Nine simple ways to make it easier to (re)use your data

Ethan P. White, Elita Baldridge, Zachary T. Brym, Kenneth J. Locey, Daniel J. McGlinn, and Sarah R. Supp

Ethan P. White (ethan.white@usu.edu), Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, USA, 84341

Elita Baldridge (elita.baldridge@usu.edu), Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, USA, 84341

Zachary T. Brym (zachary.brym@usu.edu), Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, USA, 84341

Kenneth J. Locey (kenneth.locey@usu.edu), Dept. of Biology, Utah State University, Logan, UT, USA, 84341

Daniel J. McGlinn (daniel.mcglinn@usu.edu), Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, USA, 84341

Sarah R. Supp (sarah.supp@usu.edu), Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, USA, 84341

Abstract

Sharing data is increasingly considered to be an important part of the scientific process. Making your data publicly available allows original results to be reproduced and new analyses to be conducted. While sharing your data is the first step in allowing reuse, it is also important that the data be easy understand and use. We describe nine simple ways to make it easy to reuse the data that you share and also make it easier to work with it yourself. Our recommendations focus on making your data understandable, easy to analyze, and readily available to the wider community of scientists.

Introduction

Sharing data is increasingly recognized as an important component of the scientific process (Whitlock et al. 2010). The sharing of scientific data is beneficial because it allows replication of research results and reuse in meta-analyses and projects not originally intended by the data collectors (Poisot, Mounce, and Gravel 2013). In ecology and evolutionary biology, sharing occurs through a combination of formal data repositories like GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and Dryad (http://datadryad.org/), and through individual and institutional websites.

While data sharing is increasingly common and straightforward, much of the shared data is not easily reused because it does not follow best practices in terms of data structure, metadata, and licensing (Jones et al. 2006). This makes it more difficult to work with existing data and therefore makes the data less useful than it could be (Jones et al. 2006; Reichman, Jones, and Schildhauer 2011). Here we provide a list of 9 simple ways to make it easier to reuse the data that you share.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

Our recommendations focus on making your data understandable, easy to work with, and available to the wider community of scientists. Following these practices makes it easier for anyone to reuse your data including other members of your lab and even yourself.

1. Share your data

The first and most important step in sharing your data is to share your data. The recommendations below will help make your data more useful, but sharing it in any form is a big step forward. So, why should you share your data?

Data sharing provides substantial benefits to the scientific community (Fienberg and Martin 1985). It allows 1) the results of existing analyses to be reproduced and improved upon (Fienberg and Martin 1985; Poisot, Mounce, and Gravel 2013), 2) data to be combined in meta-analyses to reach general conclusions (Fienberg and Martin 1985), 3) new approaches to be applied to the data and new questions asked using it (Fienberg and Martin 1985), and 4) approaches to scientific inquiry that couldn't even be considered without broad scale data sharing (Hampton et al. 2013). As a result, data sharing is increasingly required by funding agencies (Poisot, Mounce, and Gravel (2013); e.g., NSF, NIH, NSERC, FWF), journals (Whitlock et al. 2010), and potentially by law (e.g. FASTR).

Despite these potential benefits to the community, many scientists are still reluctant to share data. This reluctance is largely due to perceived fears of 1) competition for publications based on the shared data, 2) logistical barriers, and 3) a lack of recognition for sharing data (Palmer et al. 2004; Hampton et al. 2013). These concerns are often not as serious as they first appear, and the minimal costs associated with data sharing are frequently offset by individual benefits to the data sharer (Parr and Cummings 2005; Hampton et al. 2013). Many data sharing initiatives allow for data embargoes or limitations on direct competition that can last for several years while the authors develop their publications and thus avoid competition for deriving publications from the data. Also, logistical barriers to data sharing are diminishing as data archives become increasingly common and easy to use (Parr and Cummings 2005; Hampton et al. 2013). Datasets are now considered citable entities and data providers receive recognition in the form of increased citation metrics and credit on CVs and grant applications (Piwowar, Day, and Fridsma 2007; Piwowar and Vision; Poisot, Mounce, and Gravel 2013). In addition to increased citation rates, shared datasets that are documented and standardized are also more easily reused in the future by the original investigator. As a result, it is increasingly beneficial to the individual researcher to share data in the most useful manner possible.

2. Provide metadata

The first key to using data is understanding it. Metadata is information about the data including how it was collected, what the units of measurement are, and descriptions of how to best use the data. Clear metadata makes it easier to figure out if a dataset is appropriate for a project. It also makes data easier to use by both the original investigators and by other scientists by making it easy to figure out how to work with the data. Without clear metadata, datasets can be overlooked or not used due to the difficulty of understanding the data (Fraser and Gluck 1999; Zimmerman 2003), and the data becomes less useful over time (Michener et al. 1997).

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

Metadata can take several forms, including descriptive file and column names, a written description of the data, images (*i.e.,* maps, photographs), and specially structured information that can be read by computers. Good metadata should provide 1) the what, when, where, and how of data collection, 2) how to find and access the data, 3) suggestions on the suitability of the data for answering specific questions, 4) warnings about known problems or inconsistencies in the data, and 5) information to check that the data is properly imported, such as the number of rows and columns in the dataset and the total sum of numerical columns (Michener et al. 1997; Zimmerman 2003; Strasser et al. 2012).

Just like any other scientific publication, metadata should be logically organized, complete, and clear enough to enable interpretation and use of the data (Zimmerman 2007). Specific metadata standards exist (*e.g.,* Ecological Metadata Language EML, Directory Interchange Format DIF, Darwin Core DWC (Wieczorek et al. 2012), Dublin Core Metadata Initiative DCMI, Federal Geographic Data Committee FGDC (Reichman, Jones, and Schildhauer 2011; Whitlock 2011). These standards are designed to provide consistency in metadata across different datasets and also to allow computers to interpret the metadata automatically. This allows broader and more efficient use of shared data (Brunt et al. 2002; Jones et al. 2006). While following these standards is valuable, the most important thing is to have metadata at all.

You don't need to spend a lot of extra time to write good metadata. The easiest way to develop metadata is to start describing your data during the planning and data collection stages. This will help you stay organized, make it easier to work with your data after it has been collected, and make eventual publication of the data easier. If you decide to take the extra step and follow metadata standards, there are tools designed to make this easier including: KNB Morpho, USGS xtme, and FGDC workbook.

3. Provide an unprocessed form of the data

Often, the data used in scientific analyses are modified in some way from the original form in which they were collected. This is done to address the questions of interest in the best manner possible and to address common limitations associated with the raw data. However, the best way to process data depends on the question being asked and corrections for common data limitations often change as better approaches are developed. It can also be very difficult to combine data from multiple sources that have each been processed in different ways. Therefore, to make your data as useful as possible it is best to share the data in as raw a form as possible.

This is not to say that your data are best suited for analysis in the raw form, but providing it in the raw form gives data users the most flexibility. Of course, your work to develop and process the data is also very important and can be quite valuable for other scientists using your data. This is particularly true when correcting data for common limitations. Providing both the raw and processed forms of the data, and clearly explaining the differences between them in the metadata, is an easy way to include the benefits of both data forms. An alternate approach is to share the unprocessed data along with the code that process the data to the form you used for analysis. This allows other scientists to assess and potentially modify the process by which you arrived at the values used in your analysis.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

4. Use standard data formats

Everyone has their own favorite tools for storing and analyzing data. To make it easy to use your data it is best to store it in a standard format that can be used by many different kinds of software. Good standard formats include the type of file, the overall structure of the data, and the specific contents of the file.

Use standard file formats

You should use file formats that are readable by most software and, when possible, are non-proprietary (Borer et al. 2009; Strasser et al. 2011; Strasser et al. 2012). Certain kinds of data in ecology and evolution have well established standard formats such as FASTA files for nucleotide or peptide sequences (http://zhanglab.ccmb.med.umich.edu/FASTA/) and the Newick files for phylogenetic trees (http://evolution.genetics.washington.edu/phylip/newicktree.html). Use these well defined formats when they exist, because that is what other scientists and most existing software will be able to work with most easily.

Data that does not have a well defined standard format is often stored in tables. Tabular data should typically be stored as text files because they can be opened by any type of software. These text files use delimiters to indicate different columns, and commas are the most commonly used delimiter (i.e., comma-delimited text files with the .csv extension). In contrast to plain text files, proprietary formats such as those used by Microsoft Excel can be difficult to load into other programs. In addition, these types of files can become obsolete, eventually making it difficult to open the data files at all if the newer versions of the software no longer support the original format (Borer et al. 2009; Strasser et al. 2011; Strasser et al. 2012).

When naming files you should use descriptive names so that it is easy to keep track of what data they contain (Borer et al. 2009; Strasser et al. 2011; Strasser et al. 2012). If there are multiple files in a dataset, name them in a consistent manner to make it easier to automate working with them. You should also avoid spaces in file names, which can cause problems for some software (Borer et al. 2009). Spaces in file names can be avoided by using camel case (e.g, RainAvg) or by separating the words with underscores (e.g., rain_avg).

Use standard table formats

Data tables are ubiquitous in ecology and evolution. Tabular data provides a great deal of flexibility in how to structure the data, which makes it easy to structure the data in a way that is difficult to (re)use. We provide three simple recommendations to help ensure that tabular data is properly structured to allow the data to be easily imported and analyzed by most data management systems and common analysis software, such as R and Python.

- Each row should represent a single observation (i.e., a record) and each column should represent a single variable or type of measurement (i.e., a field) (Borer et al. 2009; Strasser et al. 2011; Strasser et al. 2012). This is the standard format for tables in the most commonly

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

used database management systems and analysis packages and makes the data easy to work with in the most general way.

- Every cell should contain only a single value (Strasser et al. 2012). For example, do not include units in the cell with the values (Figure 1) or include multiple measurements in a single cell. Violating this rule makes it difficult to process or analyze your data using standard tools, because there is no easy way for the software to treat the items within a cell as separate pieces of information.

- There should only be one column for each type of information (Borer et al. 2009; Strasser et al. 2011; Strasser et al. 2012). The most common violation of this rule is cross-tab structured data (http://en.wikipedia.org/wiki/Cross_tabulation), where different columns contain measurements of the same variable (e.g., in different sites, treatments, etc.; Figure 1).

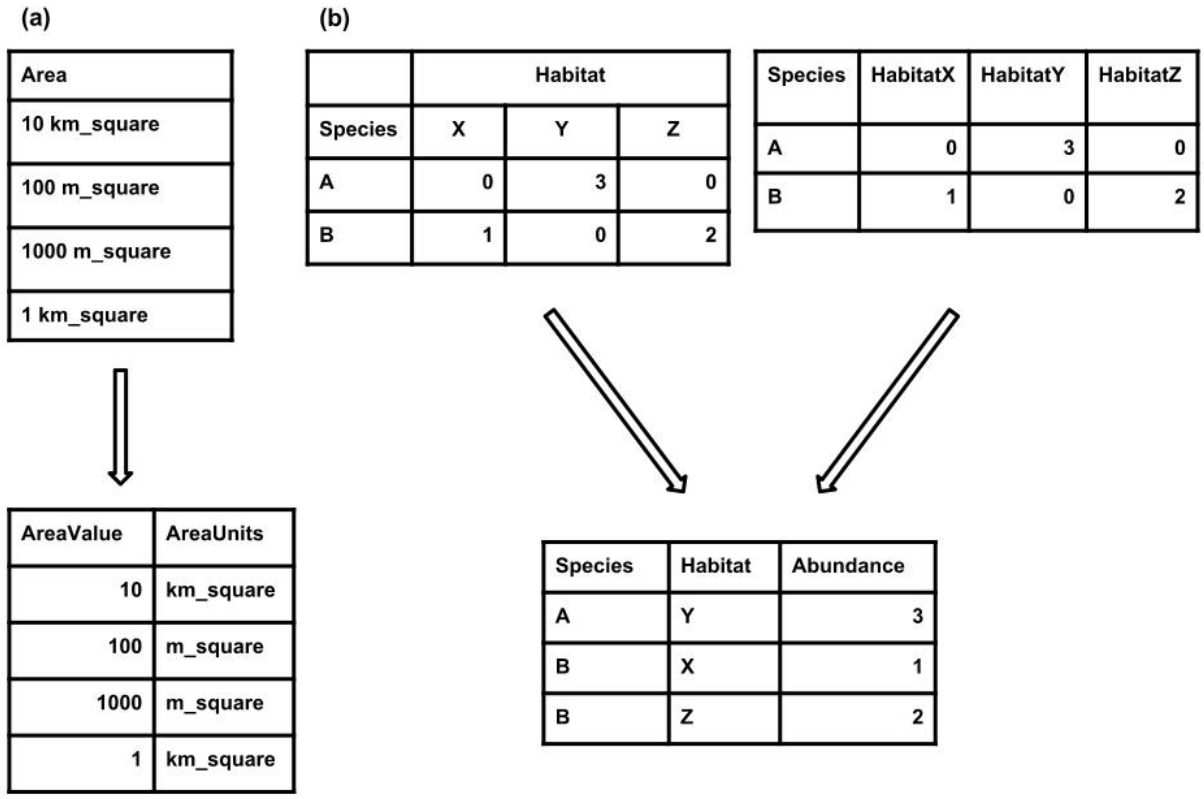

Figure 1: Examples of how to restructure two common issues with tabular data. (a) Each cell should only contain a single value. If more than one value is present then the data should be split into multiple columns. (b) There should be only one column for each type of information. If there are multiple columns then the column header should be stored in one column and the values from each column should be stored in a single column.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

While cross-tab data can be useful for its readability, and may be appropriate for data collection, this format makes it difficult to link the records with additional data (e.g., the location and environmental conditions at a site) and it cannot be properly used by most common database management and analysis tools (e.g., relational databases, dataframes in R and Python, etc.). If tabular data are currently in a cross-tab structure, there are tools to help restructure the data including functions in Excel, R (e.g., melt() function in the R package reshape; Wickham (2007)), and Python (e.g., melt() function in the Pandas Python module http://pandas.pydata.org/).

In addition to following these basic rules you should also make sure to use descriptive column names (Borer et al. 2009). Descriptive column names make the data easier to understand and therefore make data interpretation errors less likely. As with file names, spaces can cause problems for some software and should be avoided.

Use standard formats within cells

In addition to using standard table structures it is also important to ensure that the contents of each cell don't cause problems for data management and analysis software. Specifically, we recommend:

- Be consistent. For example, be consistent in your capitalization of words, choice of delimiters, and naming conventions for variables.
- Avoid special characters. Most software for storing and analyzing data works best on plain text, and accents and other special characters can make it difficult to import your data (Borer et al. 2009; Strasser et al. 2012).
- Avoid using your delimiter in the data itself (e.g., commas in the notes filed of a comma-delimited file). This can make it difficult to import your data properly.
- When working with dates use the YYYY-MM-DD format (i.e., follow the ISO 8601 data standard).

5. Use good null values

Most ecological and evolutionary datasets contain missing or empty data values. Working with this kind of "null" data can be difficult, especially when the null values are indicated in problematic ways. Unfortunately, there are many different ways to indicate a missing/empty value, and very little agreement on which approach to use.

We recommend choosing a null value that is both compatible with most software and unlikely to cause errors in analyses (Table 1). The null value that is most compatible with the software commonly used by biologists is the blank (i.e., nothing; Table 1). Blanks are automatically treated as null values by R, Python, SQL, and Excel. They are also easily spotted in a visual examination of the data. Note that a blank involves entering nothing, it is not a space, so if you use this option make sure there aren't any hidden spaces. One potential issue with blanks is that it can be difficult to know if a value is missing or was overlooked during data entry.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

NA and NULL are reasonable null values, but they are only handled automatically by a subset of commonly used software (Table 1). NA can also be problematic if it is also used as an abbreviation (e.g., North America, Namibia, *Neotoma albigula*, sodium, etc.). We recommend against using numerical values to indicate nulls (e.g., 999, -999, etc.) because they typically require an extra step to remove from analyses and can be accidentally included in calculations. We also recommend against using non-standard text indications (e.g., No data, ND, missing, —) because they can cause issues with software that requires consistent data types within columns). Whichever null value that you use, only use one, use it consistently throughout the data set, and indicate it clearly in the metadata.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
999, -999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Can cause problems with data type	Python	Avoid
No data	Can cause problems with data type, contains a space		Avoid
Missing	Can cause problems with data type		Avoid
-,+,.	Can cause problems with data type		Avoid

Table 1: Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as being a null value for specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented as part of the table.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

6. Make it easy to combine your data with other datasets

Ecological and evolutionary data is often most valuable when combined with other kinds of data (e.g., taxonomic, environmental). You can make it easier to combine your data with other data sources by including the data that is common across many data sources (e.g., Latin binomials, latitudes and longitudes) It is common for data to include codes or abbreviations. For example, in ecology and evolution codes often appear in place of site locations or taxonomy. This is useful because it reduces data entry (e.g., DS instead of *Dipodomys spectabilis*) and redundancy (a single column for a species ID rather than separate columns for family, genus, and species). However, without clear definitions these codes can be difficult to understand and make it more difficult to connect your data with external sources. The easiest way to link your data to other datasets is to include additional tables that contain a column for the code and additional columns that describe the item in the standard way. For example, you might include a table with the species codes followed by their most current family, genus, and specific epithet. For site location, you could include a table with the site code followed by latitude and longitude. Linked tables can also be used to include additional information about your data, such as spatial extent, temporal duration, and other appropriate details.

7. Perform basic quality control

Data, just like any other scientific product, should undergo some level of quality control (Reichman, Jones, and Schildhauer 2011). This is true regardless of whether you plan to share the data because quality control will make it easier to analyze your own data and decrease the chance of making mistakes. However, it is particularly important for data that will be shared because scientists using the data won't be familiar with quirks in the data and how to work around them.

At its most basic, quality control can consist of a few quick sanity checks of the data. More advanced quality control can include automated checks on data as it is entered and double-entry of data (Lampe and Weiler 1998; Paulsen, Overgaard, and Lauritsen 2012). This additional effort can be time consuming, but is valuable because it increases data accuracy by catching typographical errors, reader/recorder error, out-of-range values, and questionable data in general (Lampe and Weiler 1998; Paulsen, Overgaard, and Lauritsen 2012).

Before sharing your data we recommend performing a quick "data review". Start by performing some basic sanity checks on your data. For example:

- If a column should contain numeric values, check that there are no non-numeric values in the data.
- Check that empty cells actually represent missing data, and not mistakes in data entry, and indicate that they are empty using the appropriate null values (see recommendation 6).
- Scan your data for consistency in unit of measurement, data type (e.g., numeric, character), naming scheme (e.g., taxonomy, location), etc.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

Then ask someone else to look over your metadata and data and provide you with feedback about anything they didn't understand. In the same way that friendly reviews of papers can help catch mistakes and identify confusing sections of papers, a friendly review of data can help identify problems and things that are unclear in the data and metadata.

8. Use an established repository

For data sharing to be effective, data should be easy to find, accessible, and stored where it will be preserved for a long time (Kowalczyk and Shankar 2011). To make your data visible and easily accessible, and to ensure a permanent link to a well maintained website, we suggest depositing your data in one of the major well-established repositories. This guarantees that the data will be available in the same location for a long time, in contrast to personal and institutional websites that do not guarantee the long-term persistence of the data. There are repositories available for sharing almost any type of biological or environmental data. Repositories that host specific data types, such as molecular sequences (e.g., DDBJ, GenBank, MG-RAST), are often highly standardized in data type, format, and quality control approaches. Other repositories host a wide array of data types and are less standardized (e.g., Dryad, KNB, PANGAEA). In addition to the repositories focused on the natural sciences there are also all purpose repositories where data of any kind can be shared (e.g., Figshare).

When choosing a repository you should consider where other researchers in your discipline are sharing their data. This helps you quickly identify the community's standard approach to sharing and increases the likelihood that other scientists will discover your data. In particular, if there is a centralized repository for a specific kind of data (e.g., GenBank for sequence data) then you should use that repository.

In cases where there is no *de facto* standard it is worth considering differences among repositories in terms of use, data rights, and licensing (Table 3) and whether your funding agency or journal has explicit requirements or restrictions related to repositories. We also recommend that you use a repository that allows your dataset to be easily cited. Most repositories will describe how this works, but an easy way to guarantee that your data is citable is to confirm that the repository associates it with a digital object identifier (DOI). DOIs are permanent unique identifiers that are independent of physical location and site ownership.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

Repository	License	DOI	Metadata	Access	Notes
Dryad	CC0	Yes	Suggested	Open	Ecology & evolution data associated with publications
Ecological Archives	No	Yes	Required	Open	Publishes supplemental data for ESA journals and stand alone data papers
Knowledge Network for Biocomplexity	No	Yes	Required	Variable	Partners with ESA, NCEAS, DataONE
Paleobiology Database	Various CC	No	Optional	Variable	Paleontology specific
Data Basin	Various CC	No	Optional	Open	GIS data in ESRI files, limited free space
Pangaea	Various CC	Yes	Required	Variable	Editors participate in QA/QC
Figshare	CC0	Yes	Optional	Open	Also allows deposition of other research outputs and private datasets

Table 2: Popular repositories for scientific datasets. This table does not include well-known molecular repositories (e.g. GenBank, EMBL, MG-RAST) that have become *de facto* standards in molecular and evolutionary biology. Consequently, several of these primarily serve the ecological community. These repositories are not exclusively used by members of specific institutions or museums, but accept data from the general scientific community.

9. Use an established and liberal license

Including an explicit license with your data is the best way to let others know exactly what they can and cannot do with the data you shared. We recommend using well established licenses (e.g., Creative Commons licenses) in order to clearly communicate the rights and responsibilities of both the people providing the data and the people using it. We also recommend using the most open license possible, because even minor restrictions on data use can have unintended consequences for the reuse of the data (Schofield et al. 2009; Poisot, Mounce, and Gravel 2013). The Creative

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

Commons Zero license (CC0) places no restrictions on data use and is considered by many to be the best license for sharing data (e.g., (Schofield et al. 2009; Poisot, Mounce, and Gravel 2013), http://blog.datadryad.org/2011/10/05/why-does-dryad-use-cc0/). Having a clear and open license will increase the chance that other scientists will be comfortable using your data.

Concluding remarks

Data sharing has the potential to transform the way we conduct ecological and evolutionary research (Fienberg and Martin 1985; Whitlock et al. 2010; Poisot, Mounce, and Gravel 2013). As a result, there are an increasing number of initiatives at the federal, funding agency, and journal levels to encourage or require the sharing of the data associated with scientific research (Piwowar and Chapman 2008; Whitlock et al. 2010; Poisot, Mounce, and Gravel 2013). However, making the data available is only the first step. To make data sharing as useful as possible it is necessary to make the data usable with as little effort as possible (Jones et al. 2006; Reichman, Jones, and Schildhauer 2011). This allows scientists to spend their time doing science rather than cleaning up data.

We have provided a list of 9 practices that require only a small additional time investment but substantially improve the usability of data. These practices can be broken down into three major groups.

1. Well documented data is easier to understand.
2. Properly formatted data is easier to use in a variety of software.
3. Data that is shared in established repositories with open licenses is easier for others to find and use.

Most of these recommendations are simply good practice for working with data regardless of whether that data is shared or not. This means that following these recommendations (2-7) make the data easier to work with for anyone, including you. This is particularly true when returning to your own data for further analysis months or years after you originally collected or analyzed it. In addition, data sharing often occurs within a lab or research group. Good data sharing practices make these in-house collaborations faster, easier, and less dependent on lab members who may have graduated or moved on to other things.

By following these practices we can assure that the data collected in ecology and evolution can be used to its full potential to improve our understanding of biological systems.

Acknowledgments

Thanks to Karthik Ram for organizing this special section and inviting us to contribute. Carly Strasser and Kara Woo recommended important references and David Harris and Carly Strasser provided valuable feedback on null values, all via Twitter. The writing of this paper was supported by a CAREER grant from the U.S. National Science Foundation (DEB 0953694) to EPW.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7

References

Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. 2009. "Some simple guidelines for effective data management." *Bulletin of the Ecological Society of America* 90: 205–214.

Brunt, James W., Peter McCartney, Karen Baker, and Susan G. Stafford. 2002. "The future of ecoinformatics in long term ecological research." In *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics: SCI*, 14–18.

Fienberg, Stephen E., and Margaret E. Martin. 1985. *Sharing research data*. Natl Academy Pr.

Fraser, Bruce, and Myke Gluck. 1999. "Usability of Geospatial Metadata or Space-Time Matters." *Bulletin of the American Society for Information Science and Technology* 25: 24–28.

Hampton, Stephanie E., Carly A. Strasser, Joshua J. Tewksbury, Wendy K. Gram, Amber E. Budden, Archer L. Batcheller, Clifford S. Duke, and John H. Porter. 2013. "Big data and the future of ecology." *Frontiers in Ecology and the Environment*. doi:10.1890/120103. http://www.esajournals.org.dist.lib.usu.edu/doi/abs/10.1890/120103.

Jones, Matthew B., Mark P. Schildhauer, O. J. Reichman, and Shawn Bowers. 2006. "The new bioinformatics: integrating ecological data from the gene to the biosphere." *Annual Review of Ecology, Evolution, and Systematics*: 519–544.

Kowalczyk, Stacy, and Kalpana Shankar. 2011. "Data sharing in the sciences." *Annual Review of Information Science and Technology* 45: 247–294. doi:10.1002/aris.2011.1440450113. http://dx.doi.org/10.1002/aris.2011.1440450113.

Lampe, A. J., and J. M. Weiler. 1998. "Data capture from the sponsors' and investigators' perspectives: Balancing quality, speed, and cost." *Drug information journal* 32: 871–886.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. 1997. "Nongeospatial metadata for the ecological sciences." *Ecological Applications* 7: 330–342.

Palmer, Margaret A., Emily S. Bernhardt, Elizabeth A. Chornesky, Scott L. Collins, Andrew P. Dobson, Clifford S. Duke, Barry D. Gold, et al. 2004. "Ecological Science and Sustainability for a Crowded Planet." http://www.esa.org/ecovisions/ppfiles/EcologicalVisionsReport.pdf.

Parr, Cynthia S., and Michael P. Cummings. 2005. "Data sharing in ecology and evolution." *Trends in Ecology and Evolution* 20: 362–362.

Paulsen, Aksel, Søren Overgaard, and Jens Martin Lauritsen. 2012. "Quality of Data Entry Using Single Entry, Double Entry and Automated Forms Processing–An Example Based on a Study of Patient-Reported Outcomes." *PloS one* 7: 35087.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. 2007. "Sharing detailed research data is associated with increased citation rate." *PLoS One* 2: 308.

Piwowar, Heather A., and Wendy W. Chapman. 2008. "A review of journal policies for sharing research data." In *ELPUB2008*.

Piwowar, Heather A., and Todd J. Vision. "Data reuse and the open data citation advantage." *PeerJ PrePrints* 1: 1. http://dx.doi.org/10.7287/peerj.preprints.1.

Poisot, Timothée, Ross Mounce, and Dominique Gravel. 2013. "Moving toward a sustainable ecological science: don't let data go to waste!." https://github.com/tpoisot/DataSharingPaper/blob/master/DataSharing-MS.md.

Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. "Challenges and opportunities of open data in ecology." *Science(Washington)* 331: 703–705.

Schofield, Paul N., Tania Bubela, Thomas Weaver, Lili Portilla, Stephen D. Brown, John M. Hancock, David Einhorn, Glauco Tocchini-Valentini, Martin Hrabe de Angelis, and Nadia Rosenthal. 2009. "Post-publication sharing of data and tools." *Nature* 461: 171–173.

Strasser, Carly A., Robert B. Cook, William K. Michener, Amber Budden, and Rebecca Koskela. 2011. "Promoting Data Stewardship Through Best Practices." In *Proceedings of the Environmental Information Management Conference 2011 (EIM 2011)*.

Strasser, Carly A., Robert Cook, William K. Michener, and Amber Budden. 2012. "Primer on Data Management: What you always wanted to know." doi:10.5060/D2251G48. http://dx.doi.org/doi:10.5060/D2251G48.

Whitlock, Michael C. 2011. "Data archiving in ecology and evolution: best practices." *Trends in ecology & evolution* 26: 61–65.

Whitlock, Michael C., Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore. 2010. "Data archiving." *The American Naturalist* 175: 145–146.

Wickham, Hadley. 2007. "Reshaping data with the reshape package." *Journal of Statistical Software* 21. http://www.jstatsoft.org/v21/i12/paper.

Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An evolving community-developed biodiversity data standard." *PloS one* 7: 29715.

Zimmerman, Ann S. 2003. "Data sharing and secondary use of scientific data: Experiences of ecologists."

———. 2007. "Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse." *International Journal on Digital Libraries* 7: 5–16.

PeerJ PrePrints | https://peerj.com/preprints/7 | v1 received: 11 Apr 2013, published: 11 Apr 2013, doi: 10.7287/peerj.preprints.7