A small horizontal bar with a teal-to-orange gradient.

Clean Data: Organizing Data in Spreadsheets Part 1

SFU Library Data Services Team

Website: www.lib.sfu.ca/data

Email: data-services@sfu.ca



Part 1: Basic Spreadsheet Structures

- Values, variables, and observations
- 3 rules of basic spreadsheet structures
- How to spot structure errors and how to fix them



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>
- 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>
- Wickham, Hadley. R for Data Science. Hadley Wickham.
<https://r4ds.had.co.nz/tidy-data.html> (online, free)
- Wickham, Hadley. (2014). Tidy Data.
<https://vita.had.co.nz/papers/tidy-data.pdf>



INTRO QUESTION

- Do you work, or have worked, with data in spreadsheets?
- Could you share your data with someone?
- Could you use your data in 5, 10, or 15 years?
- Could you use your data with a different software?

KEEP TRACK OF WHAT YOU DO



- Create new file to do cleanup, don't modify the original

The screenshot displays two windows. The background window is an Excel spreadsheet titled 'survey_data - Excel'. It contains a table with columns A through H: Date, Year, Month, Day, Plot, Sex, Weight_g, and Collector. The data rows show dates from 12/8/2020 to 12/19/2020, with corresponding year, month, day, plot, sex, weight, and collector names.

The foreground window is a Notepad application titled '*README_survey_data - Notepad'. It contains the following text:

```
File Edit Format View Help
Processing notes on survey_data.csv

2020-12-08 work done -----

1. Created Year, Month, Day fields: used =YEAR(),
=MONTH(), and =DAY() formulas to pull information
from Date field

2. Created Plot field: moved information from
survey_data_plots.csv spreadsheet

3. Standardized all measurements in Weight field to
grams and updated field name to Weight_g
```

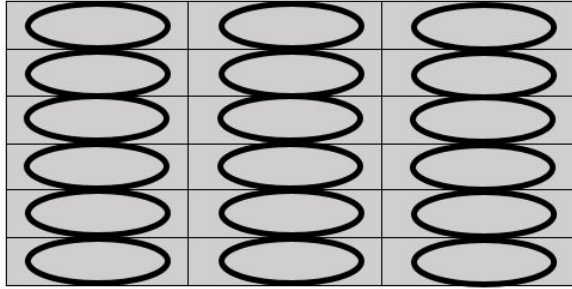
- Keep track of your clean-up steps in a separate .txt file



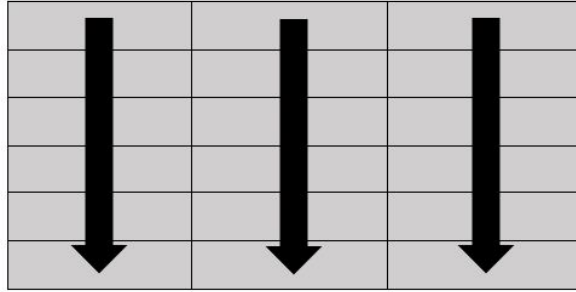
BASIC SPREADSHEET STRUCTURES

1. Don't combine multiple pieces of information in one cell.
2. Put each variable in one column
 - What you're measuring/documenting, (e.g. 'weight' or 'temperature'.)
3. Put each observation in its own row.

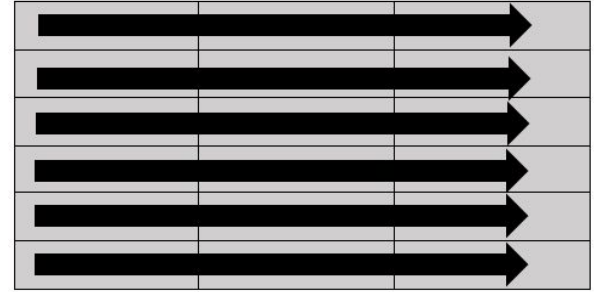
BASIC SPREADSHEET STRUCTURES



Values



Variables



Observations

MULTIPLE VALUES IN CELLS

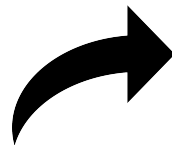
Sales
USD 10
CAD 100
GBP 1000
JPY 1



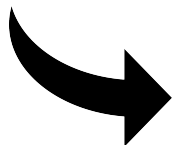
Sales	Currency
10	USD
100	CAD
1000	GBP
1	JPY

VARIABLES IN MULTIPLE COLUMNS

	Site		
Species	BR	NV	PC
Cat	0	3	4
Dog	7	2	0



Species	Site	Count
Cat	BR	0
Cat	NV	3
Cat	PC	4
Dog	BR	7
Dog	NV	2
Dog	PC	0



Species	SiteBR	SiteNV	SitePC
Cat	0	3	4
Dog	7	2	0

OBSERVATIONS IN ROWS

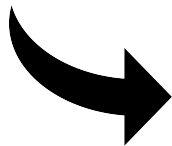
Country	1999	2000
Canada	400	600
USA	700	300
China	200	100



Country	Year	Cases
Canada	1999	400
Canada	2000	600
USA	1999	700
USA	2000	300
China	1999	200
China	2000	100

OBSERVATIONS IN ROWS

Country	Year	Type	Count
Canada	1999	Cases	700
Canada	1999	Population	29000000
Canada	2000	Cases	800
Canada	2000	Population	30000000

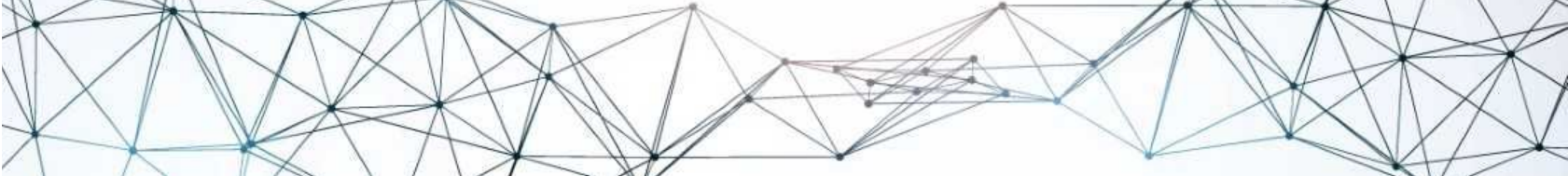


Country	Year	Cases	Population
Canada	1999	700	29000000
Canada	2000	800	30000000

RECOMMENDATIONS

- Be consistent
- Put fixed variables first and then measured variables
- Only include unique information in the data

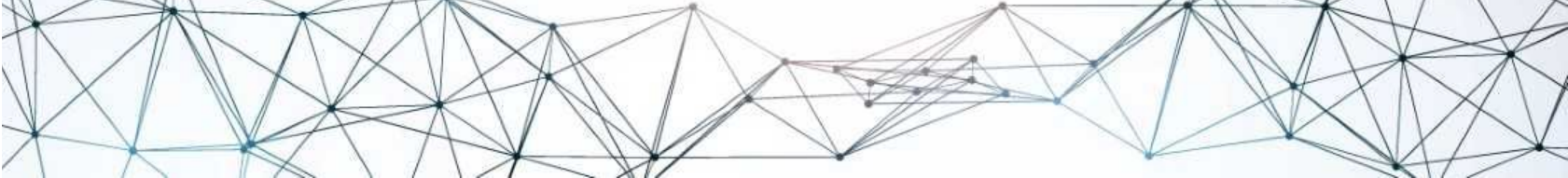
Species	Height	Height Unit
Cat	15	Centimeters
Dog	30	Centimeters



FOR MORE INFORMATION

FIND	HELP	BORROW	FACILITIES	ABOUT
RESEARCH ASSISTANCE	CITE + WRITE	WORKSHOPS + CONSULTATIONS	ACADEMIC INTEGRITY	PUBLISH
Find materials by subject + course	Citation + style guides	All workshops + classes	Copyright	Scholarly publishing + Open access
Find materials by format + type	Citation management software	Undergraduate workshops	Avoiding plagiarism	Summit Research Repository
Research tutorials	Undergraduate writing + learning (SLC)	Graduate workshops		Research data management
Frequently asked questions	Graduate writing, learning + research (RC)	Undergraduate consultations (SLC)		Digital Humanities Innovation Lab
Services for you		Graduate consultations (RC)		Thesis submission
Ask a librarian				

Contact us at: data-services@sfu.ca

A small horizontal bar with a teal-to-orange gradient.

Clean Data: Organizing Data in Spreadsheets Part 2

SFU Library Data Services Team

Website: www.lib.sfu.ca/data

Email: data-services@sfu.ca



Part 2: Common Spreadsheet Errors

- Common spreadsheet errors
- Null values
- Field names
- Dates



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>
- 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>
- Wickham, Hadley. R for Data Science. Hadley Wickham.
<https://r4ds.had.co.nz/tidy-data.html> (online, free)
- Wickham, Hadley. (2014). Tidy Data.
<https://vita.had.co.nz/papers/tidy-data.pdf>



COMMON ERRORS, PART 1

- Multiple tables in one sheet
- Multiple tabs
- Not filling in zeros
- Using bad null values (or none at all)
- Using formatting to convey information
- Using formatting to make the data sheet look 'nice'



COMMON ERRORS, PART 2

- Placing comments or units in cells
- More than one piece of information in a cell
- Field name problems
- Special characters in data
- Inclusion of metadata in data table
- Date formatting

[illegible]

MULTIPLE TABS



Plot: 4			gray cell means my meas		
Date collect	species	sewgt			
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			

2013 2014 dates +

FORMATTING (TO CONVEY INFORMATION)

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			



Date collected	Plot	Species	Sex	Weight	Calibrated
2014-01-08	2	NA			
2014-01-08	2	DM	M	44	Y
2014-01-08	2	DM	M	38	Y
2014-01-08	2	OL			
2014-01-08	2	PE	M	22	Y
2014-01-08	2	DM	M	38	Y
2014-01-08	2	DM	M	48	Y
2014-01-08	2	DM	M	43	Y
2014-01-08	2	DM	F	35	Y
2014-01-08	2	DM	M	43	Y
2014-01-08	2	DM	F	37	Y
2014-01-08	2	PF	F	7	Y
2014-01-08	2	DM	M	45	Y
2014-01-08	2	OT			
2014-01-08	2	DS	M	157	N
2014-01-08	2	OX			
2014-02-18	2	NA	M	218	N
2014-02-18	2	PF	F	7	Y
2014-02-18	2	DM	M	52	Y

NULL VALUES

Null values	Problems	Recommend?
0	Cannot be distinguished from actual zero value	✗
Blank	Is it a missing value? A mistake? Might be ignored in calculations? Create a new column called DataMissing to explain blanks.	✓
NULL	Might cause problems depending on data type	✓
NA or na	Could turn a numerical column into text, could also be an abbreviation (North America?)	✓
N/A	Not compatible with most statistical software	✗
-999 or 999	Could be accidentally used in calculations!	✗

Table adapted from White, E. P., Baldrige, E., Brym, Z. T., Locey, K. J., McGlinn, D. J., & Supp, S. R. (2013). Nine simple ways to make it easier to (re) use your data. *Ideas in Ecology and Evolution*, 6(2). doi:10.4033/iee.2013.6b.6.f

IMPROVING FIELD NAMES

- ✗ spaces
- ✗ special characters
- ✗ abbreviations
- ✓ underscores (_)
- ✓ camelCase

Avoid	Good option	Alternative
Cell Type	cell_type	CellType
Maximum temp (°C)	max_temp_C	MaxTemp
Mean growth/year	Mean_year_growth	MeanYearGrowth
1st Obs	observation_01	first_observation
precmm	precipitation_mm	Precipitation
w.	weight	weight

DATES AS DATA

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

Putting dates in a single column is not great!

DATES AS DATA: OPTION 1

1. Store dates in separate year, month, day columns using the *YEAR()*, *MONTH()*, *DAY()* functions.

Date	Month	Day	Year
2015-06-08	=MONTH(A2)	=DAY(A2)	=YEAR(A2)
2015-06-09	6	9	2015
2015-06-10	6	10	2015
2015-06-11	6	11	2015
2015-06-12	6	12	2015
2016-08-07	8	7	2016
2016-08-08	8	8	2016

DATES AS DATA: OPTION 2

2. Store dates in two columns for the year and the day-of-year (DOY). Use $A_3 - DATE(YEAR(A_3), 1, 0)$ for the day-of-year.

Date	Year	DOY
6/8/2015	=YEAR(A2)	=A2 - DATE(YEAR(A2), 1, 0)
6/9/2015	2015	160
6/10/2015	2015	161
6/11/2015	2015	162
6/12/2015	2015	163
6/13/2015	2015	164

DATES AS DATA: OPTION 3

3. Store dates in the *YYYYMMDDhhmmss* format, e.g.,
February 14, 2020 09:45:14 would be 20200214094514

Date	Date
3/24/2015 5:25:35 PM	=TEXT(A2,"yyyymmddhhmmss")
3/25/2015 7:16:32 PM	20150325191632
3/26/2015 7:09:35 AM	20150326070935
3/30/2015 8:52:53 AM	20150330085253



SPREADSHEET EXAMPLE

Workshop materials:

https://1sfu-my.sharepoint.com/:f:/g/personal/libdata_sfu_ca/EnEdJDHd_21BpUCUIQu7H9ABjtjhd7LTGjsAyDUEanm2g?e=LzoNTm

1. messy_data.xlsx
2. cleaned_data.xlsx

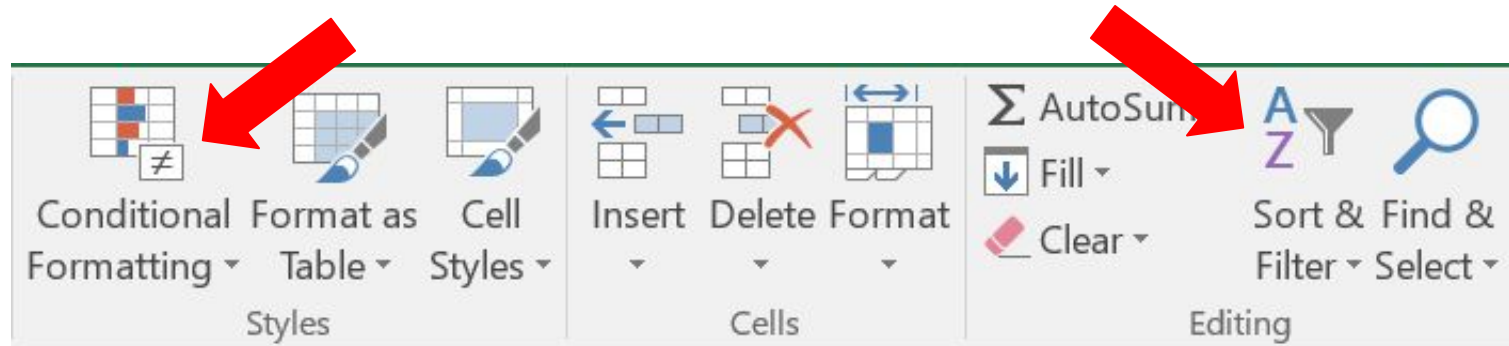


FIXING PROBLEMS

- Leave raw data raw, create a new file altogether, and document changes
- Can use:
 - *Reshape* package in R to clean cross-tabular data
 - Or fix it in Excel:
<https://youtu.be/FozNs-cyGqA>

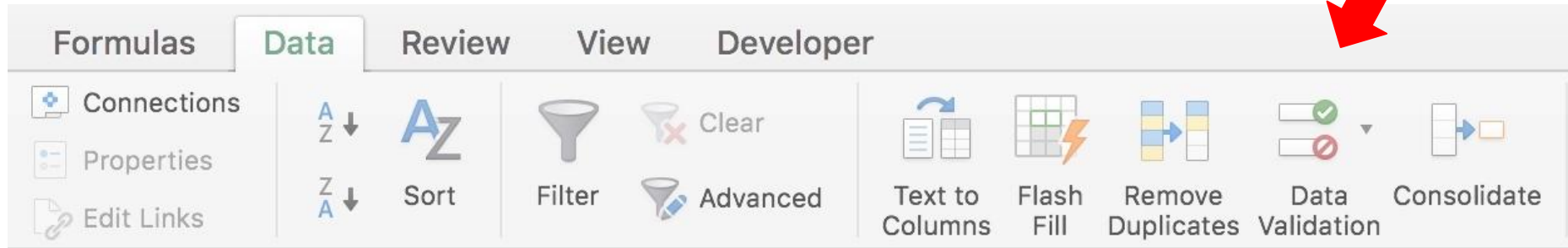
QUALITY CONTROL

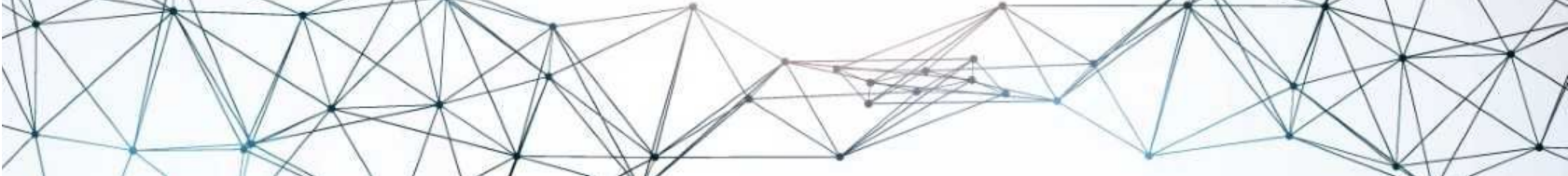
1. **Filter columns**
 - a. Displays unique values: use to scan for errors
 - b. Sort: problem values often sort to the bottom or top
2. **Conditional formatting** can flag inconsistent values



DATA VALIDATION

Data validation can limit errors in data entry, by controlling the values allowed in each column.

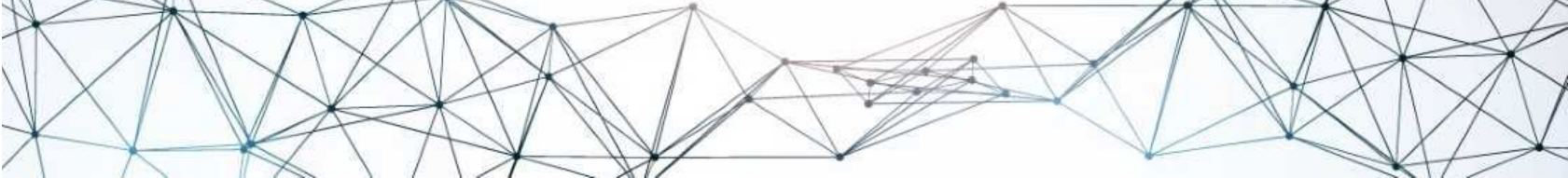




FOR MORE INFORMATION

FIND	HELP	BORROW	FACILITIES	ABOUT
RESEARCH ASSISTANCE	CITE + WRITE	WORKSHOPS + CONSULTATIONS	ACADEMIC INTEGRITY	PUBLISH
Find materials by subject + course	Citation + style guides	All workshops + classes	Copyright	Scholarly publishing + Open access
Find materials by format + type	Citation management software	Undergraduate workshops	Avoiding plagiarism	Summit Research Repository
Research tutorials	Undergraduate writing + learning (SLC)	Graduate workshops		Research data management
Frequently asked questions	Graduate writing, learning + research (RC)	Undergraduate consultations (SLC)		Digital Humanities Innovation Lab
Services for you		Graduate consultations (RC)		Thesis submission
Ask a librarian				

Contact us at: data-services@sfu.ca

A small horizontal bar with a teal-to-orange gradient.

Clean Data: Organizing Data in Spreadsheets Part 3

SFU Library Data Services Team

Website: www.lib.sfu.ca/data

Email: data-services@sfu.ca



Part 3: Documentation and Exporting

- Data documentation
 - ReadMe files
 - Data Dictionaries
- Exporting data from spreadsheets
- Finishing touches



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>
- 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>
- Wickham, Hadley. R for Data Science. Hadley Wickham.
<https://r4ds.had.co.nz/tidy-data.html> (online, free)
- Wickham, Hadley. (2014). Tidy Data.
<https://vita.had.co.nz/papers/tidy-data.pdf>

DOCUMENTATION: README FILE

- ReadMe are plain text files that document:
 - General project Info
 - Data and file overview
 - Methodology
 - Sharing and access info
- [Cornell University Guide to writing "readme" style metadata](#)

DOCUMENTATION: DATA DICTIONARY

- Gives info about variables
 - Variable names
 - Format
 - Measurement unit
 - Expected values (nulls, mix/max, list)

Variable	Variable Name	Format	Measurement Unit	Allowed Values	Description
Date_Collected	Date	yyyy-mm-dd		2015-03-12 to 2016-05-06	When the data was collected
Species	Species	Text		Cat, Dog, Raccoon	Species that was observed
Sex	Sex	Numeric		1 = Female, 2= Male	Sex of animal
Hgt.	Height	Numeric	Centimeters	0-999	Height of animal in centimeters

DOCUMENTATION: DATA DICTIONARY

- Resources for making data dictionaries
 - [U.S. Geological Survey Data Management: Data Dictionaries](#)
 - [Open Science Framework: How to Make a Data Dictionary](#)

EXPORTING FROM SPREADSHEETS

Excel (or OpenOffice) default file format not ideal

- proprietary
- data inconsistencies
- not accepted in repositories

Save your spreadsheet in CSV or TSV format

WHAT IF COMMAS ARE VALUES?

genus, species, taxa

Amphispiza, bilineata, Bird

Ammospermophilus, harrisi, Rodent-not, censused

Ammodramus, savannarum, Bird

Baiomys, taylori, Rodent

genus	species	taxa	
Amphispiza	bilineata	Bird	
Ammospermophilus	harrisi	Rodent-not	censused
Ammodramus	savannarum	Bird	
Baiomys	taylori	Rodent	

WHAT IF COMMAS ARE VALUES?

genus, species, taxa

Amphispiza, bilineata, Bird

Ammospermophilus, harrisi, "Rodent-not, censused"

Ammodramus, savannarum, Bird

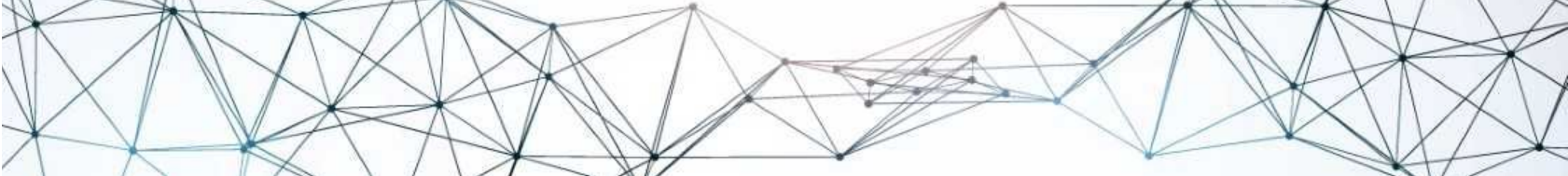
Baiomys, taylori, Rodent

genus	species	taxa
Amphispiza	bilineata	Bird
Ammospermophilus	harrisi	Rodent-not, censused
Ammodramus	savannarum	Bird
Baiomys	taylori	Rodent



FINISHING TOUCHES

1. Ensure files are organized consistently
2. Keep backup copies
3. Document the steps you take/changes you make
4. Maintain the raw data file and don't modify it
5. For major manipulations, save as separate files and use naming and versioning conventions
6. Turn formulas into static values after you use them



FOR MORE INFORMATION

FIND	HELP	BORROW	FACILITIES	ABOUT
RESEARCH ASSISTANCE	CITE + WRITE	WORKSHOPS + CONSULTATIONS	ACADEMIC INTEGRITY	PUBLISH
Find materials by subject + course	Citation + style guides	All workshops + classes	Copyright	Scholarly publishing + Open access
Find materials by format + type	Citation management software	Undergraduate workshops	Avoiding plagiarism	Summit Research Repository
Research tutorials	Undergraduate writing + learning (SLC)	Graduate workshops		Research data management
Frequently asked questions	Graduate writing, learning + research (RC)	Undergraduate consultations (SLC)		Digital Humanities Innovation Lab
Services for you		Graduate consultations (RC)		Thesis submission
Ask a librarian				

Contact us at: data-services@sfu.ca