


# Six Simple Steps to Share Your Data When Publishing Research Articles

Patricia A. Soranno 

## Abstract

In this article, I provide six simple steps for authors to make their data available and meet the most common data publishing requirements for both open-access and non-open-access journals. These steps are simple and can easily be integrated into authors' workflows when conducting research and writing research articles. The steps include: deciding authorship of the dataset vs. authorship of the research article; exporting the data needed for all article analyses into archivable (preferably plain text) files using simple table formats; writing the metadata; depositing the data and metadata in a data repository; writing a data availability statement in the article; and citing the dataset in the methods and literature cited sections. Authors who implement these guidelines will reap many benefits for their science while also following best practices for open, robust, and reproducible science.

## Introduction

As Editor-in-Chief for an open-access journal, *Limnology & Oceanography: Letters*, part of my job is to work with authors to ensure that they are meeting minimum standards to foster open and reproducible science by placing their data in a permanent and public data repository upon publication of their research articles. To this end, I am providing this simple how-to-guide to help authors more easily prepare their data and make them available for publication in not only our journal but in any journal. Although not all journals require that authors share their data upon publication, most, if not all, open-access journals currently require it; and, I believe it will become standard practice for all publications in the near future. Thus, now is the time for all scientists, regardless of journal requirements, to learn

the simple steps to make their data available in a publicly accessible data repository for all of their publications. The intended audience for this article is students who have not published yet or scientists who have published research articles before but not their data.

Although scientists (i.e., you) can decide to only share your data for articles published in journals that require it, I recommend that you always publish your data with your research articles to not only benefit science but also because it benefits you and your career. First, this practice ensures that your data are archived for future use by other researchers who want to build upon your work. Doing so will also increase the future citations of your original article that provided the data (Piwowar and Vision 2013). Also, by archiving your data (Fig. 1), you will have a well-documented dataset for your future work. Too often, authors themselves have trouble reproducing or finding the data that underlie their past publications. By following these steps for all research articles, your career's worth of data will always be available to help advance science and be part of your lasting legacy in addition to your articles and ideas.

In this article, I provide six simple steps for authors to make their data available (Fig. 2) and to meet the most common data publishing requirements for both open-access and non-open-access journals. I briefly describe each step below and then provide additional resources for those interested in more in-depth treatment of these topics.

### Step 1: Decide authorship for the dataset vs. authorship for the research article

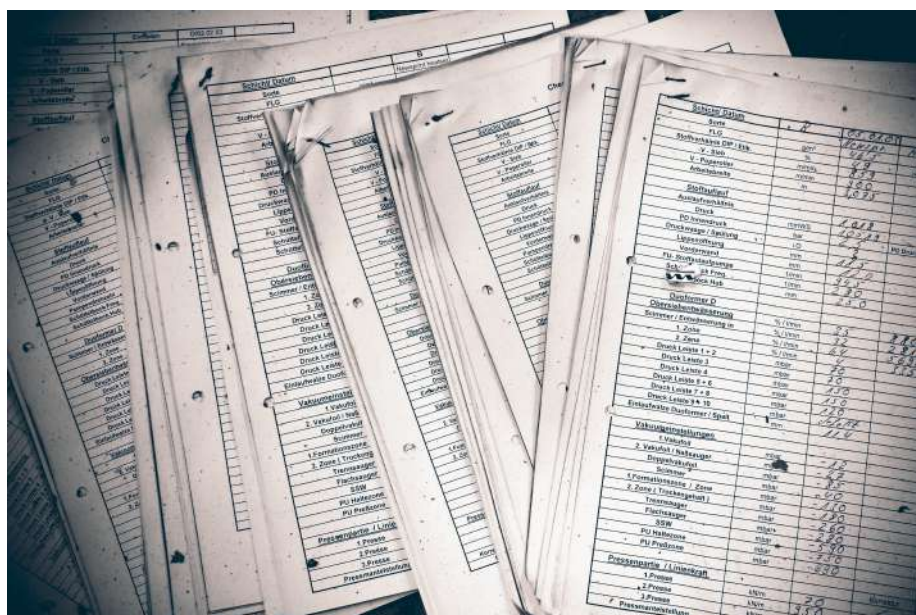
There is ongoing discussion about what constitutes authorship in scientific publications today.

One topic of debate is whether simply contributing to data collection should count toward authorship (e.g., Duke and Porter 2013). Although it is beyond the scope here to review the different arguments, I propose that one often-overlooked option in this debate is that data providers or those collecting data can be listed as an author on the dataset but not the publication when they do not make any further contributions to the manuscript in question. In other words, the dataset that gets deposited in the data repository (see step 4) can have a different list of coauthors (and even a different lead author) than the research article itself. This option is another strategy that scientists can use to ensure that all participants in science get recognized for their contributions. Either way, authorship of both the dataset and the article should happen early in the research process and be discussed throughout the manuscript-development phase, particularly after data have been collected or compiled and after analyses have been conducted. The issue of authorship for both the manuscript and the data should be discussed at those points again and contributions should feed into authorship decisions (Fig. 2).

### Step 2: Export all the data needed for the article analysis, tables, and figures in a simple and archivable format

For some research articles, all data can be provided in one file. For other articles with multiple types of data, it will be necessary to provide the data in multiple files. Either option is fine as long as you write the metadata for each file (see below) and as long as each file follows four important principles:

(a) Data are as "raw" as necessary to recreate the analysis: You should include the data



**FIG. 1.** Ensuring that research data collected on paper or in digital files can be found and reused in the future means following important steps to make the data accessible in digital online repositories for future users. Image credit: Michael Gaeda, Pixabay.

in the form that someone else could recreate your analysis presented in your article. In other words, do not only provide the data in a highly “reduced” form—such as population averages or principal components analysis (PCA) scores. Rather, you should provide the underlying data to calculate the averages or the PCA scores as well as your calculated values. It is also expected that you provide the methods or code to process the data that you report in the article (e.g., show how you calculate the values in your article from raw machine outputs). Keep in mind that every “column” in your data table will also require a description in the metadata section (see below). Finally, it is not necessary for you to include all data from the study if it was not included in the article. If there are questions about which data to include in the data repository, it is best to contact the editor.

(b) Data are in a simple table format: In general, you want your data to be easily interpreted by most if not all standard software programs. For data that are best stored in a table format (which most data are unless they are images, video, or geographic data), it is highly recommended that you follow some very simple practices. For those using a spreadsheet program like Microsoft (MS) Excel, Broman and Woo (2018) provide an extremely helpful and simple explanation for creating these types of tables, that include: be consistent, do not leave cells

empty, put one thing in a cell, organize data with subjects as rows and variables as columns and with a single header row, do not include calculations in the data file, do not use color or highlighting, choose good names, and more. I strongly recommend this article for most scientists to make sure common practices are being used, even if they are not using a spreadsheet. The fact that this valuable article was published quite recently (2018) suggests that many seasoned scientists may not be following these important guidelines. Also, for scientists using the R program and looking for a little more advanced treatment of this topic, Wickham (2014) has developed the “tidy” data framework that follows many of these basic principles and that facilitates more advanced analysis of your data.

(c) Data are in an archivable format: If your data are currently in MS Excel, it is fine. However, you must first export your data in a format that does not rely on proprietary software (such as MS Excel or Access) or a particular software version to open (like MS Excel 2013). It is easy to export your data from MS Excel into a plain text format that can be read by any software now and hopefully well into the future. Two of the more common formats are \*.csv and \*.txt files. These files also have the advantage that they are “machine readable” and can be easily processed by computer programs. There will be other types of data (such as images) that

will require different formats. For more special types of data, refer to your community standards for the format that is non-proprietary and considered archivable.

(d) Data that are downloaded from other sources are included in your dataset: It is important that authors provide an exact copy of their dataset, even if it includes data that can be downloaded from other data repositories or from government agency websites. Thus, it is not acceptable to simply state that data are available for download at another repository, partly because most researchers conduct some form of preprocessing of the data that are not always documented and because it is not always possible to ensure that other data sources will be available into the future. This practice ensures that other researchers will be able to more fully reproduce your analysis with the exact dataset used in the study. It is essential that the original data sources be documented in the metadata (see below).

### Step 3: Write the metadata

Once you get into the habit of writing metadata, it becomes easier with time and practice. At its simplest, you can download a metadata template, fill it in with the requested information, and deposit the metadata file along with the data in step 4. I recommend the template provided by the Environmental Data Initiative (EDI; [environmentaldatainitiative.org](http://environmentaldatainitiative.org)), which is also available in a modified form in the author guidelines for *Limnology & Oceanography: Letters*. These templates are easy-to-use text documents in which you can enter the required elements. The most important factor is to ensure that this metadata file is always stored in the same repository as your dataset. There are two major components of research metadata that you must be sure to document when writing metadata.

(a) Metadata for the dataset origin and context: This type of information provides details on the study itself, the keywords associated with the dataset, the authors, organizations, funders, timeframe, organisms, location of the study, and the study design and methods. This information allows other scientists to “discover” the data through searches and identify the basic information of the study itself.

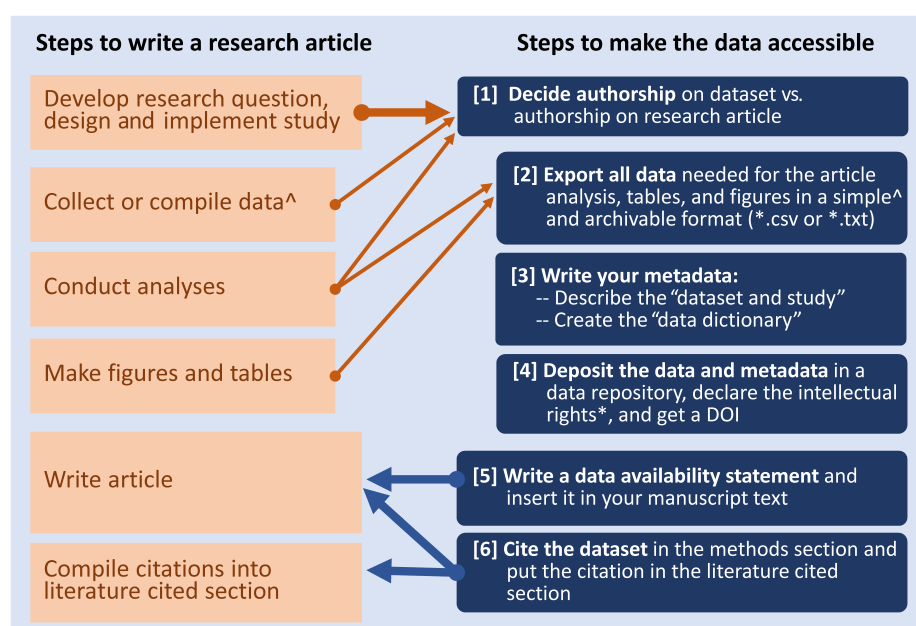
(b) Metadata for the specific variables in the dataset, called the “data dictionary”: This type of metadata provides the key information for future researchers to understand, interpret, and use the

data in your dataset. This information includes date and time formats, unique identifiers, definitions of variables, units of measurement, missing data codes, and other factors. Without such detailed information, the data will be difficult if not impossible for others to use.

However, there are two additional factors that you should consider. The first is whether there is a metadata standard in your disciplinary research area. If there is, then it is strongly recommended that you use it. The second is that it is best practice for metadata to be machine readable, which means that the metadata is written in a way that a computer can easily find it and process it through automated searches. Making the metadata machine readable means converting the information provided in the template into a file format, such as XML, that is easily readable and interpretable by computers. An example metadata format is Ecological Metadata Language, which uses XML. There are resources described on the EDI website to convert your metadata to EML format using R and Excel ([environmentaldatainitiative.org](http://environmentaldatainitiative.org)).

#### Step 4: Deposit the data and metadata in a data repository and declare the intellectual rights

One of the more important reasons that open-access journals require data to be deposited in a data repository is that repositories give your dataset a permanent identifier, most commonly a digital object identifier (DOI). A permanent identifier means your data can always be found, even if the website shuts down. Web pages and their URLs are not permanent. However, an independent organization maintains DOIs, and even if the actual location of your data changes, the DOI will always point directly to your dataset. So, how do you choose a data repository? Your first choice should be one that maintains data in your disciplinary area, which you can search for at [re3data.org](http://re3data.org). However, if there is no such repository, then select a general data repository that comes recommended by others and that meets your minimum requirements needed for your data such as dataset size limits, cost, or file format requirements. Currently, I recommend Dryad ([datadryad.org](http://datadryad.org)) and the EDI ([environmentaldatainitiative.org](http://environmentaldatainitiative.org)), which are excellent options for ecological or biological data.



<sup>^</sup> Follow the simple rules for data organization from Broman and Woo (2018): be consistent, do not leave cells empty, put one thing in a cell, organize data with subjects as rows and variables as columns and with a single header row, do not include calculations in the data file, do not use color or highlighting, choose good names.

\* Most commonly, "public domain" under Creative Commons CC0 1.0 "No Rights Reserved."

**FIG. 2.** Description of the six simple steps for authors to make their data available and meet the most common data publishing requirements for both open-access and non-open-access journals. Note that the steps to write a research article can occur in any order. However, it is strongly recommended that authors begin to think about sharing their data at the beginning of the research study rather than at the end.

#### Step 5: Write a data availability statement and insert it into your manuscript text

Data availability statements should be written in the manuscript text itself (typically immediately after the title and affiliations). The statement should be short and say something like: *Data and metadata are available in the Environmental Data Initiative repository at [insert dataset DOI].* It should say no more, and it should say no less!

#### Step 6: Cite the dataset in the methods section and put the citation in the literature cited section

In addition to the data availability statement, it is standard practice to refer to the dataset in your methods section as you would cite a research article (i.e., "The data were collected in summer 2014 (Wang et al. 2017)."). Thus, the Wang et al. (2017) citation is the citation to the data that are located in the data repository and its full citation should be in the literature cited section. This citation allows the authors to track the use of the data by tracking how often the dataset gets cited.

#### But, what about ....?

There are three additional factors that I have not addressed in the above simple guidelines. First, there will always be research projects that use some types of data that complicate these simple steps. For example, proprietary data can have restricted use; data from human subjects have their own restrictions and must be deidentified; data that are high-risk to share such as locations of endangered species; and datasets that are so large that they cannot be accessed online and may require human intervention such as mailing of the data. So, what are you to do in these cases? First, I recommend you consult experts in your immediate research area who have dealt with similar challenges. Second, contact the editor. It is the editor's job to help authors abide by the guidelines of a journal. Nevertheless, despite these potential complications, I propose that most research articles will fall well within these simple guidelines and that they can serve as a starting point for most scientists to begin to adopt these practices as part of their everyday workflow so that this practice becomes second nature.

The second issue that I have not addressed is that of sharing research scripts, code, or



software. I have not addressed this issue in this article because it is far more common for journals to require that authors share their data rather than their code. Nevertheless, I recommend that you choose one of two options to incorporate code into your best practices. The first option would be to learn and practice these steps to share your data, and then once it becomes second nature, then apply similar principles to the sharing of your code. The second alternative is to share your code alongside the data (which many data repositories already allow). This second alternative is highly recommended for promoting a more fully reproducible science.

The third issue that I have not addressed is the rich, detailed context behind each of these steps, which I have left out to provide the minimum information needed for those who are new to data sharing. However, several excellent recent articles and websites provide more detail on these topics that I highly recommend. The articles and links below are highly readable with concrete suggestions that should be especially helpful for students and new researchers. These resources also provide more advanced topics for researchers interested in taking the next step along the path to more open and reproducible science.

- For more information about open science in general, Wilkinson et al. (2016) provide an overview of the FAIR data principles that form a critical foundation for open science and data reuse. And, for those interested in advanced strategies of making their science more open, see Hampton et al. (2015).
- For recommendations about best practices for structuring your data tables in spreadsheets, see Broman and Woo (2018), although the basic strategies work for structuring data tables in any software.
- For additional information about ensuring that your data will be useable by future users, see White et al. (2013) and Goodman et al. (2014).
- For specific guidelines for sharing genomics data, see Brown et al. (2018) for an

excellent description of dealing with this specific type of datasets.

- For more advanced users who are collaborating (and sharing their data) with data analysts, see “How to share data with a statistician” by J. Leek, L. Collado-Torres, N. Reich, and N. Horton (<https://github.com/jtleek/datasharing>).
- For further details about publishing your data, see EDI’s website on the five phases of data publishing: <https://environmentaldatainitiative.org/resources/five-phase-s-of-data-publishing/phase-5/>.

## Summary

The six steps to make data available in a public repository include deciding on authorship of the dataset, exporting the data in plain text format using simple table formats, writing the metadata, depositing the data and metadata in a data repository, writing a data availability statement, citing the dataset in the methods section, and providing a citation in the literature cited section. By following these steps, authors will be ready to submit their manuscript to most journals that strongly recommend or require that they share their data at the time of submission or the resubmission stage. At the same time, authors will be following best practices for managing their data and for robust, reproducible science. I hope that in the coming years, the practices outlined in this article will become so common that this article becomes obsolete and young scientists wonder why there was ever a need for it to be written.

## Acknowledgments

Thanks to Ian McCullough for writing the title of this article and for feedback on the article; to Nicole Smith for feedback on the article and information on the article topic, metadata standards, and key resources; to Joe Stachelek for key resources and feedback on the article; and to Ellie Phillips for feedback on an earlier draft. Thanks especially to

Corinna Gries for feedback on the article as well as educating me on best practices related to data provenance, sharing, access, and documentation and for her leadership in the EDI, which is setting a high bar for data repositories and is doing much to foster open science through data access and reuse. Thanks also to EDI for the metadata template that I modified for the *L&O: Letters* author guidelines.

## References

- Broman, K. W., and K. H. Woo. 2018. Data organization in spreadsheets. *Am. Stat.* 72: 2–10. <https://doi.org/10.1080/00031305.2017.1375989>.
- Brown, A. V., J. D. Campbell, T. Assefa, D. Grant, R. T. Nelson, N. T. Weeks, and S. B. Cannon. 2018. Ten quick tips for sharing open genomic data. *PLoS Comput. Biol.* 14: e1006472. <https://doi.org/10.1371/journal.pcbi.1006472>.
- Duke, C. S., and J. H. Porter. 2013. The ethics of data sharing and reuse in biology. *Bioscience* 63: 483–489. <https://doi.org/10.1525/bio.2013.63.6.10>.
- Goodman, A., and others. 2014. Ten simple rules for the care and feeding of scientific data. *PLoS Comput. Biol.* 10: e1003542. <https://doi.org/10.1371/journal.pcbi.1003542>.
- Hampton, S. E., and others. 2015. The Tao of open science for ecology. *Ecosphere* 6: 1–13. <https://doi.org/10.1890/ES14-00402.1>
- Piwowar, H. A., and T. J. Vision. 2013. Data reuse and the open data citation advantage. *PeerJ* 1: e175. <https://doi.org/10.7717/peerj.175>.
- White, E., E. Baldridge, Z. Brym, K. Locey, D. McGlinn, and S. Supp. 2013. Nine simple ways to make it easier to (re)use your data. *Ideas Eco. Evol.* 6:1–10. doi: 10.4033/iee.2013.6b.6.f.
- Wickham, H. 2014. Tidy data. *J. Stat. Softw.* 59: 1–23. doi: 10.18637/jss.v059.i10.
- Wilkinson, M. D., and others. 2016. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>.

**Patricia A. Soranno**, Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI; [soranno@msu.edu](mailto:soranno@msu.edu)