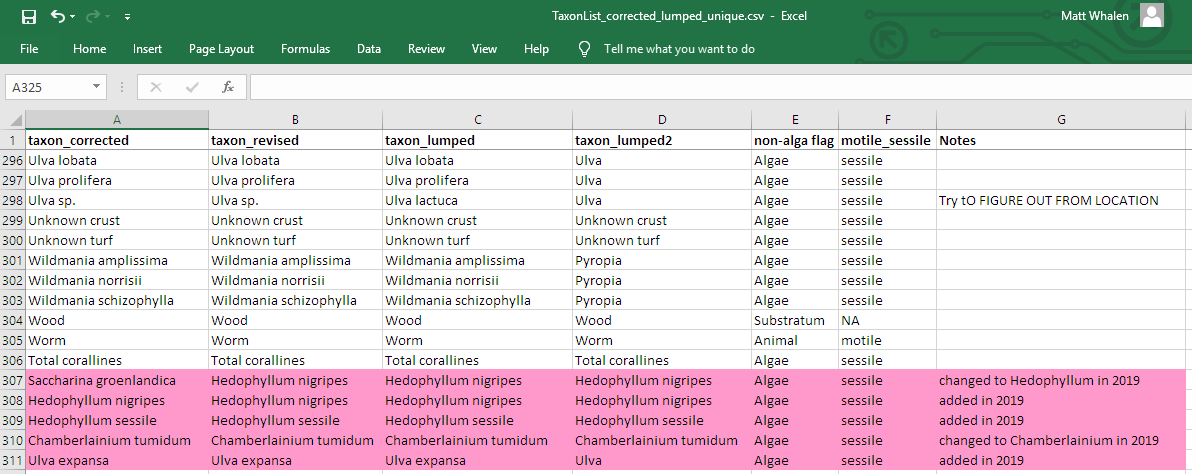
**Martone Lab – UBC – Hakai Rocky Shore Seaweed Surveys**

**Protocol for preparing tidy data for manipulation and analysis**

How to prepare raw data (as entered into Excel from field sheets) for import into R for combining data from transects at every site and shore height in every year (i.e., every tab within each Excel file).

For more information about tidy data see <https://en.wikipedia.org/wiki/Tidy_data> and <https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>

**Basic Steps:**

1. Make sheet names consistent: Use same site names (e.g. West Beach) and zone names (e.g. HIGH) for every tab, noting that capitalization matters. The only thing that should differ from year to year is the year suffix (e.g. “West Beach HIGH 2017”), which should be added to sheet names.
2. Control the dimensions of the data
   1. Delete any columns after column K. This is usually a data summary appended as new columns. Delete several columns after the data summaries to control for errant data entry.
   2. Delete any blank rows. Often seaweed and invertebrate data are separated by a blank row (usually highlighted in grey). These will mess up the data input.
   3. Resolve herbivore data into tidy rows. This is difficult in many cases because the data are somewhat ambiguous as entered. However, this must be done to keep the data in a consistent format. Also, be sure to separate rows that contain both percent cover and abundance data (e.g., *Katharina*).
   4. Try not to change species names in the raw files. They will be cross referenced using the full species key spreadsheet. So, be sure to use the names that are considered correct at the time of the survey (i.e., in the field when data are being entered)
   5. Make sure each tab/sheet in the excel file has the same number of header rows. Recently, clam shell hash has become a measure of substrate at Meay Channel, but this row should be added to every sheet in the file, so that they all have the same number of header rows. If not, then it won’t read into R.
3. Make sure any new species that have not appeared before (e.g., *Hedophyllum* in 2019) are entered into the spreadsheet “Data/taxa/TaxonList\_corrected\_lumped\_unique.csv.”
   1. New species can be added to the bottom/end of the spreadsheet. See image below for recent examples from 2019
   2. Be sure to add new names to the column taxon\_revised, taxon\_lumped, and taxon\_lumped2 as appropriate (e.g., change *Saccharina* to *Hedophyllum* for *S. sessilis* and *S. groenlandica*).
4. In the R script “Data/R Code for Data Prep/Martone\_Hakai\_Intertidal\_Data\_1\_CombineAll....csv” make sure to update the number of elements in the vector named header so that it properly describes the number of header rows for each year. NOTE that 2011 and 2012 data are combined in the same excel file, so the length of the vector header is the same as the number of years in the dataset minus one.