Data Management Policies for the Wolkovich Lab

December 1, 2024

Contents

1 Overview **1**

2 Does this apply to me? **2**

3 Types of data **2**
3.1 Raw data . . . 2
3.2 Derived data . . . 2
3.3 Other types of data & things that are borderline data . . . 2
3.4 Received data . . . 3
3.5 Code . . . 3

4 Data organization and formats **3**
4.1 File formats . . . 3
4.2 Data formatting . . . 4
4.3 Metadata . . . 4
4.4 Cleaning data . . . 4

5 Where to store data **5**
5.1 Data storage in the lab . . . 5
5.2 Preferred repository for data archiving and publication . . . 5
5.3 Beware data provenance issues . . . 5

6 Co-authorship & data **6**

7 Redistribution policies & timelines **6**
7.1 Length of time after collection to be made available to rest of lab . . . 6
7.2 Use of lab data . . . 7
7.3 Length of time after collection for data to be made public . . . 7
7.4 Data licenses . . . 7

8 UBC Policies **7**

9 Harvard Policies **7**

1 Overview

The Wolkovich lab Data Management Policies are designed to make data produced by the lab useful both to those who collected the data and to those who may also benefit from its use over

the short and long-term. The main goal is to produce robust, re-usable data with adequate additional information and storage practices that give the data a long lifespan. To that end, this document details how data are managed within the Wolkovich lab, including: how and where data should be stored, how data are shared within the lab, and how data are managed beyond the lab, including public archiving of data. In addition to helping each member of the lab keep their data organized and preserved for their own research benefits the policies also serve researchers in the lab to help meet data management requirements of funding agencies.

2 Does this apply to me?

This document applies to data produced by anyone who is appointed in the lab under the supervision of Lizzie Wolkovich (henceforth 'Lizzie'), including undergraduates, graduate students, postdoctoral fellows and associates and any visiting students or research fellows. It applies regardless of funding source or whether you are co-supervised or have many co-authors. If you want to deviate from this plan you must discuss this with Lizzie within 6 months of starting the project that you expect will deviate from these policies. Any deviations that both you and Lizzie agree upon must be written down by you and sent to Lizzie in a format that includes your name (signature on paper or emailed from you) and the date. While lab members are encouraged to apply these policies to all their data, these policies do not apply to data you produced before arriving in the lab or to projects fully outside of the lab (discuss with Lizzie if you are unsure).

3 Types of data

These policies cover raw and derived data and code.

3.1 Raw data

Raw data (also known as primary or source data) are those collected or produced by the lab that have not been processed to averages or in other ways that would make them derived. Examples are: counts of one species in a plot, total C or N of plant tissue, biomass weights, leaf shapes. Raw data are generally numerical or character data written on paper or electronically (raw data do not include things like plant leaves you need to process to turn into data).

3.2 Derived data

Derived data are the output of raw data after some sort of processing (often statistical). Examples are: mean first flowering date, slope values from a regression for different species, standard errors.

3.3 Other types of data & things that are borderline data

Various other types of data exist, such as data produced by model simulations. This plan is not designed to cover all other types of data so when beginning a project with a novel data type the researcher should meet with Lizzie to discuss best practices for organization, storage and preservation of the data.

Additionally some things are not exactly data, or at least are not the type of data that fall under this plan. Such things include inventory lists as projects develop or field notes while setting up projects. For example, when I started a project back in 2003 I visited about 100 or so shrubs and took notes on each, only 56 shrubs ended up in the experiment so that list of 100 or so may be data, but not the type that fall under this plan. If you have any queries, ask Lizzie.

3.4 Received data

The Wolkovich lab often uses data obtained from somewhere/someone else (which I call 'received data' for now), such as from a colleague or from a national repository. Before any data enter the lab, the lab researcher(s) and data provider must agree on how data will be used and published. In general, lab researcher(s) should try to follow all the lab policy guidelines including full publication of the raw data, however this is not a requirement. What is required?

1. Data must be properly cited and acknowledged including acknowledging grant funding numbers if applicable.
2. Data must conform to any data standards or policies set out by the original data user (e.g., any policies by the US-NPN for use of their data must be followed).
3. Once any such data enters the lab it must follow the organization and internal storage guidelines.
4. Derived data from the received data must follow all of this document's policies.
5. See also 'Co-authorship & data' below.

3.5 Code

Code! If you are in the lab hopefully you love to code or will soon. All code used for publication or to clean/manage data falls under this document. Lots of code when you are playing around or such does not, but such code probably leads to code used for publication or to clean/manage data so consider that it may eventually fall under this document.

4 Data organization and formats

4.1 File formats

Your data should be kept in tabular form with accompanying metadata *from the day it is digital.* For some data this will be *the* day you collect. For other data, it may start on paper. Data collected on paper must be collected with minimal metadata immediately (reference to project, who collected it, date of collection, and where collected), then scanned (or take a very sharp photo) within *one week of collection* for field-collected data, though ideally scan it the same day; for data collected on paper in the lab, you must scan it *the day of collection.* All paper data must be transferred to electronic format within one month of the end of field season—though not to exceed 4 months after data collection. Acceptable formats for this are:

1. Comma separated (`csv`) or tab-delimited files for the tabular data and plain text (`txt`) for the metadata. The two files must be linked by common names and the tabular filenames must be listed near the top of the metadata file. One `txt` file for the metadata may suffice for multiple `csv` files if properly documented.
2. An Excel notebook (`xls`or `xlsx`) with tabs for the data and a metadata tab. If using this format though you *must* submit final files to a data repository in a nonproprietary form (see option 1) and I warn you bad things happen to data in Excel (especially to dates) so it's not recommended.
3. A Google notebook with tabs for the data and a metadata tab. If using this format though you *must* submit final files to a data repository in a nonproprietary form (see option 1).

4.2 Data formatting

To repeat, your data should be kept in tabular form with accompanying metadata *from the day you collect it.*

In general, try to keep your data in the 'long' form and keep items in as many separated columns as needed (when in doubt use more columns). This means use more rows even if it means duplicating information in some, but not all columns. For example, if you are listing all the plant species you counted in a plot you might have the following columns: `date`, `site`, `observer`, `plot`, `genus`, `species`, `varietyetc`. Under this format you would repetitively enter rows of `date`, `site`, `observer`, `plot` for each species. Also note how many columns you have: generally keep track of date, observer, and keep species names broken down into different columns—it's quicker code to put them together later than to break them apart.

In general use lowercase letters when possible and allowed and avoid spaces and special characters (accents, exclamation points).

Dates should be day-month-year (not month-day-year). To avoid confusion try to use a two-digit day, three-letter month, and four-digit year (e.g., `04-Jul-1786`).

4.3 Metadata

All your files should have metadata. Metadata is info about data. *Start your metadata as soon as you start any data file.* As you start it should include the date you started the file, the work, who is writing the metadata and then all the relevant information including:

- Why were the data collected? What project are the data for?
- Who collected the data? If applicable, who entered the data? *Record who collected each piece of data.*
- When were the data collected?
- Where were the data collected? (Reserve name, city, state, country and GPS location if possible)
- What does each column mean? In what units? What are the sites or plots or such (if mentioned)?
- What authority do the species names come from (if applicable)?
- What changes have been made to the data and by whom? (Though try to avoid this if possible, see below 4.4.)
- Any other relevant information.

You should record who wrote each bit of the metadata and when. In general you should aim to have all the information needed to follow the Ecological Metadata Language (EML), which you can learn more about here: `https://knb.ecoinformatics.org/#external//emlparser/docs/index.html`, but be forewarned, this is a lot of information.

4.4 Cleaning data

Data often need cleaning. For example, species names are mis-spelled or commas are misplaced in one cell of a large spreadsheet etc.. If possible all data should be cleaned by a coded language (such as `R`), this way it is possible to track all cleaning. Even one cell changes are best done in a code file, that ideally is dedicated to cleaning your data. If not possible keep specific notes on your cleaning, including who did it, when and what the change was *from* and *to*. For example, 'on 05-Sep-2015 Jane H. Smith changed the value for g of N (grams of nitrogen) for site X, plot 2, sampled on 03-Jul-2014 from 60 to 0.60.'

5 Where to store data

5.1 Data storage in the lab

Data in the lab may be collected on paper first or immediately into a computer. You should note in your metadata the initial mode of collection and store all paper datasheets in the appropriate place in the lab. Following the lab's backup storage policy you should back up all lab related research files daily or weekly.

All your data should be saved to the Wolkovich Lab server (midge) backup section (that's `/home/data/backup`, that's usually `cd ..//data/backup` from where you start, for example, I start in `/home/lizzie`), which is regularly backed-up on a BackBlaze account associated with the temporalecology lab email. If you choose to leave data on your personal machine you *must have those data on two backup systems (e.g., your own daily/weekly backup and GitHub).*[1] Please update files instead of keeping versions of them with different names. Ideally data on midge will also be kept in a `csv` or `txt` form on GitHub or another versioning repository. Code can be kept on Github only if this is clearly documented. For most projects, you will need to put data on midge once after collecting it and from there can modify versions of it in GitHub.

You can and should delete data files off the lab server once they are publicly archived. Some files that are too large or otherwise not possible to archive (e.g., image files or files shared by growers) can stay on the server indefinitely, check with Lizzie.

For each of your projects include a Markdown (.md) file in the `datacodemgmt/wheredatacodelive` folder on the labgit repo (until 2023, this was all on Google drive). This file should be named after your project, include your name as author of the file, the names of everyone who worked on the project, and the time period of the project. It should list each raw (not derived!) datafile and where it is located as well as the names of all major code files and where they are located. Make this very clear so anyone can find the associated data and code. You must check this file and update it as needed as every 6 months. A template is provided in the `datacodemgmt/wheredatacodelive/ZeTemplate` folder called `ZeTemplate.md` to help. Be sure to update the wheredatacodelive file once data are publicly archived and discuss with Lizzie making sure all files are properly archived.

5.2 Preferred repository for data archiving and publication

All lab members who use data in the lab agree to make the data public (see 7 below) and are responsible for publication of the data. The recommended repository is the Knowledge Network for Biocomplexity (KNB), because it includes excellent metadata standards and is a central repository for ecological data. Other repositories for data are allowed if approved by Lizzie and given that they are part of DataOne, this includes Dryad. Code related to data should ideally be published with the data on KNB or other repositories. It can also be published on GitHub, if clearly linked and noted from the data publication site.

5.3 Beware data provenance issues

Provenance of your data includes from where it *originally* comes. Provenance issues crop up when data show up multiple places and it is unclear whether you are looking at the same data or different data or ... who knows. Avoid creating such issues by publishing your data *once* in *one place*. So, if your data must be published on the Harvard Forest LTER site *do not* re-publish them on KNB or other sites. Avoid committing to duplicate publication and make sure you know your funders' and field sites' requirements for publication to avoid re-publication. If

[1] GitHub is not a backup service so it should only ever (ever!) be used as such in addition to a real backup service such as Backblaze or Time Machine used with multiple hard drives (see databackuplab.txt in the labgit).

needed ask sites to point to the one place you publish the data (for example, publish the data on KNB and ask journals or other repositories to point to the data on KNB).

6 Co-authorship & data

Authorship requires substantial intellectual contribution, which includes the reading and editing of all manuscripts on which you are a co-author through the submission-for-publication stage (when you submit a paper you must say that 'all co-authors approve of the submitted version,' thus if you are a co-author you must do this). In general, if you have helped *design* collection of data then you have made an intellectual contribution that is often expected within the lab to lead to co-authorship on an initial resulting manuscript, but co-authorship will be handled on a manuscript-by-manuscript basis. In general, the Wolkovich lab does not give co-authorship for collection or donation (aka sharing or loaning or giving) of data.

First authors in the lab must talk about co-authorship expectations and agreements early and often with all possible and existing co-authors. If you are a potential co-author and feel a first author is not meeting this expectation, bring it up to Lizzie.

If you agree to take on existing data you cannot offer co-authorship for use of the data unless four criteria are met:

1. The co-author agrees to (and does) make substantial intellectual contribution to the work, which includes the reading and editing of all manuscripts on which you are a co-author through the submission-for-publication stage. This includes helping with interpretation of the data, system, study questions.

2. Agreement of co-authorship is made at the start of the project.

3. Agreement is approved of by Lizzie.

4. All data-sharers are given an equal opportunity at authorship. It is not allowed to offer or give authorship to one data-sharer unless all other data-sharers are offered an equal opportunity at authorship—this *includes* data that are publicly-available, meaning if you offer authorship to one data-sharer and were planning to use publicly-available data you must reach out to the owner of the publicly-available data and strongly offer equivalent authorship as offered to the other data-sharer. As an example, if five people share data freely with you for a meta-analysis and and a sixth wants authorship you either must strongly offer equivalent authorship to all five or deny authorship to the sixth person. Note that the above requirements must also be met in this situation. If one or more datasets are more central or critical to a paper to warrant selective authorship this must be discussed and approved by Lizzie (and has not, to date, occurred within the lab).

7 Redistribution policies & timelines

The Wolkovich lab is committed to data sharing and open data access. This means all your data related to the lab must be made public, whether you publish work from the data or not.

7.1 Length of time after collection to be made available to rest of lab

Data should be available for anyone in the lab to look at immediately upon completion of the experiment or field sampling and you should be happy to show anyone the data at any point (failure to do this will freak out Lizzie and not be good for you). 'Experiment' or 'field sampling' are considered to be events lasting less than one year in general and include: a growth chamber study, a summer field season (e.g., May-September in Quebec or March-October for winegrapes

in California). Please note your electronic data should be properly stored at all times in the lab (see 5.1).

Data you collected and planned to publish are available to others in the lab for use at the time of publication of a related manuscript or **one year** after completion of the project if no manuscript is currently submitted, whichever comes first. What timeframe delimits a 'project' should be defined upfront with Lizzie, otherwise it will be assumed to fall under an 'experiment' or 'field sampling' as mentioned above. So, for example, if you plan to sample a field site for three years and will then analyze the resulting data you must have this timeline approved by Lizzie before the start of the first field season. Deviations from the one year requirement may be approved if discussed early and openly, and depending on circumstances.

Unless otherwise agreed, Lizzie can use data collected in the lab immediately for grants; she must inform others using the data when doing this.

7.2 Use of lab data

Sharing and use of lab data are encouraged, given that you are open about your plans to use the data. If you plan to use lab data that do not fall under your project you must openly discuss this with Lizzie and whoever else is using the data before starting any work and you must be open about your use of the data during your work. Lizzie has the right to ask you to defer from using the data if your use may conflict with already-underway-use by others in the lab.

7.3 Length of time after collection for data to be made public

Data you use related to the lab should be made public: immediately at the time of publication, or **two years** after completion of the project—whichever comes first. Please review the definition of project above.

7.4 Data licenses

Data from the Wolkovich lab should generally be made public under a Creative Commons license. In general the CC0 is recommended, this license clarifies that the data are in the public domain, and one relies on scientific norms for attribution. An alternative option is the CC-BY, in which the data do not go into the public domain. An overview of the issues is provided by the JISC (click here).

8 UBC Policies

Though outside of the reach of this plan, remember that all data containing student information is personal information and must not be shared or preserved on a non-secure system (or even a secure system that operates outside of Canada). More details here.

9 Harvard Policies

Here are ye Harvard Policies and how data in the Wolkovich lab fits into ye policies. All data in the Wolkovich lab fall under these policies. In addition, you may need this verbaige for official grants. Enjoy!

Harvard University has developed and adopted a Research Data Security Policy requiring appropriate technical and process protections for confidential research data: `http://www.security.harvard.edu/research-data-security-policy`.

Under the policy, research data is classified into levels to indicate the sensitivity of the data. The levels range from Level 1 for "De-identified research information about people and other non-confidential research information" to Level 5 for "Extremely sensitive information about individually identifiable people."

Data at each level is subject to a set of detailed technical and administrative process protections that are appropriate to protect the security, privacy and confidentiality of the particular type of data. This policy has been formally adopted by the University and applies to all research involving confidential data whether funded or not.

Harvard has a university-wide policy addressing the required protections for intellectual property arising from research. These policies can be found here: `http://otd.harvard.edu/resources/policies/`.

Data will not be encumbered with intellectual property rights (including copyright, database rights, license restrictions, trade secret, patent or trademark) by any party (including the investigators, investigators' institutions, and data providers); nor is it subject to any additional legal requirements.

All data in the Wolkovich lab are considered as non-confidential research information and openly-available upon request (if requested before public archiving).