Title:
“DataDeps.jl: Automatic Data Setup for Replicable Computational and Data Science Research”

The open science movement has made great strides in making it possible to reproduce scientific works. The release of data and of source code for research scripts allows simulations and statistical analysis to be repeated. However, in practice it is often difficult even given the source code and documentation to actually replicate someone else’s work. The issue is often not in replicating their results, but rather in getting their research scripts to run at all. These issues come down to problems in matching their computational environment. All non-trivial research code has some dependencies: on software and on data.

Software dependencies are now well managed by modern package managers. However, data dependencies have not enjoyed the same treatment. Instead dependencies on data are managed either by ad-hoc solutions, or far more commonly via manual methods. This manual data management normally involves placing instructions in a readme to for example: download a file from a site, extract it either to a particular location, and/or reconfigure the script to reference its location. This manual process introduces opportunities for mistakes, prevents fully automated testing, and takes up time that could be better spent on scientific pursuits. This paper highlights this need for fully automated data dependency setup, and discussed our solution for the Julia programming language: DataDeps.jl.

The DataDeps.jl package provides data dependency management. It performed automatic download of data, when and if it is required – it does not download data again unnecessarily, nor data that is never used. It also provides automatic location of data on disk, allowing data to be storied in any of multiple locations. For example allowing the data to be moved from a large slow network store to a fast local disk without requiring reconfiguration. It also makes it simple to display information to the users, such as bibliographic details for a paper that should be cited when using the data. DataDeps.jl is useful for research scripts depending on data as it allows them to be more easily executed in new environments. DataDeps.jl is also for also for other libraries depending on data – such as a machine learning based tool which needs to download pretrained models.

This paper also discusses DataDepGenerators.jl, which allows for the easy reuse of existing data. DataDepsGenerators takes as input an id (e.g. URL) for a dataset on a major data repository (e.g. DataDryad, UCI ML repository, etc), and it outputs the code to define that dataset for DataDeps.jl. Including the URLs of each file with-in it, and a message to display to the user before it is download. This message automatically includes the author, the creation date, and any bibliographic details associated with the dataset. The user can then manually check this is correct and make any needed changes. DataDepGenerators.jl makes it trivial and reliable reuse data created by others.

We create these packages for the relatively new young julia programming language for technical computing. Julia has had rapid uptake in the scientific computing and data science communities. It has a strong culture of automated testing, with the ubiquitous use of automatic continuous integration tests (via Travis CI and App Veyor) on all newly created packages/projects. We build upon, and support, this tradition with the creation data dependency management tools.

By making use of modern software practices including end-to-end automatic testing and deployment, we ensure science can be replicated.