McEwan Lab Best Practice

Data management and sharing

Updated October 2017

McEwanlab.org

1. Data collection

Collection:

(a) If you are leading a project you must keep a field notebook in which you outline your observations and activities. These are available in the McEwan lab.

(b) Field data collection needs to be clean, legible to others, ***in pencil***, on waterproof paper, with dates and data collector information on each sheet. Data collected on paper in this fashion is timeless and *superior to digital methods*.

(c) When possible there should be uniformity in data collection system. Preferably have the same person collect the same data in the same way for the whole project and deviate from that only if needed and with caution.

(d) QA/QC your data entry. Errors are quite common in data collection and data entry. You need to practice regular QA/QC. Best is a quick scan of each sheet in the field (or lab) as the sheet is completed to see if anything absurd is written down (oak tree with 8900 cm DBH). During data entry is a good time to catch mistakes. Then after data entry, depending on the number of sheets, it is useful to go back and re-check the entry. Minimally, a random 10% sample of the data sheets and re-check with the entered data is needed. If mistakes are found, then another random 10% sample (without replacement) is needed.

2. Precautions against the loss of data

Computers are only vaguely reliable instruments.

You are required to assume that your computer, or the one you are using, will fail unexpectedly.

Examples of **unacceptable** reasons for losing data:

(a) my hard drive crashed! (of course it did!)

(b) I was really busy getting ready for "x" so I didn’t back up the data

(c) there is only one computer with a microscope camera/software package/etc and so I saved everything there and didn't back it up.

3. McEwan lab data management practices & data sharing

Here are some rules for working in the McEwan lab:

(a) Data ownership.

At the time of graduation, or the ending of any particular project, all data and other information associated with the project must be transferred to Dr. McEwan for curation, storage, publication or sharing.

Students may be entitled to authorship on subsequent publications; however, that is determined on an individual basis depending on the efforts of the students within the context of the overall project.

As Principle Investigator in the lab, and the person responsible to funding agencies, etc, Dr. McEwan is the ultimate owner of all data collected in the lab and reserves the right to make datasets publicly available and move forward with publishing or other uses pursuant to the code of ethics surrounding scientific information put forward by the Ecological Society of America.

(b) Everyone will use Excel to enter and manipulate data and create CSV files and all analysis must be done in R.

(c) We will share data sets across the lab

(d) We will create data products (data sets) that are shared publicly

(e) Everyone will help out with analysis- share in a collaborative fashion

(f) Analysis of data sets will often be open to peers, and, in some instances, will take place live in front of the lab group. This might feel scary at times if it is your data set, but this is the way we are moving forward

Process:

(a) Every new project in the lab should begin with the creation of a folder that is shared. Dr. McEwan will create that folder and share it with the participants in the project. If you are involved in a project and do not have a folder, let Dr. McEwan know via email or in person and he will create the folder. All activities related to the project must take place within this shared folder. The shared folder should have a logical sub-folder structure and can include "Literature" "Analysis," "Writing," "Scripts," etc.

(b) Data products. The first priority for any project, once data have been collected is to create data products for the project they are working on. These data products will be shared in the

Project Folder and consist of a written description of the methods, data entered into the appropriate format for sharing and analysis in R, and explanatory text in the form of meta-data.

Features of a required data product

- a journal-quality description of methods. This may need to include images, etc.
- a Final CSV folder that contains the perfectly cleaned and organized R-ready files
- metadata for the CSV files

(c) Projects will advance forward, following the completion of the data products, following normal processes including exploratory data analysis, final analyses and writing. This could include writing a Thesis, a paper for publication, or preparing various presentations including posters and presentations.

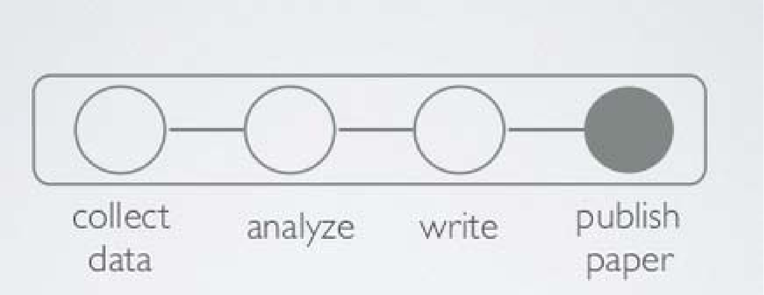
4. Data sharing outside the lab.

(a) Concept

Below is an image describing the way things have traditionally worked in science- data are collected, analyzed and then a paper is written. The paper including the outcomes of data analysis is how the data are shared with the public:

(note images are borrowed from a Chris Lortie Ignite talk)

Traditional Workflow



But there are a lot of people now who think this model stinks.

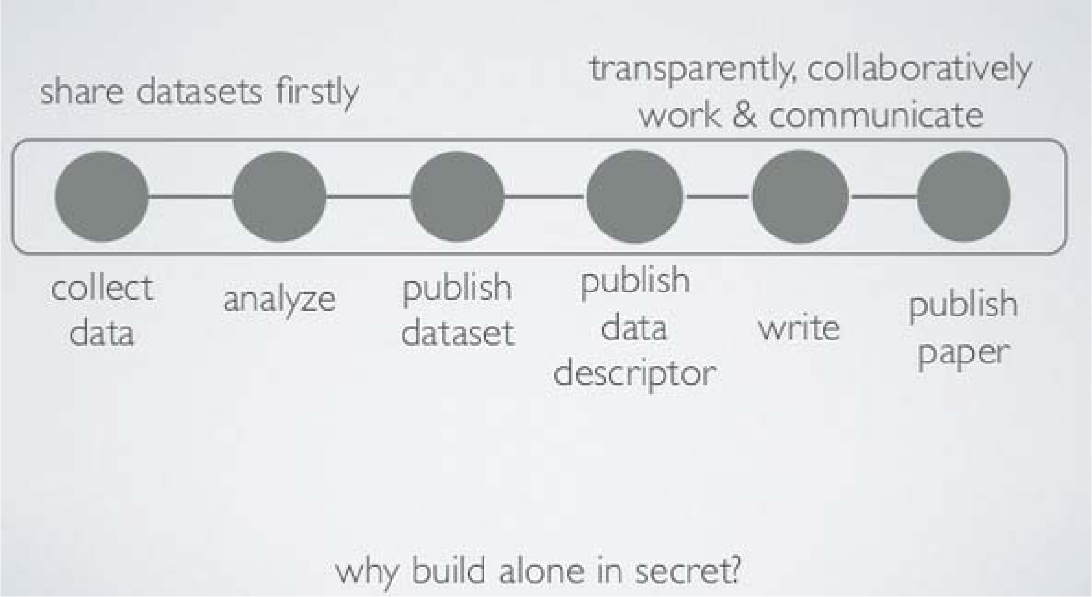
It is not transparent,
it hinders progress,
it makes the public wary of science,
data get lost
etc.

From a practical perspective I can tell you that this model is no good for a PI of a lab because, unless a good plan is in place and things are followed up on carefully, data can get lost.

Example: A student graduates, the PI is too busy to pay attention, the student gets a job, 13 months pass, then the PI is "where are the data, we need them as a baseline for a new study!"...the student via email from Colorado does not reply for 7 weeks and then writes "the data are on the laptop in the lab!" But the laptop in the lab is gone, it crapped out a while back...so no data...

Here is what many believe is the future for intelligent and transparent science:

Modernized Workflow



share datasets firstly

transparently, collaboratively work & communicate

collect data

analyze

publish dataset

publish data descriptor

write

publish paper

why build alone in secret?

The main point is that the raw data set is published, with a citable DOI before or ***<u>at the same time as</u>*** the paper. Then the data are forever available, not only to people in the lab but to anyone!

This is called "Open Science."

This is not a new idea...in fact, it is similar to the International Tree-ring Data Bank which has been around since the 1980s:

http://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring

And Vegbank:

http://vegbank.org/vegbank/index.jsp

And others.

You can get data from these repositories and analyze them yourself for whatever reason you might have. You simply cite the data set in the article!

You can write your own publication using these data...or sweep them into an analysis you are doing. Below is a figure from a publication of mine which is nearly all data that I collected but I included this piece (highlighted in red) of data that was collected by someone I never met before and posted in the tree-ring data bank.

R. W. McEwan and B. C. McCarthy

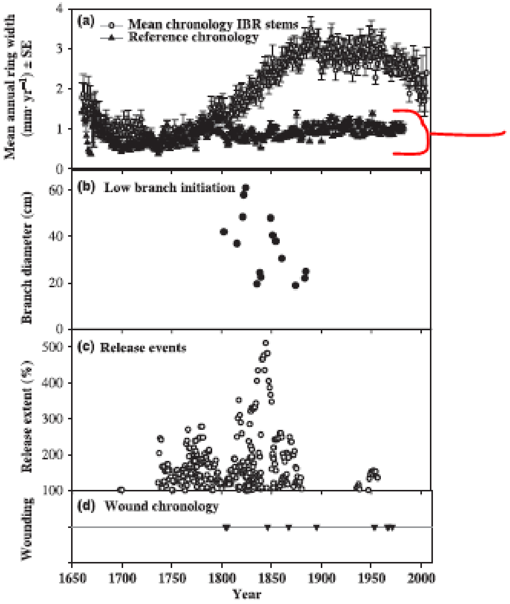


Figure 6 Composite growth chronology and ecological features of remnant savanna trees in the Inner Bluegrass Region (IBR) of Kentucky, USA. (a) Line and scatter plots representing mean annual ring width (± 1 SE) for all IBR samples (○) and *Quercus alba* from an old-growth forest in eastern Kentucky (▲). (b) Pith dates and diameters of low branches collected from remnant IBR savanna trees. (c) All growth releases from all IBR samples. (d) All wounds recorded on all IBR samples are plotted as downward-pointing triangles on the same line.

This idea seems to be gaining momentum. Many opportunities exist for sharing data broadly. For instance, a group that originates from the [national center for ecological analysis and synthesis]() is running a system called **KNB** that promises to create a repository and network for data from a wide array of investigators:

https://knb.ecoinformatics.org/

The Ecological Society of America have now created an alternative route for publishing which is called "Ecological Archives"

This is a place where you can submit a "data paper" that is basically a description along with a data set, (or sets). They sometimes publish raw data from papers published in an ESA journal but they also consider stand-alone data papers...

http://esapubs.org/archive/

These data papers are peer reviewed, and basically represent raw data along with descriptions of the data.

There are many other avenues in all fields.

(b) McEwan Lab Data Archive

Dr. McEwan is in the process of creating a data archive through the University of Dayton library. The idea is to store curated data set on the site in such a way that they are publicly available and have a citable DOI. This means that we can site the data sets and others can use them. It also means that McEwan lab folks will always have access to the data sets. This is a permanent archive. Dr. McEwan will contact you about storing your data sets in the archive if appropriate.

LINK: http://demo.udayton.bepress.com/mcewan/