NSF-EAR Postdoc Fellowship

A Data Management Plan created using DMPTool

Creator: Sheila Saia

Affiliation: Cornell University

Template: National Science Foundation (NSF)

Last modified: 01-04-2017

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

NSF-EAR Postdoc Fellowship

Types of data

My proposed research will generate continuously monitored data from probes installed in each of the two experimental wetlands cells. Data loggers will record probe measurements every 15 minutes for redox potential, pH, temperature, inflow and outflow water height of each wetland cell, groundwater height, as well as wetland and groundwater salinity. These data will be saved in the .csv and .txt file formats. In addition to continuous data, I will also generate data from grab samples (i.e., 4 grab samples per experiment for each of the two wetland treatments for a minimum of 3 experiments). For each grab sample, I will collect inflow and outflow water from each wetland cell that will be analyzed for dissolved phosphorus (DP) and total phosphorus (TP) concentration, d18O value, and salinity. I will collect wetland soil cores and will analyze these for the following data variables: core depth, DP and TP concentration for all Hedley and oxalate fractions, d18O value for Hedley (i.e., resin P, microbial P, HCl P) and oxalate fraction, percent organic matter, soil pH, enzyme activities for select enzymes outlined in the proposal, and 31P-NMR spectra and associated P concentrations. I will analyze the fertilizer added to each experiment as a tracer for the following data variables: DP and TP concentration, d18O value, and pH. These data will be saved in the .csv and .txt file formats. I will generate a syllabus and lesson plans for the data management mini-course, which will be saved in the .doc format.

The Duke University Wetland Center (DUWC) has already collected groundwater height and TP concentration measurements for the proposed wetlands cells. I will obtain these data from my mentor, Dr. Curtis Richardson, and plan to merge with the relational database discussed below. To ensure data consistency in these data over time, I plan to measure groundwater height and TP concentration using the same approach as the DUWC.

I will process all .csv and .txt data in the R and Python scripting languages. Specifically, I plan to create a relational database, indexed by experiment, that can be queried using standard R packages. I will double check hand-entered variables (e.g. soil pH) upon data entry and will use R to carry out quality assurance/quality control on all data before further data analysis.

Data and metadata standards

The metadata required to interpret the data I will collect include identification of variable units, time, sample depth, location, treatment type, description of the analysis method used, and any additional notes pertaining to grab sample collection. For continuous data (i.e., probes), I will adhere to the CUAHSI metadata and data standards for time series data also known as Water metadata language (WaterML). WaterML is considered the standard for time series data related to water research. For grab sample data, I will provide a ReadMe file for each data type using a recommended template such as the one provided by Cornell University at https://data.research.cornell.edu/content/readme. As there is no standard metadata for syllabi and lesson plans, I also provide a ReadMe file as described previously; this file will also include any notes about the effectiveness of teaching strategies I used as well as suggestions to improve the class in the future.

Policies for access and sharing

Wetland-related data will be made available publically either upon publication in a peer-reviewed journal or within two years after the end of the experiment as outlined in the NSF-EAR guidelines; whichever comes first. Upon publication, data will be made available via a GitHub repository and associated DOI issued through Zenodo. Data, ReadMe files, data processing code, and publications can be easily downloaded from GitHub. The foreseeable data users of these data include other scientists, Durham City officials, and Duke University Facilities staff. Using the data will require R and Python software, which are both open-source programs that can be downloaded for free.

Prior to becoming publically available on GitHub, data will be saved and managed on my personal PC. I will continuously work to manage these data over the course of this project as indicated in my project timeline. If interested parties would like to request these data before it is published on GitHub, these requests can be made by email through me (sms493@cornell.edu), Dr. Curtis Richardson (curtr@duke.edu), or Dr. Christian von Sperber (csperber@uni-bonn.de). I will keep two additional copies of all data on an external hard-drive and on Duke Box. Duke Box is similar to DropBox but offers unlimited data storage and keeps all versions of uploaded data for researchers affiliated with Duke University.

Data management mini-course related data will be publically accessible as soon as it is posted on a GitHub-powered course website. I will build this site with the help of my collaborators at the Duke University Data and Visualization Services Department so students can access and interact with course materials. Therefore, proposed users of these data include students taking the course, course instructors, as well as future students and instructors. I will also keep two back-up copies of these data on an external hard-drive and DukeBox.

Policies and provisions for re-use, re-distribution

All data will be posted publically according to the timing outlined in the previous section of this data management plan. Until that time, my mentors (Dr. Curtis Richardson and Dr. Christian von Sperber) and I retain the right to use the data to prepare peer-reviewed journal articles. Once posted publically, interested parties may use these data according to the Creative Commons 4.0 (CC-BY) license, which allows "others [to] distribute, remix, tweak, and build upon [our] work, even commercially, as long as they credit [us] for the original creation". I am motivated to use this license because I believe it is important that data relating to climate change can be easily accessed by other scientists. We expect no embargo periods outside the timing discussed previously. We also do not expect to have any IRB obligations given the nature of the proposed research; my proposed research does not involve human subjects or animals.

Plans for archiving and preservation of access

Wetland-related data including the relational database files, WaterML files, ReadMe files, data analysis scripts, and publication proofs will be publically available in a data repository on GitHub through a DOI generated by Zenodo. Zenodo is internationally recognized and will ensure long-term data storage (20 years). Given that collaborators on this project are USGS members (Dr. Megan Young and Dr. Carol Kendall), long-term data storage is mandated by the USGS. Continuous wetland data will also be posted on the CUAHSI hydrologic information systems (HIS) HydroClient data repository so other scientists can search for it. Because all data will be saved in non-proprietary formats (i.e., .csv and .txt), no transformations should be needed for long-term data storage. Data management mini-course related data will be stored long-term on a GitHub website.