Research Ideas and Outcomes 2: e10708
doi: 10.3897/rio.2.e10708

Data Management Plan (Biosciences)

Data Management Plan for Moore Investigator in Data Driven Discovery Grant

Ethan P White ‡,§

‡ Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, United States of America
§ Informatics Institute, University of Florida, Gainesville, United States of America

Corresponding author: Ethan P White (ethan@weecology.org)

Reviewable v1

Received: 03 Oct 2016 | Published: 04 Oct 2016

Citation: White E (2016) Data Management Plan for Moore Investigator in Data Driven Discovery Grant. Research Ideas and Outcomes 2: e10708. doi: 10.3897/rio.2.e10708

Abstract

Background

This Data Management Plan (DMP) was created for Ethan White's Moore Investigator in Data Driven Discovery award. It describes the management and sharing of all data and code associated with Gordon and Betty Moore Foundation grant GBMF4563 (White 2014). This includes raw data collected as part of the proposal, data compilations, and software. Research associated with this award is related to data-intensive approaches to studying ecological systems and the development of software for automating the cleaning, restructuring, and integration of heterogenous data sources.

New information

Detailed descriptions of data and software management, archiving, and publishing are provided.

Keywords

ecology, data driven discovery, data-intensive research, quantitative ecology

Data Description

The majority of the data involved in this project will be pre-existing data collected by other individuals and organizations. Much of this data is already openly available. In cases where it is not we will work with the relevant stake-holders to make as much of that data available as possible. All pre-existing data that my group has collected have already been made openly available.

We will likely collect some additional data during this project by compiling existing data from the literature and other publicly available sources. This data will typically be tabular data involving information on ecological systems and the traits of individual organisms. It will be stored in standard data formats such as csv and HDF5.

In terms of data products more broadly (i.e., including databases, analyses, and software tools) this project will generate both project specific computational analyses and general use computational tools. The computational tools will be software designed to make working with the diversity dimension of big data easier by automating the cleanup and restructuring of data sets and the combination of multiple data sets for analysis.

Metadata will be provided for all data products. In the case of newly collected data this will include both well written documentation and Ecological Metadata Language files. In the case of software it will include documentation in the source code as well user focused documentation and tutorials. For data provided by other individuals or institutions we will help that individual or institution develop publicly available metadata, even in cases where the data itself is not public.

As described above, much of the actual data used in this project will be owned by other individuals and organizations. Data compilations that we build and software that we develop will be owned by the PI and the University of Florida, but be made publicly available under open source and open access licenses.

Data Management

During development the data, software, and associated metadata that we produce will be stored on GitHub or another Git-based hosting service. It will be automatically backed up nightly to both a lab server and to the University of Florida High Performance Computing Center's Replicated Long-Term Storage. Data and software access and distribution will be managed using Git and the centralized Git host. Following development data and software will be archived in more permanent locations (see Data Sharing).

Data obtained from other individuals and organizations will be stored in PostgreSQL databases on a lab server and backed up nightly to the University of Florida High Performance Computing Center's Replicated Long-Term Storage. Some of this data is proprietary (e.g., the Audubon Society's Christmas Bird Count, eBird, and the North American Butterfly Associations Butterfly Count). Permission to use this data is obtained

through the use of Memorandums of Understanding or other data use agreements. These agreements typically prevent the redistribution of the raw data. There is also data that is publicly available but lacks fully open licenses. In many cases this data cannot be redistributed, but the software we will develop circumvents this issue by automating the downloading and installation of this data. As such, we can provide easy to use copies of assembled databases involving this type of data by distributing open source code that will download and assemble the datasets locally.

All personel on the project will be responsible for entering and maintaining data archives. The data and software archives will be archived indefinitely by using existing archives that are intended to be self-sustaining and are backed by CLOCKSS (see details in Data Sharing).

Newly collected data will be independently entered from by two separate individuals and then compared to identify any errors (i.e., double data entry or two pass verification) and then entered into the main database. Once in the database, the data will be checked again using an automated data validation step with constrained data values to flag errors (e.g., outside of range limits, inconsistent taxonomy, or incorrect data types). Quality control and assurance for software will be accomplished using a combination of code review, unit testing, and integration testing. Code review will be conducted prior to changes being accepted into the core repository and tests will be run automatically using Travis CI or a similar continuous integration service.

Data Sharing

Potential data users for the data we will compile include ecologists, environmental scientists, managers, and conservation planners. Potential users for our software include anyone working with long-tailed data.

Data and software will be developed in the open, with new data and changes to software being made public (typically via GitHub) in near-real-time. Formal releases of raw data and software products will be conducted as these products reach meaningful milestones. At each of these releases the relevant product will be archived in a long-term repository that includes meaningful assurances that the archive will be available even if the repository shuts down (e.g., CLOCKSS). Documentation, tutorials, and metadata for using both data and software will be archived with the relevant product. Examples of data repositories we will use include Dryad, Figshare, Ecological Archives and DataONE. Examples of software repositories include Dryad, Zenodo and Figshare. The original software and data sources will also be maintained on theGit-repository host to facilitate further development. The goal of these combined efforts is to maintain archives of all data and software products indefinitely. By using centralized archives hardware maintainence is not required.

All software will be released under permissive open source licenses (MIT orBSD). All data will be published using the CC0 Public Domain Dedication if allowed by the repository, with CC-BY as a fallback if the repository does not allow CC0. The only exception to the use of

these permissive licenses will be incases where data or code produced by other groups needs to be included that involves non-permissive licensing. If this is the case we may be required to use GPL or CC-BY-SA licenses.

In cases where we compile data from other sources we will cite all of the relevant sources and, where appropriate based on substantial contributions, include original data collectors as authors on relevant data products.

For major releases of major data products we anticipate publishing associated Data/Software papers for referencing and to help advertise the availability ofthe data product as broadly as possible.

Acknowledgements

Thanks to Stephanie Simms and others working at the California Digital Library on the Data Management Plan tool for encouraging me to post this DMP as part of their exemplar collection at RIO Journal. This work is funded by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4563 to Ethan P. White.

Funding program

Gordon and Betty Moore Foundation Data Driven Discovery Initiative

Grant title

Moore Investigator in Data Driven Discovery (GBMF4563)

Hosting institution

University of Florida

Author contributions

Ethan P. White wrote the Data Management Plan.

References

- White E (2014) Data-intensive forecasting and prediction for ecological systems (Moore Foundation Data-Driven Discovery Investigators Proposal). Figshare 1 DOI: 10.6084/M9.FIGSHARE.1189330