIAL

THE FRONTIER OF DATA DISCOVERABILITY: WHY WE NEED TO SHARE OUR DATA[1]

THERESA M. CULLEY[2]

Editor-in-Chief, *Applications in Plant Sciences*
Department of Biological Sciences, University of Cincinnati, 614 Rieveschl Hall, Cincinnati, Ohio 45221-0006 USA

We are now in an era where sharing and making data widely accessible are not only expected within many disciplines, but also required by federal granting agencies and many scientific journals. In addition, there are practical reasons why authors should deposit their data into permanent data repositories: (1) it prevents data loss due to accidents, theft, or death of the researcher; (2) it enables published research to be reproduced by others; (3) publications associated with accessible data sets can have higher citation rates; (4) deposited data sets are increasingly recognized for scholarly recognition and professional advancement; and (5) stored and accessible data can be used in the future for projects that are unanticipated today. *Applications in Plant Sciences* requires that data underlying its articles be publicly accessible as a condition of publication to promote the continued advancement of the field of plant biology.

Key words: data accessibility; discoverability; openness; public access.

Scientific journals are increasingly emphasizing and now often requiring the deposition of analyses, raw data, and even corresponding software code in online repositories, such as Dryad and GitHub. Why is this happening and what does it mean for you? This situation has developed from a concerted effort by governmental funding agencies over the past few years to make federally funded data transparent and accessible and thus "discoverable" by the scientific community and the general public. This shift has been necessitated by rapidly changing technologies that now create vast amounts of data. Initially led by the fields of particle physics, astronomy, and genomics, large-scale data generation is now becoming the rule rather than the exception in science (Marx, 2013; May, 2014; McNutt et al., 2016). In addition, debate on topics such as global climate change in the United States has led to popular calls across the country for access to the scientific data on which models of global warming are based, to examine reproducibility of model predictions. Beginning in 2011, the National Science Foundation (NSF) began requiring formal Data Management Plans that outline not only how federally funded data will be collected and stored, but also how such data will be made available to the larger scientific community.

This approach represents a developing expansion and shift in how we conduct science. Although traditional, hypothesis-driven empirical research based on individual data sets remains common in many scientific fields, recognition of "Big Data" research is quickly increasing with an emphasis on data archiving and sharing (e.g., Whitlock et al., 2010; May, 2014). This type of research comprises more than just data mining or fishing for patterns to be explored more deeply with subsequent empirical studies. It takes advantage of the unique combination of information contained across multiple large data sets to answer questions that otherwise could not be addressed. For example, analysis of functional genomics data sets coupled with computational modeling have provided recent advances in cell biology (Dolinski and Troyanskaya, 2015). Consequently, the way in which most research is being conducted today is rapidly changing. Disappearing are the days in which lone scientists worked diligently in the laboratory or field, carefully protecting their data from others to prevent being scooped by a competing researcher. Only on rare occasions would they have shared their data and even then, only through interpersonal exchanges with their most trusted colleagues (Wallis et al., 2013). In genomics, such an isolationist perspective began to change 30 yr ago with the creation of data repositories such as GenBank (https://www.ncbi.nlm.nih.gov/genbank/), the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena), and the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp). Data stored in these repositories have made possible several decades of subsequent research that spurred critical advancements in fields such as population biology and phylogenetics (Drew et al., 2013).

Today, more and more scientists are proactively collaborating, sharing their own data with the goal of answering new questions. To accomplish this, many key organizations have been formed to collect and share data, such as iDigBio (https://www.idigbio.org), which promotes digitization of museum specimens across the world, and the National Phenology Network

[1]Manuscript received 12 September 2017; manuscript accepted 20 September 2017.

The author thanks Beth Parada, Amy McPherson, and Richard Hund for their editorial support and unwavering encouragement while this paper was being written. Pam Diggle, Kent Holsinger, and Pam Soltis provided critical insight and recommendations into the topic of data accessibility and discoverability. This manuscript also benefited from key conversations with Amy Koshoffer, James Lee, Arlene Johnson, Ben Merritt, Robert Tunison, and Megan Philpott.

[2]Email: theresa.culley@uc.edu

doi:10.3732/apps.1700111

(https://www.usanpn.org), which houses data collected by researchers across the country. Although these collaborative efforts could simply represent maturation of the various fields or changes in funding directives, they may also reflect a societal shift in which the internet and social media promote a greater willingness for scientists, especially junior researchers, to share information with one another.

Recognizing the importance of openness and transparency as core values in science, in 2013 the White House Office of Science and Technology Policy released a memorandum to direct federal agencies to develop policies to ensure that "digital scientific data resulting from federal-funded scientific research are accessible to the public, the scientific community, and industry to the extent feasible and consistent with applicable law and policy" (National Institutes of Health, 2015). This emphasis on openness has also generated several agency initiatives, such as the National Institutes of Health's (NIH) Big Data to Knowledge (BD2K; https://datascience.nih.gov/bd2k/about) program, launched in 2012 to spur advances in biomedical research pertaining to human health by creating a digital research enterprise to share data and maximize engagement by the scientific community. For individual researchers who obtain federal funding, details as to how data will be obtained, stored, and made available on a suitable data repository must be included in the Data Management Plan for each proposed project. This is now required of 14 federal agencies, including NSF, NIH, the United States Department of Agriculture (USDA), and the National Aeronautics and Space Administration (NASA). Although governmental policies may shift somewhat under different administrations, the overall trajectory is an increasing emphasis on data availability. Consequently, there now exists a growing number of different data repositories for different scientific fields and data types (see Open Science list on Wiki at http://oad.simmons.edu/oadwiki/Data_repositories). This new and ever-evolving emphasis on data sharing is most evident in the NSF Guidelines (National Science Foundation, 2017), which state:

> *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing…. Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them and their products widely available and usable.*

To comply with these new requirements, many journals have begun adopting data policies that explicitly encourage or require authors to provide the data that underlie their publications (Rausher et al., 2010; Whitlock et al., 2010). The research publications of the Botanical Society of America have also worked together on their data policies over the past few years, gathering input from researchers and editors to explicitly require data accessibility (e.g., Diggle, 2017). Although *Applications in Plant Sciences* (*APPS*) has always required that genetic data be made publicly available on GenBank or other appropriate repositories (e.g., Dryad [http://datadryad.org]), we have been continually updating our data policy to respond to the needs of our professional community.

In August 2016, *APPS* went a step further and became a signatory of The Center for Open Science's Transparency and Openness Promotion (TOP) Guidelines, which promote data transparency, open sharing, and reproducibility across journals (Table 1; Nosek et al., 2015). By signing on to these guidelines, *APPS* has agreed to incorporate and promote specific standards within one of three levels of increasingly stringent requirements. For example, *APPS* already adheres to the Data Transparency standard because data are required to be posted to a suitable public repository as a condition of publication (Level 2); however, we do not require that analyses be reproduced independently before publication (Level 3). In terms of Code Transparency, *APPS* editorial staff check to see that submissions indicate whether any relevant source code is available and accessible to interested readers (Level 1); eventually we aim to require that such code be made available as a condition of publication (Level 2). Finally, *APPS* requires that voucher information (Culley, 2013; Rabeler, 2017) be included as a condition of publication (i.e., "Research Materials Transparency," Level 2). As of June 2017, *APPS* also recommends inclusion of a Data Accessibility statement with each submission, as outlined in the Instructions for Authors.

Aside from the fact that increasing numbers of journals and federal agencies are now requiring data be deposited in an accessible repository, why should *you* choose to make your data publicly available? There are several reasons:

Data permanency and prevention against loss—Deposition of data in a public repository is the ultimate way to ensure that your data are not lost over time, both for yourself and for other potential users. In a test of availability of supporting data from 2–22-yr-old articles, Vines et al. (2014) found that research data cannot be reliably preserved by individual researchers, who in some cases cannot even be located after a paper has been published; for example, the probability of obtaining a current e-mail address for the first, last, or corresponding author declines by 7% each year. While gathering data from more than 7500 papers to build the first phylogenetic Tree of Life, Drew et al. (2013) reported that only 16.7% of publications provided accessible data and that attempts to obtain data sets directly from authors were only 16% successful. Vines et al. (2014) recommend that authors share their data through public archives, rather than asking readers to contact them directly for the data (McNutt et al., 2016). This is also the best way to prevent loss through unfortunate circumstances, such as death of the researcher or computer theft (Berg, 2016; Roche, 2017). I have known at least one graduate student who diligently backed up her data to a separate external laboratory drive that was kept in the laboratory, only to have it stolen in a laboratory theft along with the original computer. Earlier in my career, I also witnessed the near destruction of a campus laboratory and all its data to California wildfires.

Accidents also happen within the laboratory; it is therefore imperative that data be stored off-site in a secure location, which is now easily possible with cloud computing and established repositories. In fact, many universities now strongly encourage data backup to university servers, and several are starting to offer their own digital repositories (e.g., University of Cincinnati's Scholar@UC [https://scholar.uc.edu]). Some researchers use platforms like GitHub (https://github.com/) to store their data and ongoing analyses so they can quickly pick up their project some time later; these data can be stored and shared with collaborators in private GitHub repositories, which can then be made public when the corresponding paper is published.

Reproducibility of research—One of the basic foundations of science is the ability of studies to be reproduced over time, and as such, the availability of the underlying data are paramount (Whitlock et al., 2010). In at least one recent example, an

TABLE 1. Transparency and Openness Promotion (TOP) Guidelines, arranged according to eight main standards and three levels (used with permission from https://osf.io/2cz65/ and also published in Nosek et al., 2015).

	Level 0	Level I	Level II	Level III
Citation standards	Journal encourages citation of data, code, and materials, or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used consistent with journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data transparency	Journal encourages data sharing, or says nothing.	Article states whether data are available, and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic methods (code) transparency	Journal encourages code sharing, or says nothing.	Article states whether code is available, and, if so, where to access it.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Research materials transparency	Journal encourages materials sharing, or says nothing.	Article states whether materials are available, and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and analysis transparency	Journal encourages design and analysis transparency, or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Study preregistration	Journal says nothing.	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration of study exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Analysis plan preregistration	Journal says nothing.	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration with analysis plan exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies, or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts results blind review.	Journal uses Registered Reports as a submission option for replication studies with peer review prior to observing the study outcomes.

editorial letter in *Science* acknowledged that the loss of raw data supporting a recent publication meant that readers may not be able to reproduce or extend the conclusions of that paper (Berg, 2016). In another example, an article in *Microchimica Acta* had to be retracted after concerns arose about potential image manipulation and the underlying data could not be produced because they had reportedly been lost (Koneswaran and Narayanaswamy, 2016; Retraction Watch, 2016). *Nature* also recently announced in May 2017 that authors now must make data easily available to readers upon publication (Nature, 2017), in an effort to increase transparency and reproducibility of research results.

Increased citation and reuse—There is evidence that compared to traditional publications, articles with accompanying publicly accessible data sets generate higher citation rates—an increase of around 20%, depending on the discipline (Dorch, 2012). In a more rigorous analysis of thousands of publications based on gene expression microarrays, Piwowar and Vision (2013) found that papers with accompanying data sets had 9% higher citation rates than other papers—smaller than 20% but an increase nonetheless. In addition, they reported that at least 20% of data sets deposited between 2003 and 2007 had been used by third parties. However, in order for archived data sets to be reused, they must be of high quality. In a review of 100 data sets supporting ecological and evolutionary publications, 56% were determined to be incomplete and 64% were archived in such a way that partially or entirely prevented their subsequent reuse (Roche et al., 2015).

Scholarly recognition—Proponents of data accessibility have suggested that the creation and posting of an existing data set should have similar value as the publication of a journal article in the consideration of the professional advancement of the researcher. In this case, a data set deposited to a curated repository such as Dryad can be assigned a digital object identifier (DOI), or the DOI can be requested directly through DataCite (https://www.datacite.org), and thus the data set can be listed on a curriculum vitae, just as for a peer-reviewed article. An alternative approach is a data paper, which details where a specific data set is stored (Rees, 2010) and its associated descriptive metadata (Riley, 2017). In fact, the journal *Genome Announcements* was created by the American Society of Microbiology specifically to house this type of publication, and data papers are now published in *Ecology* (see http://esapubs.org/archive/instruct_d.htm). In these cases, data can be easily accessed by interested researchers; this type of publication also creates professional motivation for researchers to make their data available. In fact, NSF now recommends that citable and accessible data sets be included in the required Biographical Sketch, and these are afforded the same importance as journal articles (National Science Foundation, 2017). Ideally, institutional administrators will also begin to confer professional credit for submitted data sets in decisions of promotion. As more granting agencies and

institutions provide credit for generating and preserving data, researchers will be more likely to share their data if they can also retain their rights to publish the results first (Tenopir et al., 2011; Wallis et al., 2013). In short, data sets are becoming a new currency in science.

Contribution to future research— Although data published today are intrinsically valuable, the usefulness of published data may increase in unanticipated ways in the future. In addition to meta-analyses (Rausher et al., 2010), there are long-term studies that have only been possible because the original data were available to later generations of researchers. Traditionally, these studies have been based on herbarium records or on historical data collected in very specific circumstances (such as multiyear surveys of the date of first bloom in a certain area of the country) that were left to colleagues. These investigations have focused on topics such as plant invasion (Morris et al., 2013) and plant-pollinator interactions (Burkle et al., 2013). My colleagues and I are currently studying forest succession in southwestern Ohio over the past 83 yr; this study has only been possible because a graduate student in the 1920s published his raw data in his dissertation and also left his field notebooks to the Department of Biological Sciences at the University of Cincinnati. In addition, we located a relevant 30-yr-old article based on the same site; however, the underlying data for this study have essentially been lost because the original researcher, while still alive today, has developed dementia and is unable to provide assistance.

Given that science is advancing so rapidly, there will surely be links between fields and across studies over time that are completely unanticipated today. However, the data must be deposited in a format consistent with open standards (e.g., XML, PNG) because we do not know what file formats will be in use in the future (Rees, 2010). Furthermore, in order to recreate the original analysis, researchers need to know how the data were originally collected and analyzed; therefore, the associated code and metadata should also be deposited (Riley, 2017). In short, deposition of data from long-term studies into a permanent repository ensures that a study can continue for generations to come or that the results can be combined with other data in novel ways to answer questions we cannot even anticipate today.

Conclusions— Understandably, some researchers may still be unwilling to share their data with others (Tenopir et al., 2011). However, research attitudes are changing. For example, researchers were initially reluctant to deposit their genetic sequences in GenBank or EMBL when first required by *Science* and *Nature* years ago, but this has now become commonplace. Despite these changing attitudes, some authors may still be reluctant to make their own data discoverable because of concerns that an unknown researcher may use the data and publish the analyses before the original investigator can publish his/her own research. Although this is rare, a solution is to embargo the deposited data set for a certain amount of time before it is publicly released. In Dryad, for example, data are automatically released when the accompanying article is published online, but authors can request a one-year embargo. In GenBank, authors can request an embargo (the length of time is at the author's discretion) to prevent data from being automatically released before an article is published.

Ultimately, data are becoming currency in science, and as such, data sets must be preserved in nonproprietary formats that will be useful in the future. However, researchers are currently grappling with important questions: At which stage should the data be preserved—the original raw data, or the processed and corrected data? Should the analyses be preserved as well? Should the associated metadata and documentation also be deposited to provide invaluable context for the data set? Who is responsible for checking to make sure that the data are truly accessible—authors, the editorial staff, reviewers, editors, or the publisher? If data include personal information, how can these data be shielded to protect individual privacy while promoting openness? How do we decide on data standards for each discipline? How do we as a research community agree on an appropriate data repository (governmental, community-maintained, commercial, etc.)? Who should pay to support ongoing curation and computing costs of data repositories? In a new era of cybersecurity and cyberterrorism, how can the integrity of the data be ensured? These and other questions need to be answered as science continues to shift toward data openness and accessibility. This is the face of the future.

LITERATURE CITED

Berg, J. 2016. Editorial expression of concern. *Science* 354: 1242.

Burkle, L. A., J. C. Marlin, and T. M. Knight. 2013. Plant-pollinator interactions over 120 years: Loss of species, co-occurrence, and function. *Science* 339: 1611–1615.

Culley, T. M. 2013. Why vouchers matter in botanical research. *Applications in Plant Sciences* 1: 1300076.

Diggle, P. K. 2017. The *American Journal of Botany* in 2017: Let's work together! *American Journal of Botany* 104: 3–4.

Dolinski, K., and O. G. Troyanskaya. 2015. Implications of Big Data for cell biology. *Molecular Biology of the Cell* 26: 2575–2578.

Dorch, B. 2012. On the citation advantage of linking to data: Astrophysics. Website https://hal-hprints.archives-ouvertes.fr/hprints-00714715v2 [accessed 11 September 2017].

Drew, B. T., R. Gazis, P. Cabezas, K. S. Swithers, J. Deng, R. Rodriguez, L. A. Katz, et al. 2013. Lost branches on the tree of life. *PLoS Biology* 11: e1001636.

Koneswaran, M., and R. Narayanaswamy. 2016. Retraction Note to: CdS/ZnS core-shell quantum dots capped with mercatpacetic acid as fluorescent probes for Hg(II) Ions. *Microchimica Acta* 183: 1519.

Marx, V. 2013. The big challenges of Big Data. *Nature* 498: 255–260.

May, M. 2014. Big biological impacts from Big Data. *Science* 344: 1298–1300.

McNutt, M., K. Lehnert, B. Hanson, B. A. Nosek, A. M. Ellison, and J. L. King. 2016. Liberating field science samples and data: Promote reproducibility by moving beyond "available upon request." *Science* 351: 1024–1026.

Morris, C., L. R. Morris, A. J. Leffler, C. D. Holifield Collings, A. D. Forman, M. A. Weltz, and S. G. Kitchen. 2013. Using long-term data sets to study exotic plant invasions on rangelands in the western United States. *Journal of Arid Environments* 95: 65–74.

National Institutes of Health. 2015. Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research. Available at: https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf [accessed 6 May 2017].

National Science Foundation. 2017. Proposal and award policies and procedures guide. NSF 17-1. Available at: https://www.nsf.gov/pubs/policydocs/pappg17_1/nsf17_1.pdf [accessed 6 May 2017].

Nature. 2017. *Scientific Reports*: Editorial and publishing policies. Website https://www.nature.com/srep/journal-policies/editorial-policies#availability [accessed 2 October 2017].

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. Promoting an open research culture: Author guidelines for journals could help promote transparency, openness, and reproducibility. *Science* 348: 1422–1425.

Piwowar, H. A., and T. J. Vision. 2013. Data reuse and the open data citation advantage. *PeerJ* 1: e175.

Rabeler, R. K. 2017. Making molecular specimen vouchers more accessible. *Taxon* 66: 537–538.

Rausher, M. D., M. A. McPeek, A. J. Moore, L. Rieseberg, and M. C. Whitlock. 2010. Data archiving. *Evolution* 64: 603–604.

Rees, J. 2010. Recommendations for independent scholarly publication of data sets. Available at: http://sciencecommons.org/wp-content/uploads/datapaperpaper.pdf [accessed 11 September 2017].

Retraction Watch. 2016. Concerns about image manipulation? Sorry, the data were lost in a flood. Website http://retractionwatch.com/2016/03/29/concerns-about-image-manipulation-sorry-the-data-were-lost-in-a-flood/#more-37805 [accessed 28 September 2017].

Riley, J. 2017. Understanding metadata: What is metadata, and what is it for? National Information Standards Organization (NISO), Baltimore, Maryland, USA. Available at: http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf [accessed 11 September 2017].

Roche, D. G. 2017. Evaluating *Science*'s open data policy. *Science* 357: 654.

Roche, D. G., L. E. B. Kruuk, R. Lanfear, and S. A. Binning. 2015. Public data archiving in ecology and evolution: How well are we doing? *PLoS Biology* 13: e1002295.

Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. 2011. Data sharing by scientists: Practices and perceptions. *PLoS ONE* 6: e21101.

Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, et al. 2014. The availability of research data declines rapidly with article age. *Current Biology* 24: 94–97.

Wallis, J. C., E. Rolando, and C. L. Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE* 8: e67332.

Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data archiving. *American Naturalist* 175: 145–146.