



# CLEAN DATA WITH OPENREFINE

SFU Library Data Services  
data-services@sfu.ca  
<http://www.lib.sfu.ca/data>

**SFU**  
**LIBRARY**

# Thanks to

- 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>
- Wickham, Hadley. (2014). Tidy Data. <https://vita.had.co.nz/papers/tidy-data.pdf>
- Broman, Karl W, & Woo, Kara H. (2017). Data Organization in Spreadsheets. The American Statistician, 72(1), 2–10. <https://doi.org/10.1080/00031305.2017.1375989>
- Evan Will (2021). Getting Started with OpenRefine: Explore, Clean, and Transform your Data. University of Idaho Library. <https://evanwill.github.io/openrefine-b/>
- University of Edinburgh (2019). OpenRefine Beginners Tutorial. [https://media.ed.ac.uk/media/OpenRefine+Beginners+Tutorial/0\\_y5bxsswq](https://media.ed.ac.uk/media/OpenRefine+Beginners+Tutorial/0_y5bxsswq)

# Objectives

- Understand principles of tidy data
- Identify common errors in messy data
- Clean data using the following OpenRefine functions:
  - Creating an OpenRefine project
  - Filters
  - Facet
  - Transform and transpose
  - Basic GREL language functions

# Intro Questions

---

Do you work with data in spreadsheets?

---

Could you share your data with someone else?

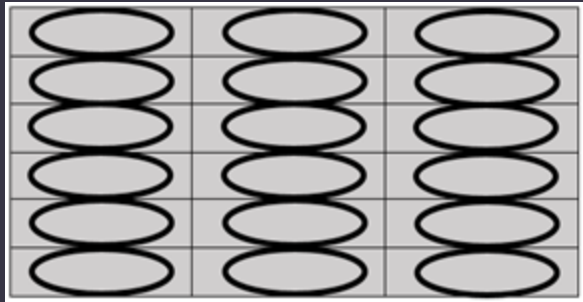
---

Could you use your data in 5, 10, or 15 years?

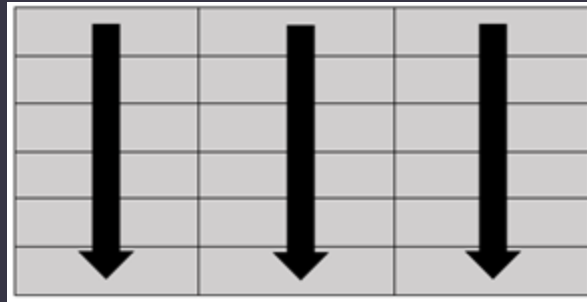
---

Could you use your data with different software?

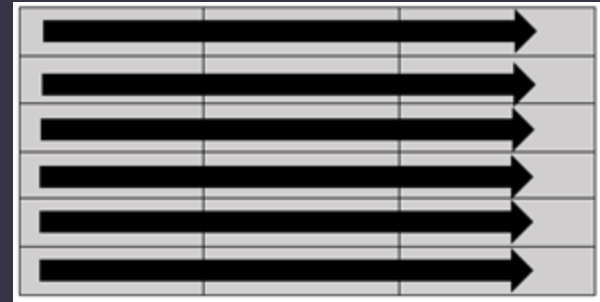
# Basics of tidy data



- One value per cell



- One variable per column



- One observation per row

# Recommendations for tidy data

- Consistent null or NA values
  - Using "0" or "999" is ambiguous and could be included as actual values in analysis
  - "NULL" or "NA" are best options
  - Make sure to document what value you choose
  - Include a notes column for information about why a value is missing

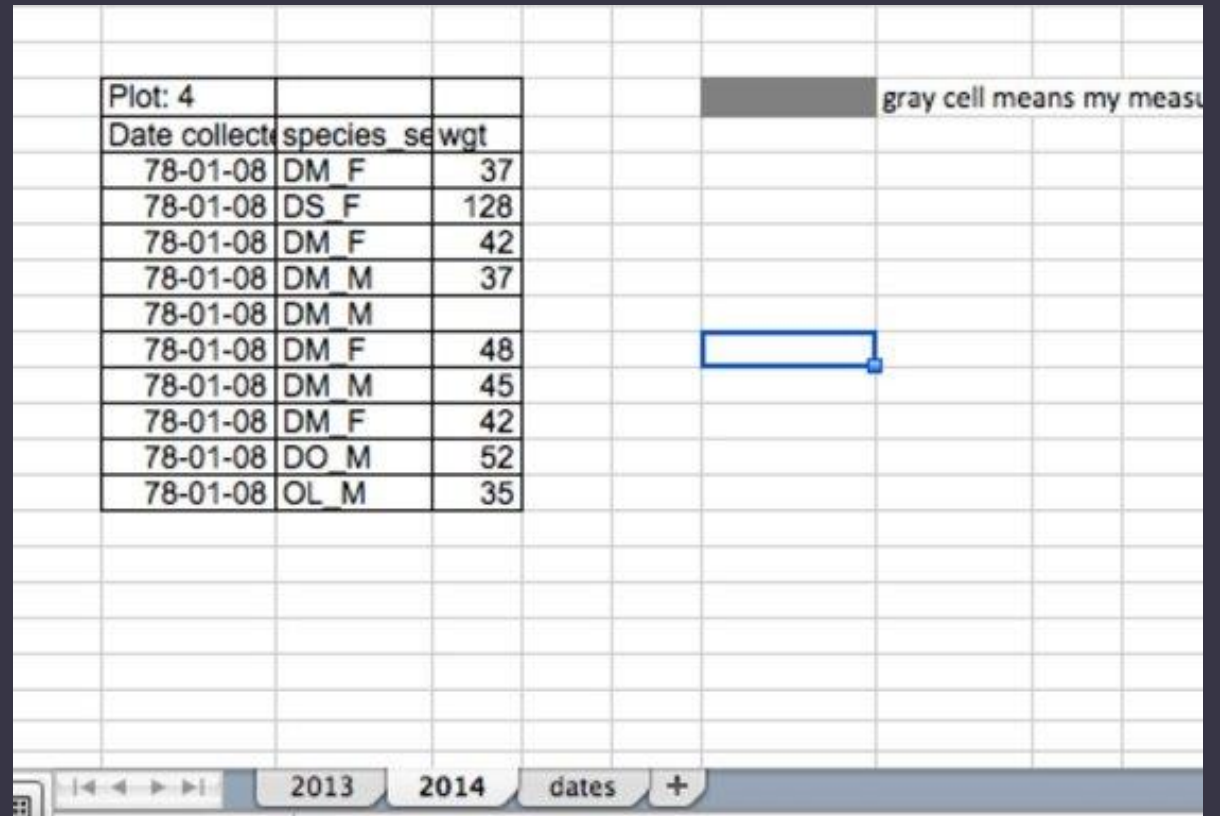
# Recommendations

- Date formatting
  - Dates are easily misinterpreted (by software and humans)
  - Choose a single format and use consistently
  - Consider storing in three different columns – year, month, day

<b>Date</b>	<b>Number</b>	<b>How it was interpreted</b>
July-10	40330	2010-06-01
July-14	41791	2014-06-01
July-15	42156	2015-06-01
July-17	42887	2017-06-01

# Recommendations

- Avoid multiple tabs or tables in a single spreadsheet
- Use columns instead of tabs
- Human-readable headers/labels are not always computer-readable



Plot: 4		
Date collect	species	se wgt
78-01-08	DM F	37
78-01-08	DS F	128
78-01-08	DM F	42
78-01-08	DM M	37
78-01-08	DM M	
78-01-08	DM F	48
78-01-08	DM M	45
78-01-08	DM F	42
78-01-08	DO M	52
78-01-08	OL M	35

gray cell means my measu



# Recommendations

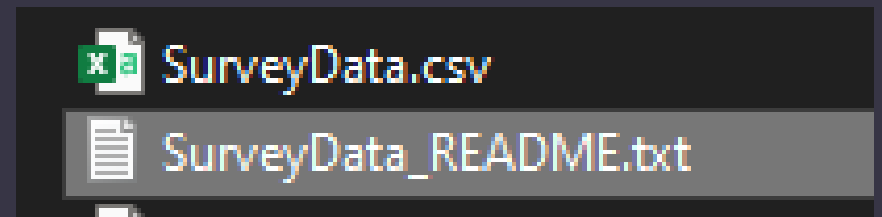
- Do not use formatting (color, bolding, etc.) to convey information
- May not be preserved across formats/software
- Convey that information in data dictionary or readme instead

Plot: 2			
Date collected	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

measurement device not calibrated

# Recommendations

- Document any abbreviations, field names, additional notes in a separate file
  - Data dictionaries, READMEs, metadata
- Include any contextual information needed to understand the data
- Make sure someone who was not involved in the project could understand your data
- Save this file alongside your data file(s)



# What is OpenRefine?

- Free, open-source java tool
- Data cleaning, exploration, transforming, editing
- Tracks any changes to datasets
- API integration for data augmentation and reconciliation with external data sources

# Why use OpenRefine?

- More user-friendly, visual than python or R
- More powerful than Excel
- Connect to online data sources
- Easily automate cleaning tasks for future datasets or projects