A Guide and Toolbox to Replicability and Open Science in Entomology

Jacob T. Wittman1,2,* and Brian H. Aukema1,*

1Department of Entomology, University of Minnesota, 1980 Folwell Avenue, St. Paul, MN 55102 and 2Corresponding author, e-mail: wittm094@umn.edu

Subject Editor: Sean ODonnell

Received 10 January 2020; Editorial decision 15 April 2020

Abstract

The ability to replicate scientific experiments is a cornerstone of the scientific method. Sharing ideas, workflows, data, and protocols facilitates testing the generalizability of results, increases the speed that science progresses, and enhances quality control of published work. Fields of science such as medicine, the social sciences, and the physical sciences have embraced practices designed to increase replicability. Granting agencies, for example, may require data management plans and journals may require data and code availability statements along with the deposition of data and code in publicly available repositories. While many tools commonly used in replicable workflows such as distributed version control systems (e.g., ‘git’) or script programming languages for data cleaning and analysis may have a steep learning curve, their adoption can increase individual efficiency and facilitate collaborations both within entomology and across disciplines. The open science movement is developing within the discipline of entomology, but practitioners of these concepts or those desiring to work more collaboratively across disciplines may be unsure where or how to embrace these initiatives. This article is meant to introduce some of the tools entomologists can incorporate into their workflows to increase the replicability and openness of their work. We describe these tools and others, recommend additional resources for learning more about these tools, and discuss the benefits to both individuals and the scientific community and potential drawbacks associated with implementing a replicable workflow.

Key words: reproducibility, open access, data curation, data management, preprint servers

Fundamental to scientific inference is the axiom that scientific findings are replicable; that is, one study conducted under identical conditions multiple times will yield consistent results. According to recent high-profile studies, a disconcerting percentage of the results from studies in the fields of medicine and social science may not be replicable, casting doubt on their veracity (Open Science Collaboration 2015, Fraser et al. 2018, Ioannidis 2018). Incentives to try to replicate results are limited, however, making it difficult to evaluate the reliability of the findings within a given field (Higginson and Munafò 2016, Lazarević and Žeželj 2018). A putative ‘reproducibility crisis’ (Ioannidis 2018) provides an opportunity to reflect on how we can improve implementation of the scientific method within the discipline of entomology, such as through the practice of ‘open science’. A European Union initiative to ‘foster the implementation of open science’, FOSTER Plus, defines open science as ‘the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods’ (FOSTER 2019). As such, open science and replicable science are tightly integrated.

The terms ‘replicability’ and ‘reproducibility’ assume different definitions in some fields but are used interchangeably in others (Barba 2018). Recent efforts to distinguish between four dimensions of reproducible research offer some helpful guidance for definitions of these terms and others (Plesser 2018, The Turing Way Community et al. 2019). For example, if a researcher can perform the same analysis on the same data and achieve the same results, that result is said to be ‘reproducible’. If the same analysis is performed on a different set of data, that result is said to be ‘replicable’. Research is deemed ‘robust’ if a different analysis can be performed with the same data and qualitatively similar results are obtained. Robust research provides proof that the results are unlikely to be due to idiosyncrasies of the analysis method. Finally, a result is ‘generalizable’ if the result is qualitatively similar to the original when a different data set and a different analysis method are used. Building toward generalizable results is an important step in building knowledge. Scientific results are the strongest when they can be generalized. To be generalized, research methods and data must be communicated and available to the fullest extent possible.

Practicing open science can benefit individual scientists, the scientific community, and society. The common adage that ‘your most important collaborator is you six months from now, and past you will not answer emails’ is easily addressed when adopting reproducible practices, as reproducible science requires thorough documentation of the processes undertaken during an experiment and analysis.
Making data and code publicly available can facilitate working with current collaborators, while also making it possible for more serendipitous collaborations to arise through repository search engines. In an open science framework, the products of research move beyond the final publication by intentionally encompassing data and code generated during the process. Availability increases the return on investment by funding agencies and increases the value of any taxpayer-supported work (Molloy 2011, Piwowar et al. 2011). Publically available code and data can also increase citation rates (Piwowar et al. 2007, Uhlir and Schröder 2007, Piwowar and Vision 2013, Drachen et al. 2016, Mislan et al. 2016). The open science movement is likely to grow as disparate fields adopt these practices.

Openness and generalizability requires an emphasis on transparency, proper documentation, and training in experimental and statistical methods. A methods section of a paper that is not completely transparent may fail to contain enough information to permit true replication of the experiment. Researchers make many small decisions over the course of an experiment, known as ‘researcher degrees of freedom’, that impact the results of the experiment or how they are interpreted (Gelman and Loken 2013, Ioannidis et al. 2014, Nakagawa and Parker 2015, Wicherts et al. 2016). Ad hoc decisions made during data analysis, for example, such as how to handle outliers or deciding how to deal with violation of statistical assumptions, can lead to false positives (Ihle et al. 2017, Munafò et al. 2017, Fraser et al. 2018).

The Center for Open Science, a nonprofit organization dedicated to increasing openness, integrity, and reproducibility of research, worked with journals, funding agencies, and scientific societies to develop Transparency and Openness Promotion (TOP) guidelines (Nosek et al. 2015). The TOP guidelines provide modular standards that journals can adopt to encourage or require different aspects of open science, including transparency standards related to data citation, data availability, analytic methods, research materials, design and analysis, study preregistration, analysis plan preregistration, and replication. Over 5,000 journals have adopted different levels of these standards as part of their publishing requirements, such as preregistering experiments before executing them, devising a detailed data management plan, making data and analysis code publicly available in an online repository, and submitting manuscripts to preprint servers at the time of submission to a journal (Hampton et al. 2015, Michener 2015, Vale 2015, Munafò et al. 2017, Elmore 2018, Powers and Hampton 2019). Some members of the open science movement are pushing for the peer review process to become even more transparent, such as removing anonymity from peer review or requiring that reviews be publicly available alongside manuscripts (Ross-Hellauer 2017).

This special collection in the Journal of Insect Science is one example of how open science is becoming more prominent within the field of entomology. For researchers unfamiliar with the methods and terminology, however, open science may appear opaque. Here, we outline best practices for open science and introduce some of the tools used, along with some of the benefits of their adoption. Readers of this article may find that they are already practicing some of the recommended best practices, even though the terminology may be unique in the emerging science of Open Entomology. Although not intended as a comprehensive treatise, we hope that this article can serve as unifying introduction to the language, tools, and processes of replication and open science, which will help facilitate learning and communicating about these processes within our field. Figure 1 provides an example of an idealized open and reproducible workflow that can guide the reader during and after the reading of this manuscript.
Replication

To preserve the possibility of replication for a given study, the study must be thoroughly documented from start to finish. There are several behaviors that can affect the probability that a study will be replicable, such as bias toward publishing positive results, sending negative or inconclusive results to the ‘file drawer’, designing studies with low power, failing to detect data collection and entry errors through appropriate quality-control measures, reporting only positive findings of several statistical analyses (i.e., \( p \)-hacking), deciding to collect more data because the results are not significant, restricting reporting to significant covariates rather than all measured covariates, treating outliers with opaque criteria, and hypothesizing after results are known (often referred to as ‘HARKing’) (Csada et al. 1996, Borer et al. 2009, Pautasso 2010, Simmons et al. 2011, John et al. 2012, Gelman and Loken 2013, Head et al. 2015, MacCoun and Perlmutter 2015, Wicherts et al. 2016, Munafò et al. 2017, Fraser et al. 2018, Lazarević and Žeželj 2018). It can be difficult to identify when research suffers from the aforementioned problems under current common practices. The methods section of most papers may not provide adequate detail; data are frequently unavailable and available data may not be reusable when descriptive metadata are insufficient or errors exist. The sections that follow describe best practices to make research more replicable and reproducible. Table 1 includes a selection of resources with more in-depth information about each of the topics discussed below.

Data Curation

Proper data curation is vital; without proper data curation, data sets decrease in information value over time. Peripheral information is lost due to accidents, changes in file storage standards, moves, and/or the human limitations of researchers who switch careers, retire, or pass away (Michener et al. 1997). Data should have appropriate metadata attached that describe the structure of the data set, as well as synopses of how each variable was collected. Proper metadata facilitates electronic searches for data sets of interest (Hampton et al. 2013). Borer et al. (2009) describe in detail important steps in the data curation process: descriptions should include each variable (e.g., ‘site_id’ is the unique identifier for each site in the study), the type of data that are entered for each variable (e.g., integers, real numbers, text, etc.), allowable values (e.g., ‘Temperatures recorded should be between 20 and 33°C’), and how missing or null values are recorded (e.g., N/A values are recorded as ‘.’). The National

| Table 1. A selection of resources to learn more about reproducibility and open science |
|-----------------------------------------------|-------------------------------------------------|----------------------------------------------------------|
| **Topic**                                     | **Resource name**                                | **Information**                                          | **URL**                                             |
| Open Science                                  | Open Science Foundation                          | A suite of information and resources about most aspects of open science. | https://cos.io/                                     |
| Open Science                                  | FOSTER                                           | FOSTER is an e-learning platform with a variety of educational resources about open science. | https://www.fosterscience.eu/about                  |
| Metadata standards                            | Digital Curation Centre – What are Metadata Standards? | An introduction to metadata standards.                     | http://www.dcc.ac.uk/resources/standards-watch-papers/what-are-metadata-standards |
| Metadata standards                            | FAIR Principles                                   | The FAIR Guiding Principles for scientific data management and stewardship. | https://www.go-fair.org                             |
| Data curation and replicable analysis         | Data curation                                    | A list of different metadata standards with links to learn more. | www.dcc.ac.uk/resources/metadata-standards/list |
| Data curation and replicable analysis         | Tidyverse                                        | A page with resources to learn more about the R package ‘tidyverse’. | https://www.tidyverse.org/learnp | |
| Data curation and replicable analysis         | Data Carpentry                                    | An organization whose mission is to ‘train researchers in the core data skills for efficient, shareable, and reproducible research practices’. Their website has educational resources available for free covering topics including relational databases, data cleaning, and data analysis and visualization. | https://datacarpentry.org/ |
| Replicable analysis                           | R for Data Science                                | An introduction to the R programming language and how to work with data in R. Available for free digitally online or a hardcopy can be purchased. | https://r4ds.had.co.nz/ |
| Replicable analysis                           | Happy R with Git                                  | An online tutorial to introduce users to using both git and GitHub, as well as working with them within R. | https://happygitwithr.com/ |
| Replicable analysis                           | Docker                                           | The ‘get started’ page for Docker, a tool used to create containers for running software in a specified computing environment. | https://www.docker.com/get-started |
| Public archival repositories                  | Data repositories recommended by the journal Scientific Data | A list of public data repositories that meet requirements for data access, preservation, and stability. | https://www.nature.com/sdata/policies/repositories |
| Public archival repositories                  | Zenodo/FigShare                                  | Free public repositories for all types of digital objects associated with a research project. | https://zenodo.org/ |
| General data science                          | DataSciGuide                                      | A searchable collection of resources maintained by data scientist Renee Teate. Available resources cover a wide variety of topics related to working with data electronically, including educational resources for different software, programming languages, statistical analyses, data storage, and data visualization. | https://figshare.com/ |

**https://r4ds.had.co.nz/**
Information Standards Organization provides a comprehensive overview of metadata (Riley and National Information Standards Organization (U.S.) 2017).

Different fields of entomology may have different standards for what type of metadata should be provided and the form in which it should be provided. For example, entomological studies that consist primarily of ecological data may wish to use the Ecological Metadata Language, described in Michener et al. (1997). Different types of data such as molecular data or genomic data frequently use their own metadata formatting conventions, many of which can be found in the list of Metadata Standards from the Digital Curation Centre (http://www.dcc.ac.uk/resources/metadata-standards/list). Using a cohesive metadata standard fosters similarity among field-specific data sets, which supports data discovery by other scientists, facilitates data acquisition and comprehension, and can further enable automatic data acquisition, cleaning, and analysis (Borer et al. 2009, Hampton et al. 2013, Michener 2015). The FAIR Guiding Principles for scientific data management and stewardship provides a guide for ensuring that data (and associated data analysis products like code and algorithms) are Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). Notably, the FAIR guidelines address the importance of making data machine-accessible so that data or other data products may be easily incorporated into downstream workflows (e.g., meta-analysis). Proper metadata and data organization makes data more valuable to the broader scientific community by making it more accessible.

Data should be stored in a ‘tidy’ format, which combines the practice of formatting data in a rectangular record format, where each variable forms a column and each observation forms a row (Table 2), with relational databases, where types of observational units form separate tables that are linked by key variables, such as plot-level and site-level ID variables (Borer et al. 2009, Wickham 2014, Hampton et al. 2015). Relational databases are discussed in the next paragraph and readers are referred to Wickham (2014) for more discussion of tidy data principles. Often, the most efficient format to record data during the course of an experiment is not rectangular record format (also sometimes referred to as ‘long’ format) but instead ‘wide’ format (Table 3). Most analytical software, however, expects that data are stored in a rectangular record format. (Exceptions exist; for example, community ecologists may often analyze species abundance data in the wide format, similar to Table 3.) Researchers who record their data in the wide format may need to transform the structure of their data from wide to long for analytical software, or vice versa. Further changes may be necessary while ‘cleaning’ the data, such as checking the data for any obvious errors or calculating new variables as a combination or transformation of recorded variables. Before any changes are made to the data, however, researchers should save a version of this ‘raw’, or unedited, data immediately after it has been entered. This practice of saving a raw version of the data ensures that any mistakes introduced through data processing can be easily remedied in the future by restoring the original data file.

Within an experiment, observations may be made on different levels of experimental unit (e.g., plot, site, and region). In such instances, a relational database should be created such that each level of observational unit should have its own table of data linked to other levels of observation by a key variable (Borer et al. 2009, Hampton et al. 2015). For example, a researcher may study pest populations at multiple plots within an agricultural field, and at multiple agricultural fields across a region. All sampling locations within a single agricultural field will share variable measurements that are recorded at the field level, but may differ in variables recorded at the plot level. Additionally, agricultural fields within a single region will share measurements recorded at the region level, and will differ in those variables from fields within a different region. It would be redundant and increase the likelihood of mistakes if the observer recorded all field- and region-level variables alongside every plot observation. Eliminating redundancies in data sets stored electronically also reduces the size of the data set, which may be important if digital storage space is limited. Instead, a researcher can keep separate data tables for each level of observation: a plot-level table, a field-level table, and a region-level table. Linking these tables with an identification variable allows the data to be combined or merged later as necessary. For example, plot-level observations can be linked to the field they were collected from by including a field ID variable, and field-level observations linked to the region they were collected from with a region ID variable. Relational database structures are not limited to data from field experiments, like in the provided example, but can be used to format any data set that exhibits such a hierarchical structure.

Software or programs that help create relational databases are referred to as relational database management systems, and include software such as Microsoft Access or MySQL, among others. Regardless of the software used to create and work with the relational databases, relational databases should be stored in a structured plain-text format such as SIARD or XML. These store the relational databases in a plain-text format that maintains the relational nature of the data while allowing it to be accessed by nonproprietary software. Storing data and other research products in nonproprietary formats helps ensure the longevity of the information contained within the data sets, in the event that proprietary software like Microsoft Excel changes or disappears (Baker et al. 2006).

When steps are taken to properly curate data and the data are made publicly available in an online repository, it increases the transparency of the data handling process and makes the data more accessible to researchers. Merging data, transforming data, or data cleaning should be performed using a script programming language, such as R (www.r-project.org) or Python (www.python.org) using the pandas and numpy libraries. The nature of programming requires that a very clear set of instructions be delivered to

### Table 2. A long format way to record species counts from two different sites, also referred to as rectangular record format

| Site | Species | Count |
|------|---------|-------|
| Site A | Species 1 | 14 |
| Site B | Species 1 | 19 |
| Site A | Species 2 | 29 |
| Site B | Species 2 | 46 |
| Site A | Species 3 | 11 |
| Site B | Species 3 | 45 |

Each variable is represented by a column and each observation by a row.

### Table 3. A wide format way to record species counts from two different sites

| Species | Site A | Site B |
|---------|--------|--------|
| Species 1 | 14 | 19 |
| Species 2 | 29 | 46 |
| Species 3 | 11 | 45 |

In this table, the site variables are the columns and the species variables represent the row. Each cell with a number is a different observation.
the computer telling it how to clean the data. This practice leaves a
clear list of actions describing how variables were transformed, how
null or missing values were handled, how potential data entry errors
were identified and fixed, how data were cleaned or summarized,
and the order in which these actions were done. Moreover, most
programming scripts may be rerun if mistakes are found later or if
one wishes to incorporate more data at a later date. Most spreadsheet
software does not leave similar footprints; that is, even with both raw
data and subsequent tidy data files available, it may not be clear
easily how the tidy data version was produced. One exception
is the program OpenRefine, which is freely available online (https://
openrefine.org). OpenRefine is a free Excel-like tool that offers tradi-
tional spreadsheet functionality while also automatically producing
a reproducible script of actions taken during cleaning and analysis.

Most new tools will have an associated learning curve, but in the
case of R, there are hundreds, if not thousands, of resources
freely available on the internet to learn the basics of working in the
R environment. Novice programmers may wish to investigate R Commander, a graphical user interface for R that provides drop-down menus for common commands used in R. For users interested in learning more about programming, a good place to start is the Data Carpentry website (https://datacarpentry.org), which offers free workshops and lessons in data analysis skills, or the book R for Data Science (Wickham and Grolemund 2016), which is available for free online (https://r4ds.had.co.nz/).

Replicable Analyses
To permit another user to replicate an analysis, it is important to
document all choices made during the analysis. Any analyses per-
formed should be curated similarly to the data to facilitate creating
results that are replicable. However, it is also important that careful
consideration to the analysis be done prior to carrying out the ex-
periment. This practice is formalized in the process of preregister-
ing a study, which is a common practice in the medical field. When a
study is preregistered, the study authors are asked to think critically,
specifically, and exhaustively about their prior hypotheses, methods, and analysis (Wicherts et al. 2016). This includes, but may not be limited to, a specific statement of the hypotheses to be tested, how the data will be analyzed, and how any issues such as violations of statistical test assumptions or how outliers will be handled. Preregistrations are submitted to a granting agency or posted publicly online and serve as a record to compare the final manuscript, if one is produced, to the original plan. These steps help separate exploratory analysis from confirmatory analysis. This separation reduces the temptation to hypothesize after the results are known, which can lead to bias in publications and inflated false-positive rates (Munafo et al. 2017). ‘Researcher degrees of freedom’, defined as the methodological and analytical choices made by an investig-
ator from the start of an investigation to the end, are often not completely documented even though seemingly arbitrary decisions may influence the final presentation of results (Gelman and Loken 2013, Wicherts et al. 2016). Thoughtful layout of the statistical plan during preregistration prevents researchers from intentionally or unintentionally abusing researcher degrees of freedom.

As with data cleaning, analysis and visualization should be done
using a script programming language. Working within a script pro-
gram serves as a notebook to document what analyses were run and how assumptions of those analyses were tested, serving as a record of all results. Maintaining a script of the analysis makes it easier to update results or graphics if errors in the data or analyses are caught at a later point, without additional intellectual overhead spent on trying to remember exactly what was done. Issues may arise, however, when rerunning scripts where versions of the operating system and software have changed. Tools that can aid in the port-
ability of analyses include Binder (https://mybinder.org) and Docker (https://www.docker.com). These tools are used to create ‘containers’ of the original running environment and any software dependencies, which allows future users to rerun analyses in the original environment without needing to laboriously recreate the original environment on their machine. This step helps ensure that if discrepancies are noticed in the future they can be attributed to something other than the computing environment. More thorough discussion of soft-
ware containers is beyond the scope of this article and are available elsewhere (Chamberlain and Schommer 2014; Boettiger 2015, 2017; Di Tommaso et al. 2015; Hung et al. 2016; Qasha et al. 2017).

Computational notebooks such as Jupyter Lab (https://jupyter.
org) allow researchers to maintain a single document that contains code and its associated output, extended prose for explaining code or documenting ideas during analyses, and other resources; these notebooks are considered by some to be the digital computational version of a lab notebook. Notebooks can be used to interweave code, text, and the results of an analysis or a graph to provide more human-readable documentation of the data exploration and analysis process. Jupyter Lab resides in ‘the cloud’ and, when coupled with Binder, can be used to recreate the appropriate computing envi-
ronment regardless of the current user’s system. These steps ensure interoperability and can address concerns of portability raised in the previous paragraph.

An additional step beyond using a scripting language for data cleaning and analyses is incorporation of a distributed version control system, such as ‘git’, that is very useful for tracking changes to programming scripts or other plain-text documents. As changes or additions are made to analyses, the user can ‘commit’ changes to a repository. A repository tracks all changes made between versions of files in the repository folder and becomes a history of file changes over time. Each commit can be accompanied with a message describing the changes that have been made during the editing process, which provides more thorough documentation regarding the differences between document versions than a simple version number. For example, when computer code to execute a statistical analysis suddenly stops working while elaborating on the script/ analysis, a researcher can easily restore the file to a commit point where the program was working. Git is also able to highlight differences between versions of a file, which helps safeguard against accidental changes to components that should have remained static.

It should be noted that git works best in tracking and comparing changes made to plain-text files (such as programming scripts). Git will keep track of different versions of more complex files, such as PDFs, but will not display comparisons between different versions of these more complicated documents.

The git program is a command-line program, requiring the user to input text commands directly into a computer terminal, which can be intimidating to a beginner. Graphical user interface (GUI) programs exist to make using git much easier. Such programs include GitKraken or GitHub Desktop, while R Studio has git GUI func-
tionality built into the program. This makes integrating git into the workflow of an R user much easier. Storage of repositories is often linked to an associated cloud service, such as GitHub, GitLab, or BitBucket, making it accessible to researchers from anywhere with internet access rather than depending upon a laptop hard drive or USB thumb drive that are prone to being misplaced. After making changes to files locally on a personal computer and committing them to the local repository, the user can ‘push’ these changes from the local repository to the cloud-based repository. Likewise, one can
‘pull’ updates from public cloud repositories to their personal computer or from private cloud repositories to which they have access. The suite of office programs provided by Google (e.g., Docs, Sheets, Slides, etc.) also provides a more accessible, albeit limited, form of version control for text documents, spreadsheets, and slide shows.

Preprint Servers

A preprint server is a webpage where researchers may upload ‘preprints’, or early drafts, of manuscripts before they are submitted to a journal. Manuscripts submitted to a preprint server are subject to a screening process to ensure that the content is scientific, inoffensive, and not plagiarized, but the maintainers of the preprint server do not conduct any official editing or peer-reviewing functions. Instead, other users are able to provide comments on manuscripts that are uploaded. Preprint servers facilitate more public discussion and criticism of works, and allow findings to become immediately available to other researchers (Ross-Hellauer 2017, Elmore 2018). Releasing a preprint on an appropriate server is later associated with as much as a 39% higher rate of citations (Fu and Hughey 2019). Preprint servers have their foundation in the physical sciences, where arXiv (pronounced ‘archive’) has been used by physicists since the early 1990s. The preprint server bioRxiv (pronounced ‘bio-archive’) was started in 2013 to serve as a preprint repository for the biological sciences. (A preprint version of this manuscript can be found on the preprint server Preprints, alongside reviewer comments and the final submission of this work.)

A manuscript on a preprint server may be updated after the authors have received feedback from the scientific community. The version of the paper is listed alongside the preprint, allowing it to serve as a record of changes between the original draft of a manuscript and the final published version. Once a version of the preprint has been published, most preprint servers allow the authors to update the original submission with a final version of the article. Authors that submit a manuscript to a preprint server retain the copyright to their work and preprint submissions are assigned a digital object identifier (DOI), allowing them to be cited. Many journals have citation guidelines for citing preprint papers, although if the preprint has since been published in a journal it is best practice to cite the version published in the journal. The preprint website bioRxiv updates the preprint manuscript with a link to the published version once available. It is always good practice to check target journal submission policies before submitting a manuscript to a preprint server. Some journals disallow submission to preprint servers before submitting to the journal, while others will not allow updates to the preprint article while the article is under review at the journal.

Open Access Publishing

Open access (OA) publishing is the process of removing barriers to accessing and sharing research published in scholarly journals. OA is often divided into two main types: Gold OA, where the publisher provides OA to the article, and Green OA, where the author(s) of the article archives the published work in an openly accessible space, such as a personal website, preprint server, or public repository (Laakso et al. 2011, Tennant et al. 2016). Other distinctions exist, such as Hybrid OA where authors can choose to pay fees to make their article open and delayed OA where all articles are converted to OA after a specified time frame (Martin-Martin et al. 2018). Gold OA often requires the author to pay extra fees associated with the open publication of their article, whereas Green OA may be free or at low cost depending on where the article is made available. With the growth of the internet and the shift from traditional print publishing to more web-based publishing, the number of OA journals and OA journal articles has increased by 18% and 30%, from 2000 to 2009, respectively (Kaiser 2010, Laakso et al. 2011). Increases in the number of OA articles have continued to grow since 2009, albeit at a slower rate (Piwowar et al. 2018).

The growth in OA publishing can be linked to growing beliefs that scientific research, especially publically funded research, is a public good that should be freely available (Paul et al. 2010, Grand et al. 2012, Tennant et al. 2016). Proponents of OA publishing cite an increase in public engagement (Strodten 2010), an increase in public trust (Grand et al. 2012), and a decrease in inequality among countries and research institutions (Odlyzko 2006) as some of the societal benefits that result from making research more readily available. Additionally, OA articles have been shown to have higher citation rates than those locked behind paywalls (Antelman 2004, Eysenbach 2006, Tennant et al. 2016, Piwowar et al. 2018). A good review of the academic, economic, and societal impacts of OA publishing can be found in Tennant (2016).

There are concerns about the cost of OA publishing. Journals may charge in excess of $1,000 USD to publish an article OA, presenting a barrier to research groups who cannot afford to pay such fees. Gold OA is more expensive than Green OA. Some institutions and funders have funds available to support OA publishing or allow researchers to expense publication fees on grants. The Public Library of Science (PloS) maintains a partial list of such funding sources from across the global (https://plos.org/open-access-funds).

Public Repositories

To better ensure the availability of digital research materials into the future, such materials should be uploaded to a public archiving repository in a nonproprietary file format. Data that consist of text and numbers in rectangular record format are commonly stored in a comma-delimited file (‘.csv’), for example. A csv file is a basic text file that contains all the data values separated or delimited by commas. Other formats exist, such as tab- or space-delimited files. Most modern spreadsheet software is able to read basic delimited text files and convert them into a spreadsheet format that is more human-readable and works with the analytical software. As a nonproprietary format for storing data, a csv file is preferred to other proprietary formats, such as ‘.xls’ files produced by Microsoft Excel.

Public archiving repositories are different from a version control repository such as git, described earlier. A git repository is a repository where the work takes place and changes are tracked over time. These repositories are meant to be interacted with frequently and will change dynamically over time. A public archiving repository is meant for long-term archival of digital products and is more static over time. They may also include files like slideshows or PDFs that are not as easily tracked by a program such as git, which works best with plain-text files.

Public archival repositories remove the requirement for the researcher to maintain and provide research material as requested. Multiple studies have shown that researchers are not always able to provide the data associated with a manuscript; uploading data and code to public repositories increases the likelihood that such materials are available (Leberg and Neigel 1999, Wicherts et al. 2006, Savage and Vickers 2009, Vines et al. 2013). One study looking at 516 articles in the field of ecology found that data availability decreases through time; authors of papers published 20 yr ago could provide the associated data less than 50% of the time (Vines et al. 2014). Making data publically available also allows researchers to show that results are robust by attempting to analyze.
the same data in different ways or incorporating the data into larger meta-analyses.

There are a variety of public archival repositories that will accept many types of electronic files. Many academic institutions operate their own public data repositories that can be used by associated researchers at minimal cost (i.e., even free). One popular repository not affiliated with a university is Dryad (https://datadryad.org), which specializes in life sciences data and code. Other repositories accept other types of data, such as GenBank (www.ncbi.nlm.nih.gov/genbank/) for genetic sequence data. There may be costs associated with submitting to certain repositories, but at least one ecological journal at the time of writing, *Oikos*, has integrated their submission system with Dryad and will cover the Data Processing Charge. The journal *Scientific Data* provides a list of public data repositories across a variety of scientific disciplines (https://www.nature.com/sdata/policies/repositories).

Zenodo (https://zenodo.org/) and FigShare (https://figshare.com) are both commonly used general purpose public repositories where any type of digital product related to research can be uploaded for free, within certain file size limits. Each object is assigned a DOI for citation purposes. Possible uploads include data and associated code, but also materials generated for conferences, workshops, or other types of projects. Both platforms allow users to store materials privately if they do not wish to share materials prior to publication. FigShare includes a desktop application which can be linked to a personal account online, making it simple to upload objects directly from one’s computer. For the advanced user, both platforms also include functionality to access data and files on their servers through their application programming interface (API).

Some data may include sensitive information that should not be made publicly available for legal or ethical reasons, such as health information, the location of endangered species, or other personal information. Steps can often be taken to deidentify data or remove the sensitive information and still post most of the data to a public repository. If the data are unusable without the sensitive information, authors should explicitly state so in a data availability statement. Data availability statements are becoming more common as publishers and funders adopt policies requiring researchers to share their data and should provide a brief description of where the data supporting results may be found (Federer et al. 2018).

Summary

Reproducible, replicable, robust open science is vital for ensuring inferences are valid and reliable. It may take extra care and time to ensure research is reproducible, but often these costs are outweighed by the benefits to ourselves and other researchers. The learning curve associated with incorporating these practices into the scientific workflow may be daunting, especially in light of the busy schedule already maintained by most scientists. It is our hope that this article will provide researchers with the foundation to begin incorporating these tools slowly and to seek out resources on their own, and as their schedule allows.

By making tools and data more widely available, the transparency and trust in science as an institution are increased (Stodden 2010, Grand et al. 2012, Grimes et al. 2018). Additionally, we increase the value of our work when we view every step in the research cycle, not just the final manuscript, as a potential valuable product and make those products publicly available (Hampton et al. 2015, Roche et al. 2015, Munafo et al. 2017). Data, code, and the specific details of our work all provide value to the scientific community when they are easily accessible. Data that are made publically available with appropriate metadata can be used by other researchers in meta-analyses or reanalyzed to ensure the robustness of results (Michener 2006, 2015; Piwowar et al. 2007). Performing analyses in a script programming language provides a record of what was done with the data, which may bolster confidence in results and allows others to apply analytic methods to their own data (Borer et al. 2009, Prager et al. 2019). Depositing all digital objects associated with a research project in a public repository and publishing manuscripts in an OA format helps ensure that mistakes are caught earlier and can help ensure others do not waste time chasing scientific dead ends (Molloy 2011, Lowndes et al. 2017).

Science serves an important role in confronting many global issues, such as resource use, invasive species, viral pandemics, and climate change, and maintaining societal trust in the scientific process is vital to ensure that scientists remain trusted sources of information on these issues (Leiserowitz et al. 2013). Practicing open science helps ensure that science, especially publicly funded science, remains transparent and accessible to the public and helps maintain public trust (Beardsley 2010).

Acknowledgments

This work was supported by United States Department of Agriculture McIntire-Stennis project MIN-17-095 and the College of Food, Agricultural, and Natural Resource Sciences at the University of Minnesota. We thank A. K. Tran, M. J. Hallinen, and S. Robinson (University of Minnesota) and two anonymous reviewers for helpful insights that improved earlier drafts of this manuscript.

Open Entomology

This paper is part of an Open Entomology collection in the *Journal of Insect Science*. Papers in the collection were submitted to a preprint server upon submission to the journal, have publicly available data (when applicable), and have peer review information published as supplemental files, including the original version of the paper, decision letters and reviewer files, and revised versions.

References Cited

Antelman, K. 2004. Do open-access articles have a greater research impact? Coll. Res. Lib. 65: 372–382.

Baker, M., M. Shah, D. S. Rosenthal, M. Roussopoulos, P. Maniatis, T. Giuli, and P. Bungale. 2006. A fresh look at the reliability of long-term digital storage, pp. 221–234. In Proceedings of the 2006 Eurosys Conference. ACM Press, New York, NY.

Barba, L. A. 2018. Terminologies for reproducible research. eprint arXiv:1802.03311. 1802.03311.

Beardsley, T. M. 2010. The biologist’s burden. BioScience. 60: 483–483.

Boettiger, C. 2015. An introduction to Docker for reproducible research. Oper. Syst. Rev. 49: 71–79..

Boettiger, C. 2017. A reproducible R notebook using docker, pp. 109–118. In J. Kitzes, D. Turek, and F. Deniz (eds.), The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. University of California Press, Oakland, California..

Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. 2009. Some simple guidelines for effective data management. Bull. Ecol. Soc. Am. 90: 203–214.

Chamberlain, R., and J. Schommer. 2014. Using Docker to support Reproducible Research. https://doi.org/10.6084/m9.figshare.1101910.v1.

Csada, R. D., P. C. James, and R. H. M. Espie. 1996. The “file drawer problem” of non-significant results: does it apply to biological research? *Oikos*. 76: 591–593.

Di Tommaso, P., E. Palumbo, M. Chatrou, P. Prieto, M. L. Heuer, and C. Notredame. 2015. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 3: e1273.
Tennant, J. P., F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, and C. H. Hartgerink. 2016. The academic, economic, and societal impacts of open access: an evidence-based review. F1000. 3: 632.

The Turing Way Community. B. Arnold, L. Bowler, S. Gibson, P. Herterich, R. Higman, A. Krystalli, A. Morley, M. O’Reilly, and K. Whitaker. 2019. The Turing Way: a handbook for reproducible data science (Version v0.0.4). Zenodo. http://doi.org/10.5281/zenodo.3233986.

Uhlir, P. F., and P. Schröder. 2007. Open data for global science. Data Sci. J. 6: OD36–OD53.

Vale, R. D. 2015. Accelerating scientific publication in biology. Proc. Natl. Acad. Sci. USA. 112: 13439–13446.

Vines, T. H., R. L. Andrew, D. G. Bock, M. T. Franklin, K. J. Gilbert, N. C. Kane, J. S. Moore, B. T. Moyers, S. Renaut, D. J. Rennison, et al. 2013. Mandated data archiving greatly improves access to research data. FASEB J. 27: 1304–1308.

Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J. S. Moore, S. Renaut, and D. J. Rennison. 2014. The availability of research data declines rapidly with article age. Curr. Biol. 24: 94–97.

Wicherts, J. M., D. Borsboom, J. Kats, and D. Molenaar. 2006. The poor availability of psychological research data for reanalysis. Am. Psychol. 61: 726–728.

Wicherts, J. M., C. L. Veldkamp, H. E. Augusteijn, M. Bakker, R. C. van Aert, and M. A. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. Front. Psychol. 7: 1832.

Wickham, H. 2014. Tidy data. J. Stat. Softw. 59: 1–23.

Wickham, H., and G. Grolemund. 2016. R for data science, 1st ed. O’Reilly Media, Sebastopol, CA.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. 2016. Comment: the FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 3: 1–9.