DATA MANAGEMENT STRATEGIES

Kim Cressman and Gabriel Kamener

Catbird Stats, LLC (KC); Florida International University (GK)

2023-11-16

GOALS FOR THIS PRESENTATION

- **NOT** to shame you
- Give you some knowledge to build better datasets
  - some = a manageable amount
  - …moving forward
- Point you toward helpful resources

SPREADSHEETS ARE USEFUL

EVEN I WILL NOT FIGHT THAT FACT

There are multiple purposes for keeping tabular data in spreadsheets:

- Data entry
- Data storage
- Data analysis
- Presentation

SPREADSHEETS *CAN BE* DANGEROUS

THINGS TO BE CAREFUL OF

- Proprietary software
- Repeating data on many rows can lead to unnoticed mistakes

REPETITION MISTAKES

TransectID	PlotID	Lat	Long	Distance	Orthometric Height	*Height Relative to MLLW*	Sp
2	7	33.34036	-79.20251	67	0.2790	1.1040	unvegetated
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	unvegetated
2	9	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	9	33.34020	-79.20233	91	-0.0930	0.7320	unvegetated
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Distichlis spic*
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Iva frutescens*
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Spartina alter*
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Spartina pate*

REPETITION MISTAKES

TransectID	PlotID	Lat	Long	Distance	Orthometric Height	*Height Relative to MLLW*	Sp
2	7	33.34036	-79.20251	67	0.2790	1.1040	unvegetated
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	unvegetated
2	9	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	9	33.34020	-79.20233	91	-0.0930	0.7320	unvegetated
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Distichlis spic*
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Iva frutescens*
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Spartina alter*
3	1	33.34056	-79.20316	2	1.0090	1.8340	*Spartina pate*

REPETITION MISTAKES

TransectID	PlotID	Lat	Long	Distance	Orthometric Height	*Height Relative to MLLW*	Sp
2	7	33.34036	-79.20251	67	0.2790	1.1040	unvegetated
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	8	33.34028	-79.20242	79	0.1993	1.0243	unvegetated
2	9	33.34028	-79.20242	79	0.1993	1.0243	*Spartina alter*
2	9	33.34020	-79.20233	91	-0.0930	0.7320	unvegetated

THINGS TO BE CAREFUL OF

- Proprietary software
- Repeating data on many rows can lead to unnoticed mistakes
- Excessive formula use can eventually lead to unseen errors

EXCEL AND DATES

- dates “seen” differently in the computer than in our brains
- dates seen differently based on operating system

BEST PRACTICES

HELPFUL SOURCES

- Broman and Woo 2018, Data Organization in Spreadsheets (open access)
- White et al. 2013, Nine simple ways to make it easier to (re)use your data (pdf)
- Tampa Bay Estuary Program Data Management SOP
- Data Carpentry Workshop on Data Organization in Spreadsheets

RECTANGLES

- One table per sheet
- When adding data, add rows, not columns
- One *type* of data per column
  - don’t type “No Data” in an otherwise numeric column
- Be thoughtful about column names
  - and don’t use special characters in them
- Be thoughtful about representation of missing data

MAKE INFORMATION EXPLICIT

- QA/QC columns, rather than comments on a cell
- Additional columns, rather than [only] color coding

DIFFERENT TABLES FOR DIFFERENT DATA TYPES

- Think: site information you only measure once (lat/long, habitat type, etc.) vs. information you measure every time

NERR Site ID	Station Code	Station Name	Latitude	Longitude	Status	Active Dates	Reserve Name	Real Time	HADS ID	GMT Offset	Station Type
noc	noclcwq	Loosin Creek	34.1722	77.8328	Active	Feb 2002-	North Carolina			-5	1
noc	nocrcmet	Research Creek	34.1555	77.8509	Active	Jan 2001-	North Carolina	R	3B02028E	-5	0
noc	nocrcnut	Research Creek	34.156	77.8499	Active	Apr 2002-	North Carolina			-5	2
noc	nocrcwq	Research Creek	34.156	77.8499	Active	Jan 1994-	North Carolina	R	3B032698	-5	1
noc	noczbnut	Zeke's Basin	33.9547	77.935	Active	Apr 2002-	North Carolina			-5	2
noc	noczbwq	Zeke's Basin	33.9547	77.935	Active	Mar 2002-	North Carolina	R	3B04523C	-5	1
owc	owcbrnut	Berlin Road	41.34889	82.51222	Active	Mar 2002-	Old Woman Creek			-5	2
owc	owcbrwq	Berlin Road	41.34889	82.51222	Active	Mar 2002-	Old Woman Creek			-5	1
owc	owcolnut	Lower Estuary	41.38167	82.51389	Active	Apr 2002-	Old Woman Creek			-5	2
owc	owcolwq	Lower Estuary	41.38167	82.51389	Active	Apr 2002-	Old Woman Creek	R	3B02849A	-5	1
owc	owcowmet	Old Woman Creek	41.37778	82.50806	Active	May 2001-	Old Woman Creek	R	3B017310	-5	0

DIFFERENT TABLES FOR DIFFERENT DATA TYPES

- Think: site information you only measure once (lat/long, habitat type, etc.) vs. information you measure every time
- Tables can be related to each other via common columns, known as “keys”
  - can even do this in Excel, with XLOOKUP

EXAMPLE: FISH MONITORING DATA

	A	B	C	D	E
1	site	habitat_type	lat	long	location
2	11	erosional edge	30.37163	-88.4438	Bayou Cumbest
3	14	erosional edge	30.3557	-88.4495	Pt aux Chens Bay
4	2	seagrass	30.38508	-88.4022	Middle Bay
5	3	seagrass	30.36205	-88.3977	Grand Bay
6	6	erosional edge	30.34905	-88.3973	Grand Battures
7	8	seagrass	30.35493	-88.4106	Jose Bay

	A	B	C	D	E	F	G
1	collection_id	site	season	year_sampled	salinity_ppt	do_mgl	water_temp_c
2	NFM08-142	2	Winter	2008	18.4	8.24	14
3	NFM08-143	3	Winter	2008	17.3	7.98	14.5
4	NFM08-146	6	Winter	2008	17.8	8.68	13.9
5	NFM08-148	8	Winter	2008	19.3	8.52	15.2
6	NFM08-151	11	Winter	2008	18.1	7.27	17.1
7	NFM08-154	14	Winter	2008	19.6	9.12	18.3
8	NFM08-156	2	Spring	2008	17.4	6.15	27.4
9	NFM08-157	3	Spring	2008	18.7	5.8	28.2
10	NFM08-160	6	Spring	2008	18.2	6.17	30
11	NFM08-162	8	Spring	2008	18.7	7.16	29.1
12	NFM08-165	11	Spring	2008	12.9	5.92	31.9
13	NFM08-168	14	Spring	2008	16.7	7.72	31.9
14	NFM08-169	2	Summer	2008	18.8	3.42	29.6
15	NFM08-170	3	Summer	2008	19.2	4.12	28.6
16	NFM08-173	6	Summer	2008	19.2	5.78	29.8
17	NFM08-175	8	Summer	2008	20.8	5.07	30.2
18	NFM08-178	11	Summer	2008	10.4	4.29	31.9
19	NFM08-180	14	Summer	2008	20.5	5.81	31.5

EXAMPLE: FISH MONITORING DATA

	A	B	C	D	E
1	site	habitat_type	lat	long	location
2	11	erosional edge	30.37163	-88.4438	Bayou Cumbest
3	14	erosional edge	30.3557	-88.4495	Pt aux Chens Bay
4	2	seagrass	30.38508	-88.4022	Middle Bay
5	3	seagrass	30.36205	-88.3977	Grand Bay
6	6	erosional edge	30.34905	-88.3973	Grand Battures
7	8	seagrass	30.35493	-88.4106	Jose Bay

one-to-many

	A	B	C	D	E	F	G
1	collection_id	site	season	year_sampled	salinity_ppt	do_mgl	water_temp_c
2	NFM08-142	2	Winter	2008	18.4	8.24	14
3	NFM08-143	3	Winter	2008	17.3	7.98	14.5
4	NFM08-146	6	Winter	2008	17.8	8.68	13.9
5	NFM08-148	8	Winter	2008	19.3	8.52	15.2
6	NFM08-151	11	Winter	2008	18.1	7.27	17.1
7	NFM08-154	14	Winter	2008	19.6	9.12	18.3
8	NFM08-156	2	Spring	2008	17.4	6.15	27.4
9	NFM08-157	3	Spring	2008	18.7	5.8	28.2
10	NFM08-160	6	Spring	2008	18.2	6.17	30
11	NFM08-162	8	Spring	2008	18.7	7.16	29.1
12	NFM08-165	11	Spring	2008	12.9	5.92	31.9
13	NFM08-168	14	Spring	2008	16.7	7.72	31.9
14	NFM08-169	2	Summer	2008	18.8	3.42	29.6
15	NFM08-170	3	Summer	2008	19.2	4.12	28.6
16	NFM08-173	6	Summer	2008	19.2	5.78	29.8
17	NFM08-175	8	Summer	2008	20.8	5.07	30.2
18	NFM08-178	11	Summer	2008	10.4	4.29	31.9
19	NFM08-180	14	Summer	2008	20.5	5.81	31.5

DATA SAFETY

- Don’t do any calculations in the raw data file!
  - Make a copy.
- Back up your data!
  - Keep it in 3 places
  - At least one in a different physical location

DOCUMENTATION (METADATA)

- Who, what, when, where, why
- How
- Data dictionary
- No universal standard, but several formats exist
- JUST DO IT

WHEN TO MOVE BEYOND SPREADSHEETS

WAIT, WHAT'S "BEYOND" A SPREADSHEET?

Relational database!

"All relational databases organize data into sets of interlinked tables." -Thomer and Wickett 2020 (open access)

"A database is, in some sense, just a collection of tables, where there's some value in the tables that allows them to be connected to each other (the 'related' part of 'relational database')." -Data Carpentry 'Data Management with SQL for Ecologists' workshop

SOME SOFTWARE EXAMPLES

- Access
- Oracle
- MySQL

ADVANTAGES

(of a well-built relational database)

- “Front end” / “Back end”
  - data entry is (can be) human-friendly
  - data validation
  - data storage is computer-friendly
  - all the linkages happen without you having to think about them

ADVANTAGES

(of a well-built relational database)

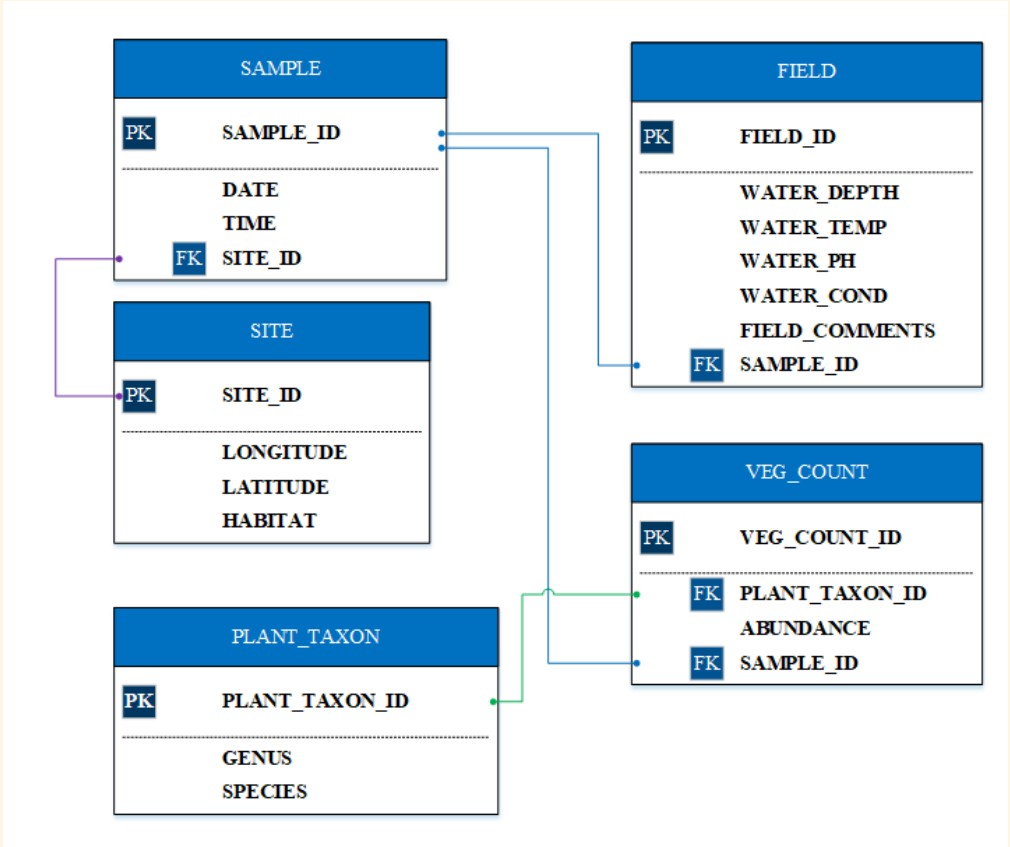
- Queries - you can pull data back out in different ways
  - e.g., if you wanted the lat/long and habitat information associated with each individual sampling event or even individual fish
    - without the errors you'd get from copying and pasting that information into every sample row
  - WITHOUT altering the original data

BARRIERS

- Not everybody has ready access to database expertise
- Not every database is designed well
- Good databases require thoughtful design, as well as ongoing maintenance

WHEN TO THINK ABOUT DATABASES

- Long-term projects
- Projects involving lots of complexity
- Projects where consistency system-wide is important


SAMPLE
PK
SAMPLE_ID
DATE
TIME
FK
SITE_ID
FIELD
PK
FIELD_ID
WATER_DEPTH
WATER_TEMP
WATER_PH
WATER_COND
FIELD_COMMENTS
FK
SAMPLE_ID
SITE
PK
SITE_ID
LONGITUDE
LATITUDE
HABITAT
VEG_COUNT
PK
VEG_COUNT_ID
FK
PLANT_TAXON_ID
ABUNDANCE
FK
SAMPLE_ID
PLANT_TAXON
PK
PLANT_TAXON_ID
GENUS
SPECIES

WRAP-UP

- Make it rectangular
- Think about simple tables, and linkages
- …even if you stay in spreadsheets
- Make it clear what's going on, through good naming systems and metadata

HELPFUL SOURCES

- Broman and Woo 2018, Data Organization in Spreadsheets (open access)
- White et al. 2013, Nine simple ways to make it easier to (re)use your data (pdf)
- Tampa Bay Estuary Program Data Management SOP
- Data Carpentry Workshop on Data Organization in Spreadsheets