Data Organization in Spreadsheets

Take-home messages

- Make your data `tidy`
  - Spreadsheets should be a rectangle, with only rows and columns.
  - Each column is a different variable (a thing you are measuring, like 'weight' or 'temperature').
  - One row per observation. Each cell has only one value.
- Column headers: Use short meaningful column names with no spaces or special characters. Don't start column names with numbers. Record units in column headers.
- Use consistent names, abbreviations/codes, and capitalization.
- Use good null values (not -999, blanks ok, some prefer `NA` or similar but this can be language specific).
- Write dates as YYYYMMDD. Better still have separate columns for Year, Month, and Day.
- don't enter the same data on multiple spreadsheets: Use one for each category of data to avoid duplicated data and to simplify corrections (e.g., taxonomy).
- Avoid using multiple tables within one spreadsheet.
- Avoid spreading data across multiple tabs (but do use a new tab to record data cleaning or manipulations).
- Record zeros as zeros.
- Use an appropriate null value to record missing data.
- Don't use formatting to convey information or to make your spreadsheet look pretty.
- Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.
- Remember that data format and excel defaults can vary by region. For example, depending on the part of the world where a user is based, the default value for the decimal and thousands operator could be a , (comma) or a . (period); some regions use mm-dd for dates while others use dd-mm.
- The reason dates in Excel are so weird is that it is *accounting software*. It counts the days from a default of December 31, 1899, and thus stores July 2, 2014 as the serial number 41822. This is so one can can easily calculate "days from a given date" for accounting purposes (like invoicing) by adding "date+XX days". Furthermore, Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.
- **Once you are done with data entry, save it as 'read only' and make *all* corrections using scripting!**

- **Entering data in tidy format will make it much easier to analyze.**
- **Collecting data in tidy format makes it easier to enter data in tidy format.**

Readings, Tools, & Resources

1. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10. [read online]
2. Data Validation in Google Sheets: blog post and video tutorial. A pdf version is available for download here.
3. Why not bypass spreadsheets like Excel and use a csv editor like [Comma Chameleon][https://comma-chameleon.io/] instead? CC and other csv editors allow you to enter data in the same way - into cells, by adding and removing rows - and then export your file. But that's about it, which means you can't do many of the things (e.g., calculations, color in cells) that cause problems down the road.
4. More advanced users comfortable with R can also look into Data Curator, with whch you can create and edit tabular data from scratch or from a template, open Microsoft Excel and CSV files, and automatically correct common problems found in these and other file types.
5. DataONE Community Engagement & Outreach Working Group (2017) "Data Quality Control and Assurance". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/05_qaqc/index on Aug 31, 2020
6. DataONE Community Engagement & Outreach Working Group (2017) "Data Entry and Manipulation". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/04_entry/index on Aug 31, 2020
7. Chris Prener, Trevor Burrows (Eds.). Data Carpentry: Data Organization in Spreadsheets for Social Scientists.
8. Peter R. Hoyt, Christie Bahlai, Tracy K. Teal (Eds.), Erin Alison Becker, Aleksandra Pawlik, Peter Hoyt, Francois Michonneau, Christie Bahlai, Toby Reiter, et al. (2019, July 5). datacarpentry/spreadsheet-ecology-lesson: Data Carpentry: Data Organization in Spreadsheets for Ecologists, June 2019 (Version v2019.06.2). Zenodo. http://doi.org/10.5281/zenodo.3269869