

# Data management and archival with APECS

*Tiff Stephens*

*10/24/2018*

Great work this summer in collecting mountains of data! I now want to make sure that we are all on the same page with how these data (and metadata) are handled for both analysis and long-term storage. I'm coordinating data management and the associated best practices to ensure that we meet the goals of our management plan but each of you will help APECS reach such goals via assisting me in the process with respect to the data that each of you have collected and are curating. You should be aware that both grants under the APECS umbrella have a data management plan. They are not complex but serve as a promise to ourselves and to NSF that we will be responsible with the expensive data that we were funded to collect. The plans outline:

- That we will be compliant with NSF data management and dissemination policies (essentially: don't lose data, keep it clean, and make it available to others).
- That all data is secure during the project (e.g. backed-up).
- That all datasets will be accompanied by appropriate metadata.
- That all data will be submitted to appropriate data repositories no later than two years after being collected. Eventually all should be publicly available.
- That we will submit data in an accessible and flexible format (i.e. .csv files).

---

## A quick summary of these guidelines

- Make sure that your entered data are accurate and that data are stored in a “clean” format.
- Have a Masterfile for each dataset that will not be manipulated.
- Make sure that each dataset has appropriate metadata.
- Store datasets in locations accessible to all APECS members (i.e. shared Google Drive and GitHub)
- Work with Tiff to archive data with selected repositories (i.e. KNB and BCO-DMO)

Onwards! Best practices for data management with APECS:

---

## 1. Get those data and build, clean, and describe a dataset

### Enter data

After fieldwork, labwork, etc., collate the raw data into a spreadsheet. Be mindful, however, how those data are entered/organized into a working dataset. A spiffy dataset is a “clean” dataset (see below). It saves *a lot* of time to enter it in such a way that the data (1) are readily accessible to whatever statistical software you plan to use and (2) readily accessible to future users. For archival, the data eventually need to be structured into a format that happens to align with structure appropriate for R.

*What is a clean dataset?* Essentially, each variable (dependent and independent) should have its own column, where the rows in each column includes a single point of information. Columns can have multiple factors included to help characterize a single datapoint – for example a “sea\_otter\_region” column can have “high”, “mid”, “low” factors within them. Further examples/guidelines are described below, paired with the following image:

**A.** Column titles and text in rows: Some packages, like R, do not like spaces in the titles for columns/variable, instead use an underscore or period between words/characters. Typically, spaces in rows/factors is just fine.

**B.** It's easiest to keep lats and longs in decimal degrees.

**C.** For dates, please annotate what format your data are in, especially in the metadata but it is also useful for column titles. e.g. “MM.DD.YY”, “YYYY.MM.DD”. Date data are notorious troublemakers in both analysis and archival, especially for international access.

**D.** These columns are examples of how multiple factors can be included in one column, notice how the information is nested...i.e. that each replicate is nested within its respective transect, nested within the respective site. Depending on your preference, these different factors can occupy separate columns (like shown with the density data in columns K and L).

**E.** Note that if there are site-level data, such needs to be filled in for EVERY unique datapoint; its OK if it looks repetitive.

**F.** Two issues here: (1) running calculations in Excel and (2) saving files with calculation sections (see bottom of columns). Unless you’re running data analysis in Excel (please don’t do this), there should be no calculations saved in the spreadsheet (these won’t be saved in a .csv, anyway). If you’re using R for analysis, please avoid calculating anything in Excel, even simple sums or means...mistakes will happen. In this example, it is better to leave the two density columns as raw and calculate their mean using code in R (this is reproducible and more reliable). The second issue will be misinterpreted by your statistical software; whatever platform you use incorporates information from the entire column called upon and thus would include “Averages”, “Nossuk”, “N Pass”, and “Shinaku” as factors in the sediment columns, and the associated mean values as new datapoints in the pits column. Just don’t do these things.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

site_name	longitude	latitude	date_MM.DD.YY	so_region	transect	replicate	sed_primary	sed_secondary	n_pits	n_otters_1	n_otters_2	avg_otters
Nossuk Bay	-133.358474	55.714689	05/19/18	high	outside	1	sand	gravel	10	63	51	=AVERAGE(K2:L2)
Nossuk Bay	-133.358474	55.714689	05/19/18	high	outside	2	sand	gravel	12	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	outside	3	sand	cobble	17	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	edge	1	sand	coarse sand	8	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	edge	2	muddy sand	coarse sand	4	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	edge	3	sand	coarse sand	7	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	inside	1	muddy sand	muddy sand	1	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	inside	2	muddy sand	muddy sand	1	63	51	57
Nossuk Bay	-133.358474	55.714689	05/19/18	high	inside	3	muddy sand	muddy sand	0	63	51	57
North Pass	-132.910143	55.228252	06/17/18	low	outside	1	sandy mud	sand	3	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	outside	2	sandy mud	sand	1	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	outside	3	sandy mud	sand	3	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	edge	1	sandy mud	sandy mud	2	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	edge	2	mud	sandy mud	0	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	edge	3	mud	sandy mud	3	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	inside	1	mud	mud	0	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	inside	2	mud	mud	1	1	0	0.5
North Pass	-132.910143	55.228252	06/17/18	low	inside	3	mud	mud	0	1	0	0.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	outside	1	mud	mud	9	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	outside	2	mud	mud	11	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	outside	3	mud	mud	4	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	edge	1	mud	mud	4	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	edge	2	mud	mud	3	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	edge	3	mud	mud	7	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	inside	1	mud	mud	0	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	inside	2	mud	mud	2	109	122	115.5
Shinaku Inlet	-133.157644	55.599299	05/30/18	mid	inside	3	mud	mud	1	109	122	115.5
Averages	Nossuk								6.66666667			
	N Pass								1.444444444			
	Shinaku								4.555555556			

No!

Don't do!

Neat. Please see Tiff or Wendel if you have any questions about this.

## Don't forget to QC/QA.

It's necessary to check ALL of the entered data with assure that it is accurate... hopefully it goes faster than the initial entry. Even after verifying everything, I like to go through the dataset in R to see if there are any odd entries or values that don't make sense before fully accepting that the QC/QA process is complete. There are a lot of ways to check data but I at least like to make sure that functions like “summary()” work without errors and manually sort each column to check for errors (e.g. odd placement of “NA” or accidental character included, like an “.” instead of “0” or “NA”).

## Metadata!

It is also critical to build metadata for each datasheet, for both APECS and for data archival. What do I mean by meta data? Two things:

- **Short, direct definitions:** This often involves transposing the column titles that you have and providing an informative definition/description of each factor in your datasheet (example below). This can be saved as a separate tab in an Excel document but will ultimately need to be a separate .csv file to pair with your data for archival.

Factor	Description
site_name	Identifying name of the field site
longitude	Longitude (E) of the site in decimal degrees
latitude	Latitude (N) of the site in decimal degrees
date_MM.DD.YY	Date of sampling expressed as month, day, and year (MM/DD/YY)
so_region	Regional categorization of sea otter presence; 3 levels: “high”, “mid”, “low”
transect	Transect location within each site relative to the seagrass bed
replicate	Replicate number within each transect
sed_primary	Qualitative characterization of the primary sediment type (e.g. sand, mud)
sed_secondary	Qualitative characterization of the secondary sediment type
n_pits	Number of pits counted within the transect replicate
n_otters1	Number of sea otters counted in the first boat survey
n_otters2	Number of sea otters counted in the second boat survey

- **Brief summaries:** This should include information about the purpose of the dataset and basic summaries of where and how it was collected. Think “briefer version of the methods section”. This is an extremely valuable description for people that will access the publicly available data years down the road, and can be the clincher for whether your data are usable/interpretable by future parties. This is a requirement for submitting to data repositories, a real-life KNB example shown below:

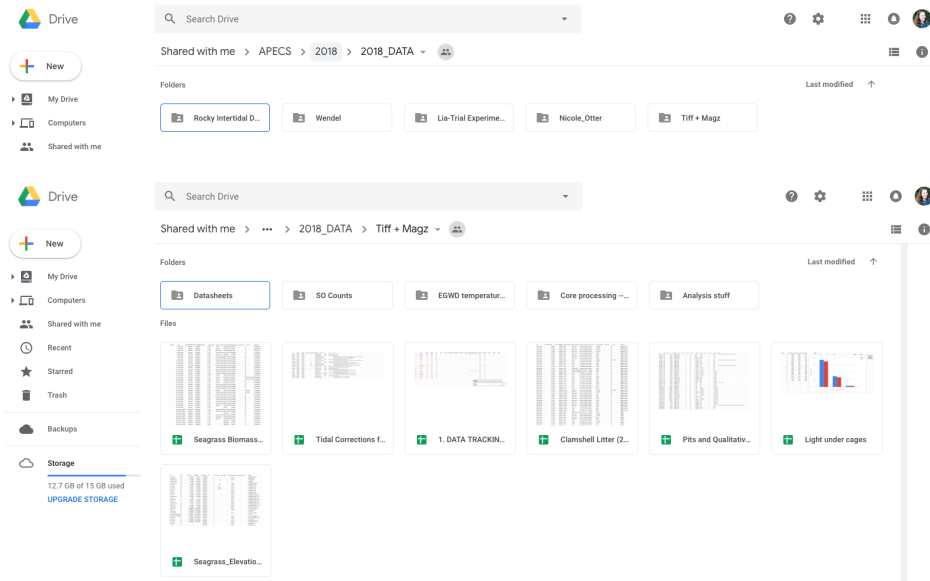
Abstract	This dataset is composed of clamshell litter data, for which the cause of death of each clam was estimated using the physical condition of the valves. Causes of death include predation by sea otter, crab, snail, and seastar/natural death. Specifically, data on the proportion of estimated cause of death of clams for each site were collected to compliment a sediment pit dataset (pits dug by identified predators above) to correct for the most likely number of pits dug by sea otters versus other animals. Clamshell litter was collected at intertidal sites that included eelgrass beds, in which three 100-m transects were placed for shell collections: within the eelgrass bed, along the edge of the eelgrass bed, and outside the eelgrass bed. Collections were conducted at 21 sites in Southeast Alaska on Prince of Wales Island. These data were collected to compliment a larger, interdisciplinary project called APECS (Apex predators, Ecosystems, and Community Sustainability), the focus of which investigated the role that sea otters have on seagrass habitats, their ecological function, and influences on traditional and subsistence harvest of specified marine organisms.
Methods	<p><b>Step 1</b></p> <p><b>Description</b> Fieldwork; collection of clamshell litter: At each site, three 100-m transects were placed parallel to the water’s edge in three distinct locations within each site: inside an eelgrass bed (‘Inside’), at the edge of an eelgrass bed (‘Edge’), and outside the eelgrass bed (‘Outside’). These three classifications required that the tidal elevation for each was different, with the ‘Edge’ being the lowest (approx. -0.37 and -1.10 MLLW) and the ‘Outside’ transect the highest (0.5 to 1.5 ft higher than the respective ‘Inside’ transect). Often, the ‘Outside’ transect corresponded with butter clam habitat. Each transect was divided into four sections: 0-25 m, 26-50 m, 51-75 m, and 76-100 m. Along each segment, clamshells were collected within a 1-m swath centered on the transect line; shells were pooled into appropriately labeled bags for each segment on each transect.</p> <p>Only shells from recently deceased clams were collected; this was determined by the amount of fouling on the shells. Large or dead barnacles, or barnacle scars, on the exterior of the valves disqualified the shells for collection. Any barnacles or barnacle scars located on the inside of the shell also disqualified the shells for collection. Mussels can quickly attach their byssus and were therefore not a disqualifying factor, nor were mobile fauna (e.g. limpets) or fast-growing seaweeds (e.g. Ulva). Additionally, only shells that still had the hinge ligament and material from both valves (even if minimal) were collected.</p> <p>All shells were transferred to the lab for estimation of the cause of death for each clam. All clam sizes were recorded to the nearest millimeter.</p> <p>Labwork; estimation of the cause of death for each clam and shell measurements: Back at the lab, clamshells were sorted by species for each segment in each transect. Shell size was then measured to the nearest millimeter. For each clam, the cause of death was first estimated for each clam; causes of death included predation by sea otter, crab, snail, or seastar/natural death. Predation by sea otter typically leaves one valve uncracked and the second valve significantly broken; often at least half of the second valve is missing. Crabs typically pinch the edges of both valves on either the anterior or posterior end of the clam, but for thin clams (e.g. bentnose Macoma), both valves may be broken anywhere along the edge. Snail predation was discernable by the characteristic circular drill hole near the umbo of one of the valves. Both seastar predation and natural death were lumped together because it was impossible to determine the difference given the nature of predation by seastars, which leaves clamshells undamaged. This combined category was recorded as ‘Whole’ for the death estimate.</p>

## 2. Day-to-day storage of data

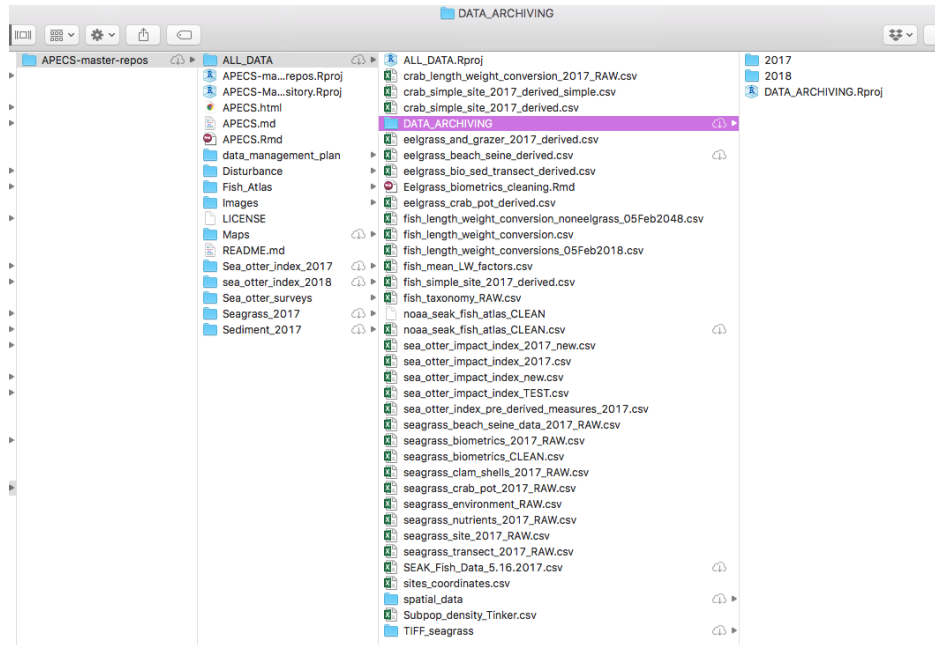
The two most important things to do remember when storing data are to keep a **Masterfile** of the each datasheet (after QC/QA) that will not be manipulated and to not lose the data. There are two places that we store data so that it is safe and accessible by APECS members: the shared Google Drive and GitHub platforms that APECS curates. Backing-up data on personal/lab hard drives and computers is also

appropriate (if not inevitable). Keep in mind, however, that it is important to be weary and conscious of multiple file versions. In the end, APECS *really* only wants two versions of the datasheet (i.e. raw and clean) for APECS-related storage and analysis, and that can get messy if you have 20 different versions of the same file.

- **Google Drive:** I think that we're all familiar with this, already, since many have their own folder within the data folder in the shared drive. If you don't have one, please start one, and if you don't have access, please ask me for it. It's fine to leave these files as google files (as opposed to .csv), but it is important to have obvious names for each file, noting whether it is raw or clean.



- **GitHub:** Tiff and Wendel have been working from the APECS GitHub folder, meaning that in addition to storage in the Google Drive, we store datasets in the “APECS-master-repos” folder to call upon in RStudio (or R) for data manipulation and analysis. We held a workshop on this last year but since few people had datasets to practice with, it is 100% understandable and recommended to consult with either Tiff or Wendel if you want to get in on the GitHub stoke. As a reminder, you'll want to have a local GitHub folder system on your local computer that links to the online GitHub repository.



### 3. Long-term archival of data

Generally, you will work with Tiff to successfully upload all APECS-collected data into our preferred data repositories. In our data management plan with NSF, we committed to archiving our data to BCO-DMO (Biological and Chemical Oceanography Data Management Office) but are also submitting to KNB (Knowledge Network for Biocomplexity). The latter is more user friendly and the archived data is accessible to BCO-DMO.

The important prep information for data archival was largely covered in the section about data structure and metadata. These are critical for archival. The following files should be submitted for archival:

- Dataset that is cleaned (required)
- Metadata (required)
- Raw dataset (optional but great for reproducibility)
- Code used to transform raw data into working data (optional but great for reproducibility)
- Diagram to help convey methods for data collection (optional)

There isn't much more to say here other than archival is a serious business and it's best to be thorough because you don't know who will access your dataset, what they'll use it for, and whether you'll still be alive if they have questions. You can look at an example from last year, where Tiff archived 2017 data using a series of submissions to KNB. *Note: KNB has been updating their user platform a lot, recently, so even Tiff needs to give it a new browse to make sure that functionality is as she knows it.*