Open Science, Open Plant Pathology

Adam H. Sparks\*,1,✉, Emerson Del Ponte\*,2, Kaique dos Santos Alves2, Zachary Foster3, and Niklaus J Grünwald4

21 April, 2020

Open research practices have been highlighted extensively during the last ten years in many fields of study as essential standards needed to promote transparency and reproducibility of scientific results. In fact, scientific claims can only be evaluated based on how protocols, materials, equipment and methods were described; data were collected and prepared; and analyses conducted. Openly sharing protocols, data and computational code are central for current scholarly dissemination and communication, but in many fields, including plant pathology, adoption of these practices has been slow. We randomly selected 300 articles published from 2012 to 2018 across 21 journals representative of the pathology discipline, and assigned them scores reflecting their openness and reproducibility. We found that most of the articles were not very open, failing to share data or code in a meaningful way. We propose that using open source tools for producing open and reproducible work and analysis is advantageous, benefiting not just readers, but the authors as well and provide ideas and tools to promote open, reproducible research practices among plant pathologists.

\* These authors contributed equally to this work.

1 University of Southern Queensland, Centre for Crop Health, Toowoomba, Queensland, Australia  
2 Departmento de Fitopatologia, Universidade Federal de Viçosa, Brazil  
3 Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA  
4 Horticultural Crops Research Unit, USDA Agricultural Research Service, Corvallis, OR, USA

✉ Correspondence: [Adam H. Sparks <[Adam.Sparks@usq.edu.au](mailto:Adam.Sparks@usq.edu.au)>](mailto:Adam.Sparks@usq.edu.au)

# Introduction

Modern plant pathological research has many facets given the array of disciplines and sub-disciplines currently involved. Collectively, they contribute to increase our basic and applied knowledge on several aspects of pathogen biology and disease development to ultimately improve plant disease management. Scientific research in the field varies from the purely observational or descriptive nature to inferential based on experimental or simulation-derived data sets. Whatever the case, research findings are verifiable based on how much of the research materials, processes and outcomes are made available beyond what is reported in the scientific article and the ability of others to make use of your methods and results. These research findings include biological materials (isolates or strains), nucleic/protein sequences, experimental and simulated raw data annotations, drawings and photographs and statistical analysis code among other materials and data generated as a course of conducting research. That is, open science leads to reproducibility and replicability.

## Definitions

In order for us to easily discuss the topic, we first must define what we mean so that we may clearly communicate. Many of the terms used in this area have varying definitions that may or may not agree with each other. For instance, reproducible research is recently highlighted by many authors (Baker 2016; Iqbal et al. 2016; Nature 2016; Patil et al. 2016; Weissgerber et al. 2016; Brunsdon 2015; Sweedler 2015; Fitzjohn et al. 2014; Ioannidis 2014; Fidler and Gordon 2013; Stodden et al. 2013) as an important issue. The International Union of Pure and Applied Chemistry provides definitions of repeatability in the “Compendium of Chemical Terminology – Gold Book” as:

“The closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time). The measure of repeatability is the standard deviation qualified with the term: `repeatability’ as repeatability standard deviation. In some contexts repeatability may be defined as the value below which the absolute difference between two single test results obtained under the above conditions, may be expected to lie with a specified probability”, (McNaught and Wilkinson 1997).

and reproducibility is defined as:

“The closeness of agreement between independent results obtained with the same method on identical test material but under different conditions (different operators, different apparatus, different laboratories and/or after different intervals of time). The measure of reproducibility is the standard deviation qualified with the term ‘reproducibility’ as reproducibility standard deviation. In some contexts reproducibility may be defined as the value below which the absolute difference between two single test results on identical material obtained under the above conditions, may be expected to lie with a specified probability. Note that a complete statement of reproducibility requires specification of the experimental conditions which differ”, (McNaught and Wilkinson 1997).

However, while these definitions are clear and well accepted in physical sciences, in biological sciences it is not always possible to use identical test material or perhaps the the time or resources are not available for full reproduction, *e.g.*, field trials that span years and locations or complex glasshouse experiments. Therefore we will follow Peng’s (2009) definition that provides clear guidelines a minimum standard of ``reproducible research’’:

“The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is ``reproducible research’’, which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses”. Peng, R. (2009). *Reproducible research and Biostatistics*. Biostatistics, 10 (3): 405-408.

### Open Science

Open science has become highlighted lately with many donors expecting data to be available (Government of Canada 2016; van Noorden 2017; Australian Research Council 2018) and other scientists interested in sharing and collaborating more widely (Wald 2010).

# A General Workflow

A general workflow for producing academic research involves clearly defining a research question, obtaining data for testing the hypothesis, summarizing/analyzing and presenting data and results, and writing the manuscript. Here we defined three levels of reproducibility which are also related with the evolution of computational methods and reproducible practices 1.

A first level of reproducibility involves openly available research materials such as isolates or strains and/or nucleic acid sequences in public collections and citations for methods used. A second level involves providing raw data and code as binary files (PDF or other non-text file) in supplemental materials, which do not allow prompt access to the data and running the code because of use of expensive commercial software or a pay-wall. The highest level includes efforts to annotate structured raw data (FAIR – Findable, Accessible, Interoperable and Re-usable) (Wilkinson et al. 2016) and fully document the analysis using open source code which are deposited in public repositories and can be run by anyone following download of data and code. The first level, as reported, is an essential step that is not substituted by the other practices and eventually researchers fail to provide sufficient description or correct citations. In the next section we present standards and tools that can be used to ensure reproducibility.

### Methods

When making your science more open and reproducible, methods, software used, this includes items such as R or Python packages that were directly used in the analysis or production of the paper, etc. should be cited properly. Deposit and annotate biological materials with an herbarium or other repository. Provide a full description of all equipment used, *e.g.* a Spectrum Technologies Watchdog 2700 weather station was used to record wind speed, direction, rainfall, temperature and relative humidity at one hour intervals. This allows end users to identify what was used and identify the methods used more accurately. Just as importantly, this acknowledges the contributions of others whose works were instrumental in your research. This also helps ensure that you as a researcher are able to reconstruct what you have done since you will have good notes and documentation and be able to identify if something changes, *e.g.* a package version, what effect it had on your research.

The use of programming or scripting languages such as R, SAS or Python enable you as a scientist to keep very detailed records of what was computationally performed. This is as opposed to using software such as spreadsheet programs like Excel, gsheets, Numbers, Calc or others which can be used for simple statistical analyses and visualization (but should not be) or other point-and-click software packages that do not enable you to keep an accurate record of the steps taken to import, format, visualise and analyse data.

### Data

Prefer text files for saving small sets of data. Data that are saved in binary formats such as PDF files are difficult to reuse because they are not easily machine readable. In many cases, data sets are small enough and curated in spreadsheets, which should be saved as a plain text file, *e.g.* comma separated (CSV) files. This also helps ensure that the data are reusable.Larger data sets may warrant the use of a proper database like MariaDB or PostgreSQL which provide users with several benefits but two important ones to mention here are data redundancy, no duplicated records; and data consistency, all of your records are sure to be recorded in the same way for every observation. While databases may offer many advantages, the trade-off is that they are more complex to set up and administer, especially for a small data set that might be only a few rows and columns in a database.

Ideally, once the data are complete, best practices for keeping your data as you perform your work include treating the raw data as read-only and using file permissions to prevent changes to the raw data files. It should be noted, that the use of a database management system also allows for both of these at the expense of added complexity. Saving files in proprietary formats such as .xls(x) can also lead to issues in the future when opening using newer (or older) software versions. Unexpected changes to values in the data (Ziemann et al. 2016) may also occur when using proprietary formats.

If steps are followed to make the data FAIR then it will be readable by humans and machines alike this will help support new discoveries and support further research. In turn sharing data will lead to new citations for your work as other discover and use it.

To make your data the most widely useable, ensure that it has a persistent identifier. A digital object identifier (DOI) is the most common (<https://www.doi.org/>) but handle is also an option (<https://handle.net/>). There are different options for generating a DOI for your data and other materials. FigShare, Zenodo and OSF all offer long-term archival along with a service to generate a DOI for your materials. The use of a persistent identifier works to ensure that even if the data are moved, they can still be located using that unique identifier.

For more on FAIR data, visit Go-Fair <https://www.go-fair.org/fair-principles/>.

### Sharing Your Research

Once you have determined how to best manage your source code for analysis and the data sets the next step is to consider how to share your data. Providers such as FigShare, Dataverse, OSF.io and Zenodo allow for you to deposit your data and generate a DOI for sharing your project once you are finished with it.

Other providers exist that allow for you to not only track changes but also to share the data openly, these include GitHub, Gitlab and Bitbucket. GitHub is arguably the most popular and widely used software development platform currently. Data that is encoded in CSV or other plain text formats can easily be deposited in a repository along with code for analysis to enable changes to be tracked and other users to download and replicate the work.

We would advise against the practice of depositing data on a laboratory website or a site such as GitHub only though. Doing either of these leaves the work in an unstable state where future users may be unable to access the work. It is a best practice to always ensure that you have deposited the data with a provider such as Zenodo, FigShare or OSF and generated a DOI for the materials to help ensure continued accessability. Many of these providers provide rather easy ways to link the project with a software development repository to help ensure that the data are available in perpetuity.

Readers should also consult with their local librarians about local resources. Most universities provide a facility for staff to deposit papers and other academic materials, but this may extend to software development respositories in some cases up as well.

Readers are encouraged to avoid using lab websites and other personal pages for sharing projects over the longer term as they are fraught with link-rot and are not an optimal way to share your projects.

# Status in Plant Pathology

* Madden et al. (2015) supply an *e-****X****tra*\* with reproducible examples for readers.
* Duku et al. (2016) provide models, data and code, (<http://adamhsparks.github.io/MICCORDEA/>) necessary to replicate the entire study modelling the effects of climate change on rice bacterial blight and rice leaf blast in Tanzania.
* Sparks et al. (2011, 2014) provide models, data and code, (<http://adamhsparks.github.io/Global-Late-Blight-MetaModelling/>) necessary to replicate model development and the subsequent the study on the effects of climate change on potato late blight.
* Del Ponte provides data and a reproducible report that explain in details all steps of the analysis and the R codes for conducting a meta-analysis for assessing heterogeneity in relationship between white mold incidence and soybean yield and between incidence and soybean tied.
* Example from Grünwald lab:
  + paper <http://apsjournals.apsnet.org/doi/full/10.1094/PHYTO-12-14-0350-FI>
  + github repo <https://github.com/grunwaldlab/Sudden_Oak_Death_in_Oregon_Forests>
* Other examples from plant pathology providing e-Xtras or supplemental material

Twenty-one plant pathology discipline journals were selected by the authors as representations of discipline-based journals target by the plant pathology research community. Among them, both fundamental and/or applied as well as journals covering specific group of pathogens/plants or broad areas were included. Two hundred articles were randomly selected from issues published from 2012 to 2018. A list of randomly selected pages was assigned to a randomized list of the 21 journals (Sparks et al. 2017) where the page number fell within an article for the given journal. In cases where an article was not suitable, *e.g.*, a review or otherwise not related to plant pathology, the next article was selected until a suitable article was found. Notes regarding the selection of articles can be found in the file, XXXX, available in this paper’s repository. The pages list was numbered from page one and went to 150. This was done since some journals restart their numbering with each issue and also ensures that the journal is more likely to have a page number corresponding to the randomly generated value. This also assumes that there is no effect or bias on reproducibility based on the time of year that an article was published, since most journals start with page number one at the beginning of the year. The list of journals was saved as a comma separated value (CSV) file and imported into R (R Core Team 2020).

# Discussion

# Acknowledgements

##### pagebreak

# References

Australian Research Council. 2018. ARC Open Access Policy. Available at: <https://www.arc.gov.au/policies-strategies/policy/arc-open-access-policy> [Accessed April 20, 2020].

Baker, M. 2016. Muddled meanings hamper efforts to fix reproduciblity crisis. Nature News.

Brunsdon, C. 2015. Quantitative methods I: Reproducible research and quantitative geography. Prog. Hum. Geogr.

Duku, C., Sparks, A. H., and Zwart, S. J. 2016. Spatial modelling of rice yield losses in tanzania due to bacterial leaf blight and leaf blast in a changing climate. Climatic Change. 135:569–583.

Fidler, F., and Gordon, A. 2013. Science is in a reproducibility crisis: How do we resolve it? Phys.org. Available at: [http://phys.org/news/2013-09-science-crisis.html?utm{\\_}content=bufferde383{\{}{\&}{\}}utm{\\_}source=buffer{\{}{\&}{\}}utm{\\_}medium=twitter{\{}{\&}{\}}utm{\\_}campaign=Buffer{\{}{\#}{\}}ajTabs](http://phys.org/news/2013-09-science-crisis.html?utm%7B\_%7Dcontent=bufferde383%7B\%7B%7D%7B\&%7D%7B\%7D%7Dutm%7B\_%7Dsource=buffer%7B\%7B%7D%7B\&%7D%7B\%7D%7Dutm%7B\_%7Dmedium=twitter%7B\%7B%7D%7B\&%7D%7B\%7D%7Dutm%7B\_%7Dcampaign=Buffer%7B\%7B%7D%7B\#%7D%7B\%7D%7DajTabs).

Fitzjohn, R., Pennell, M., Zanne, A., and Cornwell, W. 2014. Reproducible research is still a challenge. ROpensci.org. 2014 Available at: <http://ropensci.org/blog/2014/06/09/reproducibility/>.

Government of Canada, I. 2016. Tri-Agency Open Access Policy on Publications - Science.Gc.ca. Tri-Agency Open Access Policy on Publications. Available at: <http://www.science.gc.ca/eic/site/063.nsf/eng/h_F6765465.html?OpenDocument> [Accessed April 12, 2019].

Ioannidis, J. P. A. 2014. How to make more published research true. PLoS Med. 11.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. A. 2016. Reproducible Research Practices and Transparency across the Biomedical Literature. PLoS Biol. 14:1–13.

Madden, L. V., Shah, D. A., and Esker, P. D. 2015. Does the *P* value have a future in plant pathology? Phytopathology. 105:1400–1407.

McNaught, A. D., and Wilkinson, A. 1997. *Compendium of Chemical Terminology*. Second Edition. Blackwell Science Oxford.

Nature, E. at. 2016. Reality check on reproducibility. Nature. 533:437.

Patil, P., Peng, R. D., and Leek, J. 2016. A statistical definition for reproducibility and replicability. bioRxiv. Available at: <http://biorxiv.org/content/early/2016/07/29/066803>.

Peng, R. D. 2009. Reproducible research and Biostatistics. Biostatistics. 10:405–408.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Sparks, A. H., Forbes, G. A., Hijmans, R. J., and Garrett, K. A. 2011. A metamodeling framework for extending the application domain of process-based ecological models. Ecosphere. 2:art90.

Sparks, A. H., Forbes, G. A., Hijmans, R. J., and Garrett, K. A. 2014. Climate change may have limited effect on global risk of potato late blight. Global Change Biology.:3621–3631.

Sparks, A. H., Ponte, E. M. D., Foster, Z., and Grünwald, N. J. 2017. Reproducible-research-in-plant-pathology. Available at: <https://github.com/openplantpathology/Reproducible-Research-in-Plant-Pathology>.

Stodden, V., Guo, P., and Ma, Z. 2013. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. PLoS One. 8:2–9.

Sweedler, J. V. 2015. Striving for Reproducible Science. Anal. Chem. 87:11603–11604.

van Noorden, R. 2017. Gates Foundation demands open access. Nature. 541:270–270.

Wald, C. 2010. Scientists embrace openness.

Weissgerber, T. L., Garovic, V. D., Winham, S. J., Milic, N. M., and Prager, E. M. 2016. Transparent reporting for reproducible science. J. Neurosci. Res. 00:1–6.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3:1–9 Available at: <https://www.nature.com/articles/sdata201618> [Accessed April 20, 2020].

Ziemann, M., Eren, Y., and El-Osta, A. 2016. Gene name errors are widespread in the scientific literature. Genome Biology. 17:177 Available at: <https://doi.org/10.1186/s13059-016-1044-7> [Accessed April 20, 2020].

##### pagebreak

##### pagebreak

# Figures

![Figure 1: An open and reproducible research workflow.](data:application/pdf;base64,)

Figure 1: An open and reproducible research workflow.

##### pagebreak

# Colophon

This report was generated on 2020-04-21 13:45:41 using the following computational environment and dependencies:

#> R version 3.6.3 (2020-02-29)  
#> Platform: x86\_64-apple-darwin15.6.0 (64-bit)  
#> Running under: macOS Catalina 10.15.4  
#>   
#> Matrix products: default  
#> BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib  
#> LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib  
#>   
#> locale:  
#> [1] en\_AU.UTF-8/en\_AU.UTF-8/en\_AU.UTF-8/C/en\_AU.UTF-8/en\_AU.UTF-8  
#>   
#> attached base packages:  
#> [1] stats graphics grDevices utils datasets methods base   
#>   
#> other attached packages:  
#> [1] rsvg\_1.3   
#> [2] DiagrammeRsvg\_0.1   
#> [3] DiagrammeR\_1.0.5   
#> [4] Reproducibility.in.Plant.Pathology\_0.0.1-1  
#>   
#> loaded via a namespace (and not attached):  
#> [1] Rcpp\_1.0.4.8 rstudioapi\_0.11 knitr\_1.28 magrittr\_1.5   
#> [5] prompt\_1.0.0 rlang\_0.4.5 highr\_0.8 stringr\_1.4.0   
#> [9] tools\_3.6.3 visNetwork\_2.0.9 xfun\_0.13 htmltools\_0.4.0   
#> [13] yaml\_2.2.1 digest\_0.6.25 crayon\_1.3.4.9000 bookdown\_0.18   
#> [17] RColorBrewer\_1.1-2 htmlwidgets\_1.5.1 clisymbols\_1.2.0 curl\_4.3   
#> [21] glue\_1.4.0.9000 evaluate\_0.14 rmarkdown\_2.1 V8\_3.0.2   
#> [25] stringi\_1.4.6 compiler\_3.6.3 jsonlite\_1.6.1

The current Git commit details are:

#> Local: master /Users/adamsparks/Sources/GitHub/Analysis/Reproducibility\_in\_Plant\_Pathology  
#> Remote: master @ origin (git@github.com:openplantpathology/Reproducibility\_in\_Plant\_Pathology.git)  
#> Head: [90fb309] 2020-04-21: Delete source code section