For all analysis tasks, data cleaning is necessary.

For the majority of my analysis, I will be using a community matrix (containing species identity as column names, abundances as rows) and a second matrix containing details about each observation.

**To generate community matrix:**

This matrix should only contain data on the abundances of species and their identities and have no other labels or information. Total size = N species \* N samples, no NA values in data, only zeroes, maximum single species value of 100.

1. Find old Google refine script used to clean data previously:
   1. Look for repeated names, errors in names, etc.
   2. At least be consistent between sites
2. **Cast matrix long to wide format**
3. Think about normalizing percent cover values in some way (nearest 10%? Daubenmire?)
4. **Sort alphabetically**
5. **Error Check:**
   1. **Unit test that there are no species with NA values**
   2. **Unit test that there are no species with values over 100**

**To generate sample ID matrix:**

This matrix should contain data on the different sites, blocks, plots, and years when samples were taken, as well as any other relevant data (climate data, biomass, soils, etc.)

1. **Remove the first (8 or so) columns from the full cover dataset relating to unique values of:**
   1. **Year**
   2. **Site**
   3. **Plot**
   4. **Block**
   5. **Trt**
2. Error Check:
   1. **Unit test that there is a consistent number of observations per year**
   2. Unit test that there is an even number of treatments per year
   3. Unit test that there is the same unique set of observations in each year
   4. **Unit test Labels and treatments match comb-by-plot**

**One third idea (which is helpful) would be to generate species metadata:**

1. Using the data from the full-cover dataset, pull out the list of species, functional group and provenance.
2. For each unique case of ANY part of the data, create a list of species names, functional groups, and provenance.
3. Sort alphabetically
4. Error Check:
   1. Unit test that species are not considered both native and exotic within the same distribution
   2. Check that there aren’t any misspellings (by hand)
   3. Unit test that the names of species match up with column names