Data Entry, and Manipulation

DataONE Community Engagement & Outreach Working Group

Lesson Topics

- Best Practices for Creating Data Files
- Data Entry Options
- Data Integration Best Practices
- Data Manipulation Options

Learning Objectives

- Recognize and plan for inconsistencies that can make a dataset difficult to understand and/or manipulate
- Describe characteristics of stable data formats and list reasons for using these formats
- Identify data entry tools
- Identify validation measures that can be performed as data is entered
- Review best practices for data integration
- Describe the basic components of a relational database

Goals of Data Entry

- Create quality data sets that are:
  - Valid
  - Organized to support ease of use and reuse

Example: Poor Data Entry

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Wel	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1+	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	15									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

- **Inconsistency between data collection events**
  - **Location of Date information**
  - **Inconsistent Date format**
  - **Column names**
  - **Order of columns**

DataONE

Example: Poor Data Entry, continued

data.xls

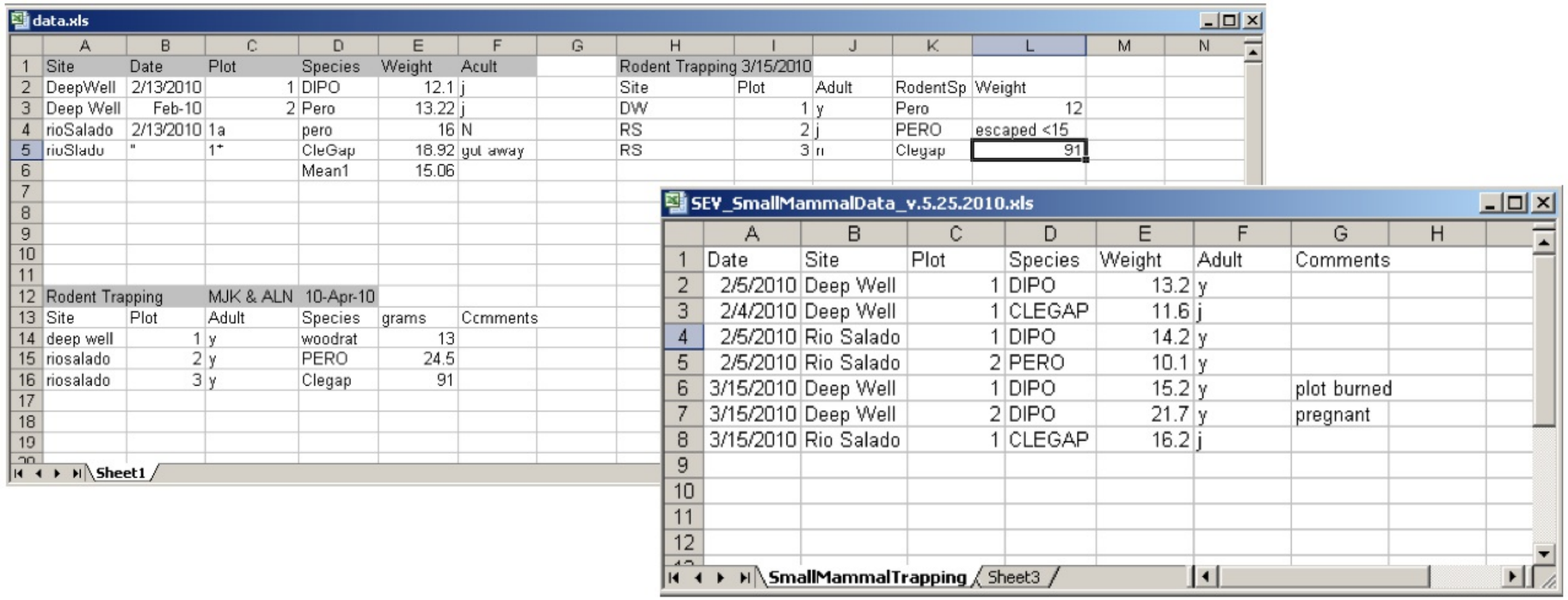
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping	3/15/2010					
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado		1+	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

- **Inconsistency between data collection events**
  - **Different site spellings, capitalization, spaces in site names—hard to filter**
  - **Codes used for site names for some data, but spelled out for others**
  - **Mean1 value is in Weight column**
  - **Text and numbers in same column – what is the mean of 12, “escaped < 15”, and 91?**

Recommended Practices

- Columns of data are consistent: only numbers, dates, or text
- Consistent Names, Codes, Formats (date) used in each column
- Data are all in one table, which is much easier for a statistical program to work with than multiple small tables which each require human intervention



data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping	3/15/2010					
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1*	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

SmallMammalTrapping | Sheet3

Recommended Practices, continued

- Create descriptive column names without spaces or special characters
  - `Soil T30` to `Soil_Temp_30cm`
  - `Species-Code` to "Species_Code (Avoid using -,+,*,^ in column names. Some software may interpret these symbols as an operator)"
- Use a descriptive file name. For instance, a file named `SEV_SmallMammalData_v.5.25.2010.csv` indicates the project the data is associated with (SEV), the theme of the data (SmallMammalData) and also when this version of the data was created (v.5.25.2010). This name is much more helpful than a file named mydata.xls.

Recommended Practices, continued

- Missing data
- Preferably leave field empty (NULL = no value)
- In numeric fields, use a distinct value such as 9999 to indicate a missing value
- In text fields, use NA (“Not Applicable” or “Not Available”)
- Use Data flags in a separate column to qualify missing value

Date	Time	NO3_N_Conc	NO3_N_Conc_Flag
20081011	1300	0.013	
20081011	1330	0.016	
20081011	1400		M1
20081011	1430	0.018	
20081011	1500	0.001	E1

M1 = missing; no sample collected

E1 = estimated from grab sample

Recommended Practices, continued

- Enter complete lines of data

sev_anpp.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	year	Site	Treat	Web	Plot	Quad	Species	winwt	spwt	fallwt	spnpp	fallnpp	anpp	
2	1999	C	U	1	E	1	CHSE7	0	0	0.05	0	0.05	0.05	
3							CHSES	0	0	0.04	0	0.04	0.04	
4							LEFE	0.37	0.17	0	0	0	0	
5							THAC	0	0.45	6.11	0.45	5.66	6.11	
6	1999	C	U	1	E	2	DAPU7	0.01	0.25	0.97	0.24	0.72	0.96	
7							LEFE	3.29	2.01	12.5	0	10.49	10.49	
8							THAC	0	1.21	17.3	1.21	16.08	17.3	
9	1999	C	U	1	E	3	CHSE7	0	0	0.01	0	0.01	0.01	
10							CHSES	0	0	0.01	0	0.01	0.01	
11							LEFE	1.32	0.07	0.7	0	0.64	0.64	
12							THAC	0	0.47	4.43	0.47	3.96	4.43	
13	1999	C	U	1	E	4	CH							
14							DA							
15							LE							
16							TH							
17	1999	C	U	1	N	1	AR							
18							AR							
19							CH							
20							DA							
21							GU							
22	1999	C	U	1	N	2	AR							

Sorting an Excel file with empty cells is not a good idea!

Book3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	year	Site	Treat	Web	Plot	Quad	Species	winwt	spwt	fallwt	spnpp	fallnpp	anpp	
2	1999	C	U	1	N	1	ARIST	0.6	2.75	4.27	2.16	1.52	3.67	
3							ARLUL2	0	0	0.95	0	0.95	0.95	
4	1999	C	U	1	E	1	CHSE7	0	0	0.05	0	0.05	0.05	
5	1999	C	U	1	E	3	CHSE7	0	0	0.01	0	0.01	0.01	
6							CHSES	0	0	0.04	0	0.04	0.04	
7							CHSES	0	0	0.01	0	0.01	0.01	
8	1999	C	U	1	E	4	CHSES	0	0	0.28	0	0.28	0.28	
9							CHSES	0	0	0.02	0	0.02	0.02	
10	1999	C	U	1	E	2	DAPU7	0.01	0.25	0.97	0.24	0.72	0.96	
11							DAPU7	0.05	0.49	0.84	0.44	0.35	0.79	
12							DAPU7	0.06	0.88	2.05	0.82	1.18	1.99	
13							GUSA2	0	0.9	0	0.9	0	0.9	
14							LEFE	0.37	0.17	0	0	0	0	
15							LEFE	3.29	2.01	12.5	0	10.49	10.49	
16							LEFE	1.32	0.07	0.7	0	0.64	0.64	
17							LEFE	2.9	0.4	0.12	0	0	0	
18							THAC	0	0.45	6.11	0.45	5.66	6.11	
19							THAC	0	1.21	17.3	1.21	16.08	17.3	
20							THAC	0	0.47	4.43	0.47	3.96	4.43	
21							THAC	0	1.5	17.26	1.5	15.76	17.26	

Sheet1 Sheet2 Sheet3

Best Practices

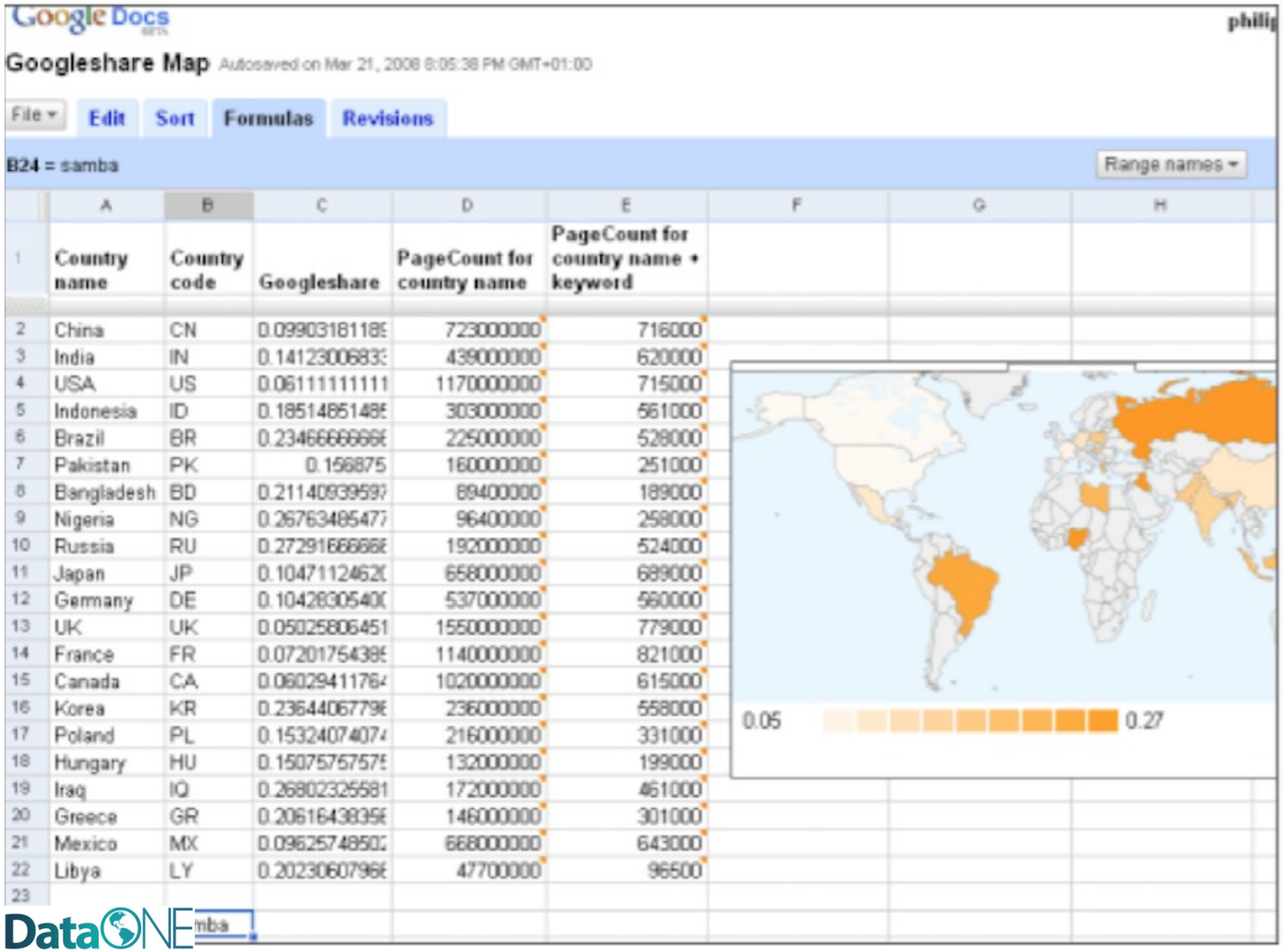
- For the long term, store data in a consistent format that can be read well in to the future and that can be used by any application now or in the future
- Appropriate file types include:
  - Non-proprietary: use an open, documented standard
  - Common usage by research community: Standard representation (ASCII, Unicode)
  - Unencrypted
  - Uncompressed
- ASCII formatted files are likely to be readable into the future
  - Use ASCII (comma-separated) for tabular data

Resources

- Best Practices for Preparing Environmental Data Sets to Share and Archive. September 2010. Les A. Hook, Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. https://daac.ornl.gov/PI/BestPractices-2010.pdf
- Preparing Data for Sharing. 2015. Libbie Stephenson. https://doi:10.7910/DVN/BJNXVQ

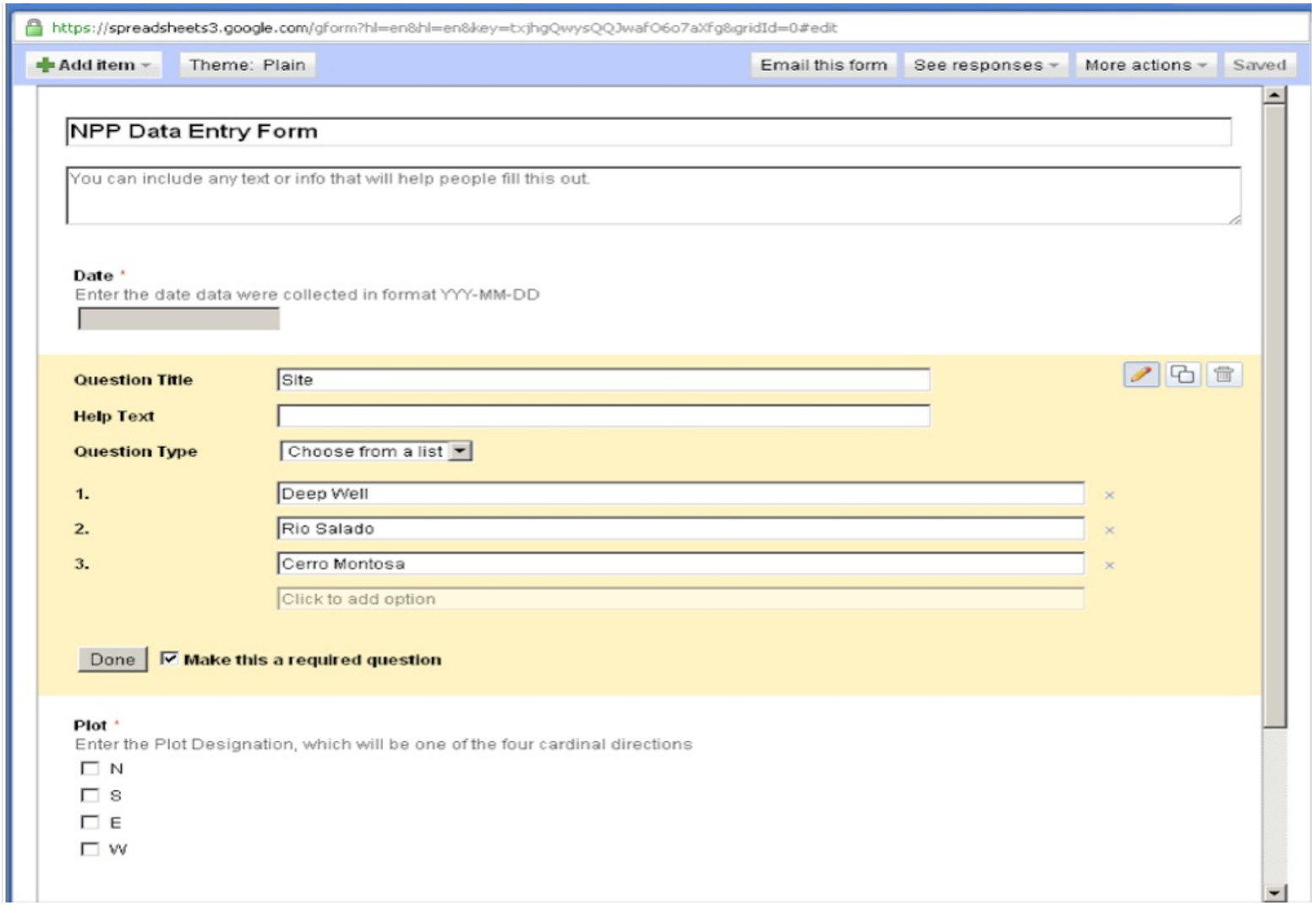
Data Entry Tools

- Two common tools: Google Docs, Excel

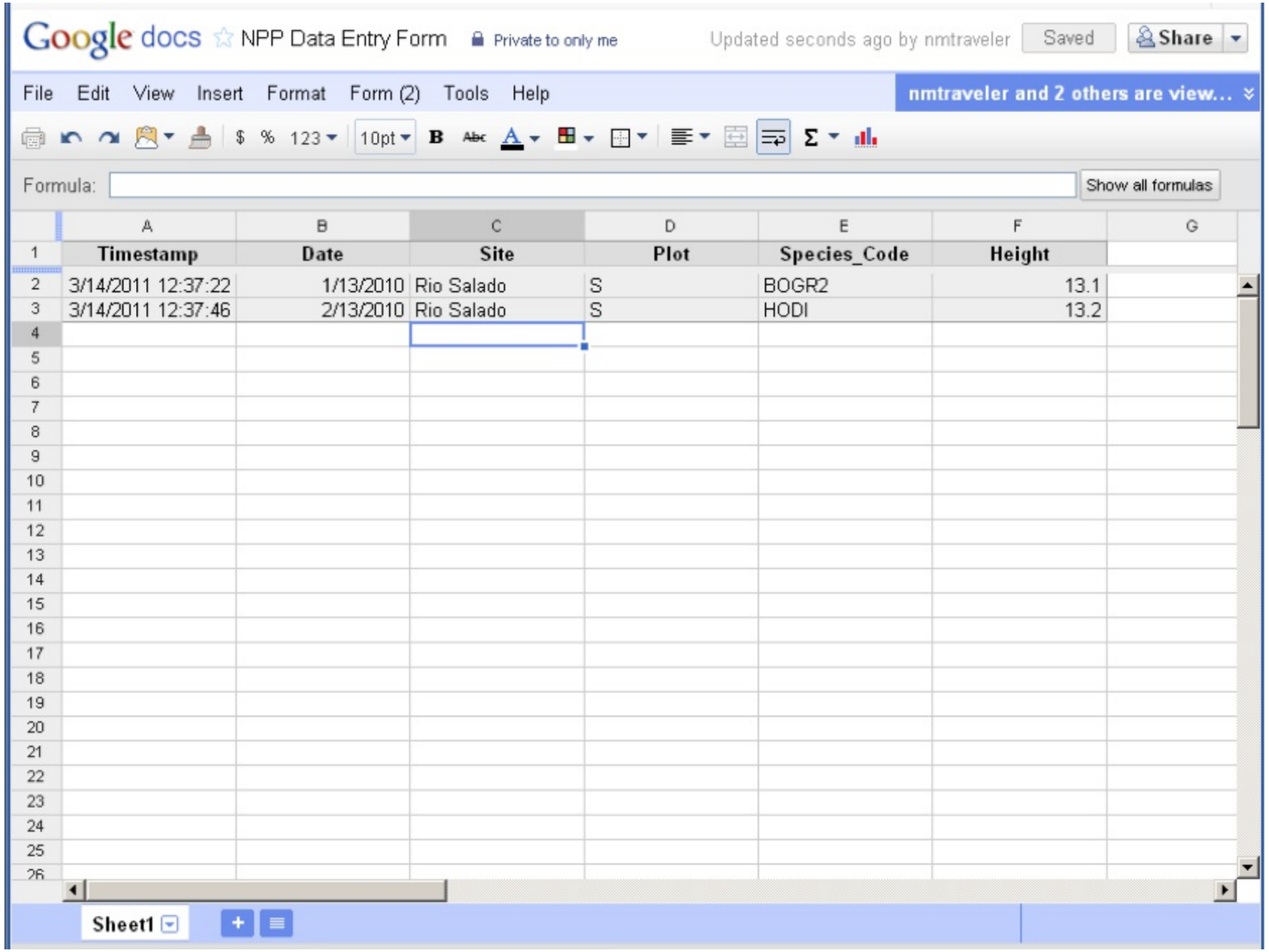


	A	B	C	D	E
1	Country name	Country code	Googleshare	PageCount for country name	PageCount for country name + keyword
2	China	CN	0.0990318118	723000000	716000
3	India	IN	0.1412300683	439000000	620000
4	USA	US	0.0611111111	1170000000	715000
5	Indonesia	ID	0.1851485148	303000000	561000
6	Brazil	BR	0.2346666666	225000000	528000
7	Pakistan	PK	0.156875	160000000	251000
8	Bangladesh	BD	0.2114093959	89400000	189000
9	Nigeria	NG	0.2676348547	96400000	258000
10	Russia	RU	0.2729166666	192000000	524000
11	Japan	JP	0.1047112462	658000000	689000
12	Germany	DE	0.1042830540	537000000	560000
13	UK	UK	0.0502580645	1550000000	779000
14	France	FR	0.0720175438	1140000000	821000
15	Canada	CA	0.0602941176	1020000000	615000
16	Korea	KR	0.2364406779	236000000	558000
17	Poland	PL	0.1532407407	216000000	331000
18	Hungary	HU	0.1507575757	132000000	199000
19	Iraq	IQ	0.2680232558	172000000	461000
20	Greece	GR	0.2061643835	146000000	301000
21	Mexico	MX	0.0962574850	668000000	643000
22	Libya	LY	0.2023060796	47700000	96500
23					

Google Docs Forms

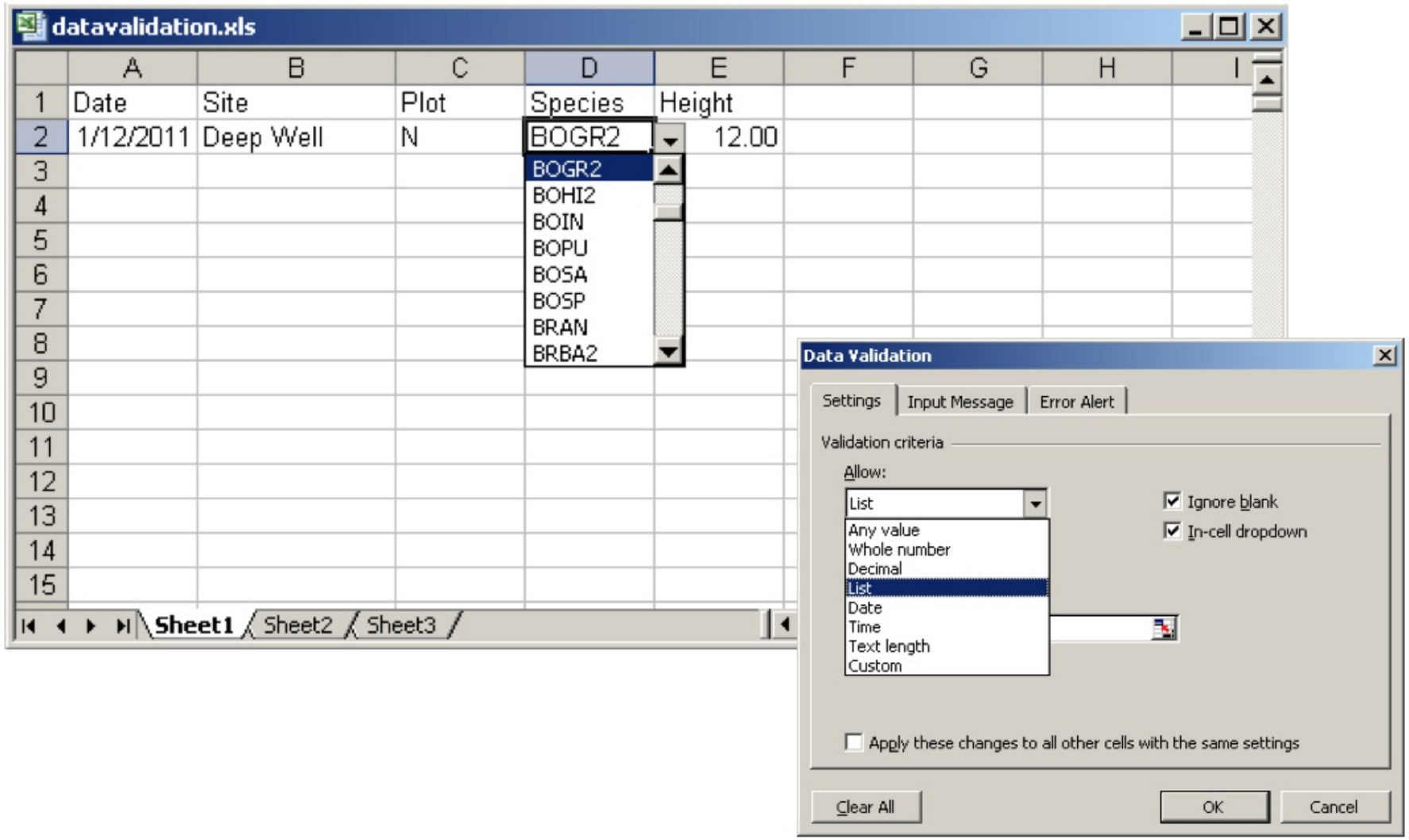


Google Docs Spreadsheet

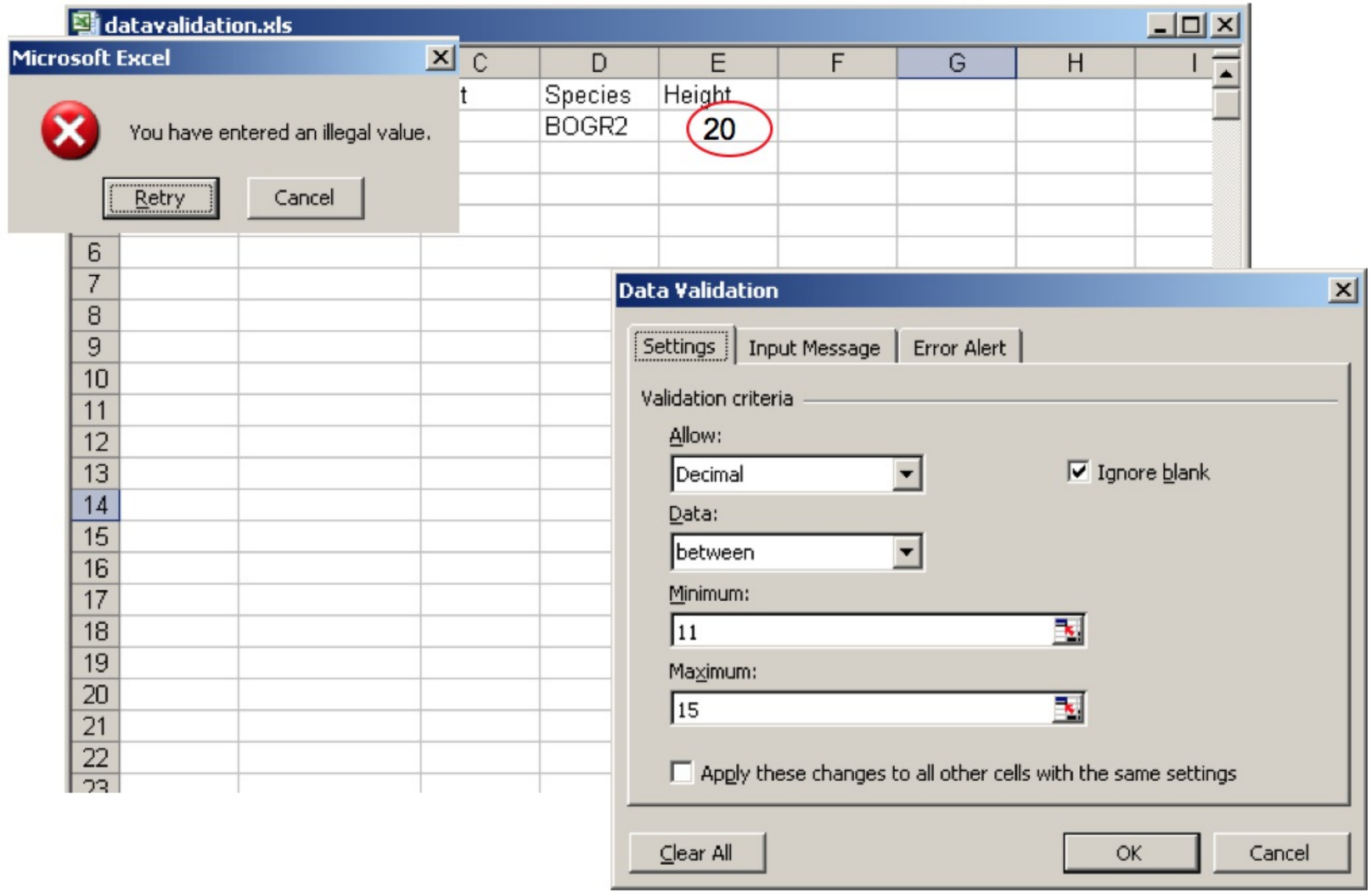


	A	B	C	D	E	F	G
1	Timestamp	Date	Site	Plot	Species_Code	Height	
2	3/14/2011 12:37:22	1/13/2010	Rio Salado	S	BOGR2	13.1	
3	3/14/2011 12:37:46	2/13/2010	Rio Salado	S	HODI	13.2	
4							

Excel


datavalidation.xls
A
B
C
D
E
F
G
H
I
1
Date
Site
Plot
Species
Height
2
1/12/2011
Deep Well
N
BOGR2
12.00
3
4
5
6
7
8
9
10
11
12
13
14
15
BOGR2
BOHI2
BOIN
BOPU
BOSA
BOSP
BRAN
BRBA2
Sheet1
Sheet2
Sheet3
Data Validation
Settings
Input Message
Error Alert
Validation criteria
Allow:
List
Ignore blank
In-cell dropdown
Any value
Whole number
Decimal
List
Date
Time
Text length
Custom
Apply these changes to all other cells with the same settings
Clear All
OK
Cancel

Excel: Data Validation

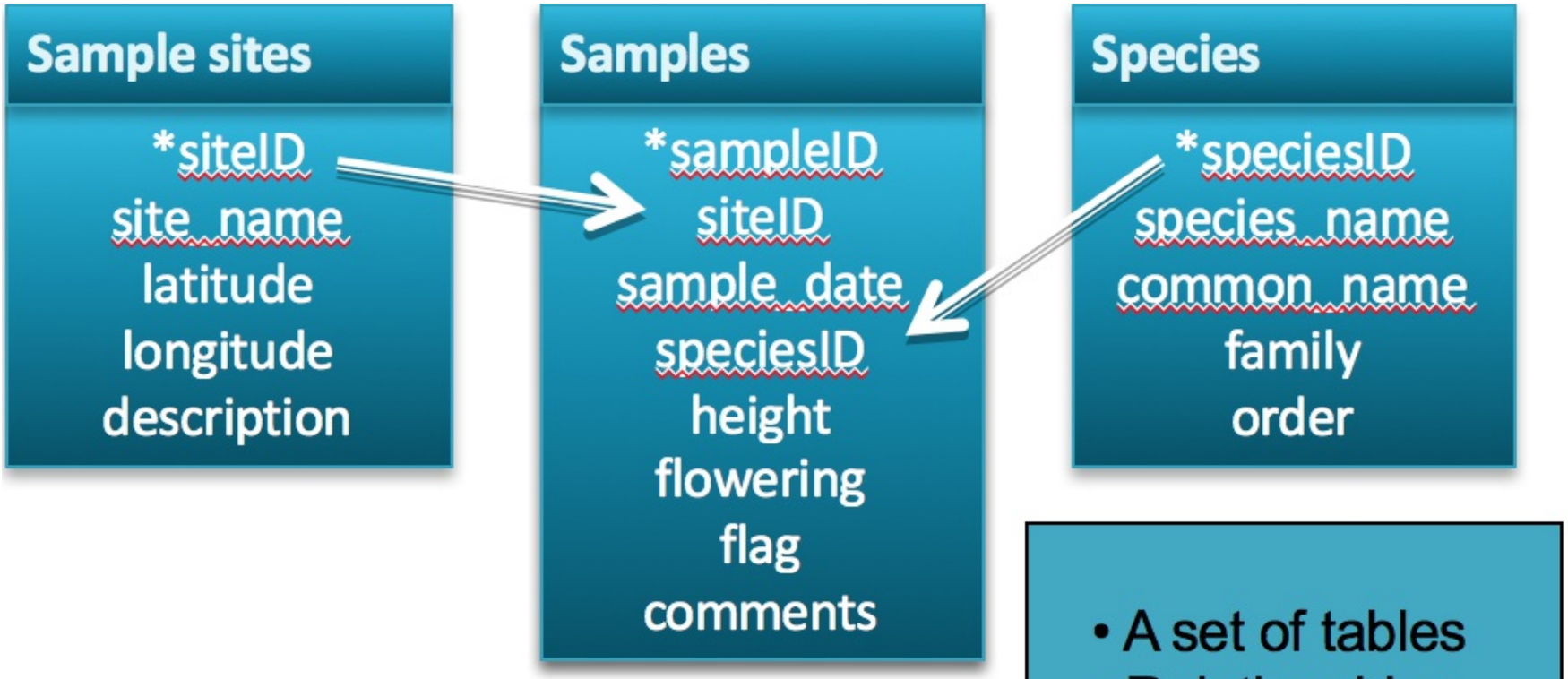


Spreadsheet versus Relational Database

- Great for charts, graphs, calculations
- Flexible about cell content type—cells in same column can contain numbers or text ** Easy to use – but harder to maintain as complexity and size of data grows

- Easy to query to select portions of data
- Data fields are typed – For example, only integers are allowed in integer fields
- Columns cannot be sorted independently of each other
- Steeper learning curve than a spreadsheet

What is a relational database?



- A set of tables
- Relationships
- A command language

Database Features: Explicit control over data types

Date	Site	Height	Flowering
<dates only>	<text only>	< real numbers only>	< ‘y’ and ‘n’ only>

Advantages
- **quality control**
- **performance**

Relationships are defined between tables

Date	Site	Species	Flowering?
2/13/2010	A	BOGR2	y
2/13/2010	B	HODR	y
4/15/2010	B	BOER4	y
4/15/2010	C	PLJA	n

Site	Latitude	Longitude
A	34.1	-109.3
B	35.2	-108.6
C	32.6	-107.5

Mix and Match data on the fly

Date	Site	Species	Flowering?	Latitude	Longitude
2/13/2010	A	BOGR2	y	34.1	-109.3
2/13/2010	B	HODR	y	35.2	-108.6
4/15/2010	B	BOER4	y	35.2	-108.6
4/15/2010	C	PLJA	n	32.6	-107.5

Using Structured Query Language (SQL)

This table is called SoilTemp

Date	Plot	Treatment	SensorDepth	Soil_Temperature
2010-02-01	C	R	30	12.8
2010-02-01	B	C	10	13.2
2010-02-02	C	R	0	6.3
2010-02-02	A	N	0	15.1

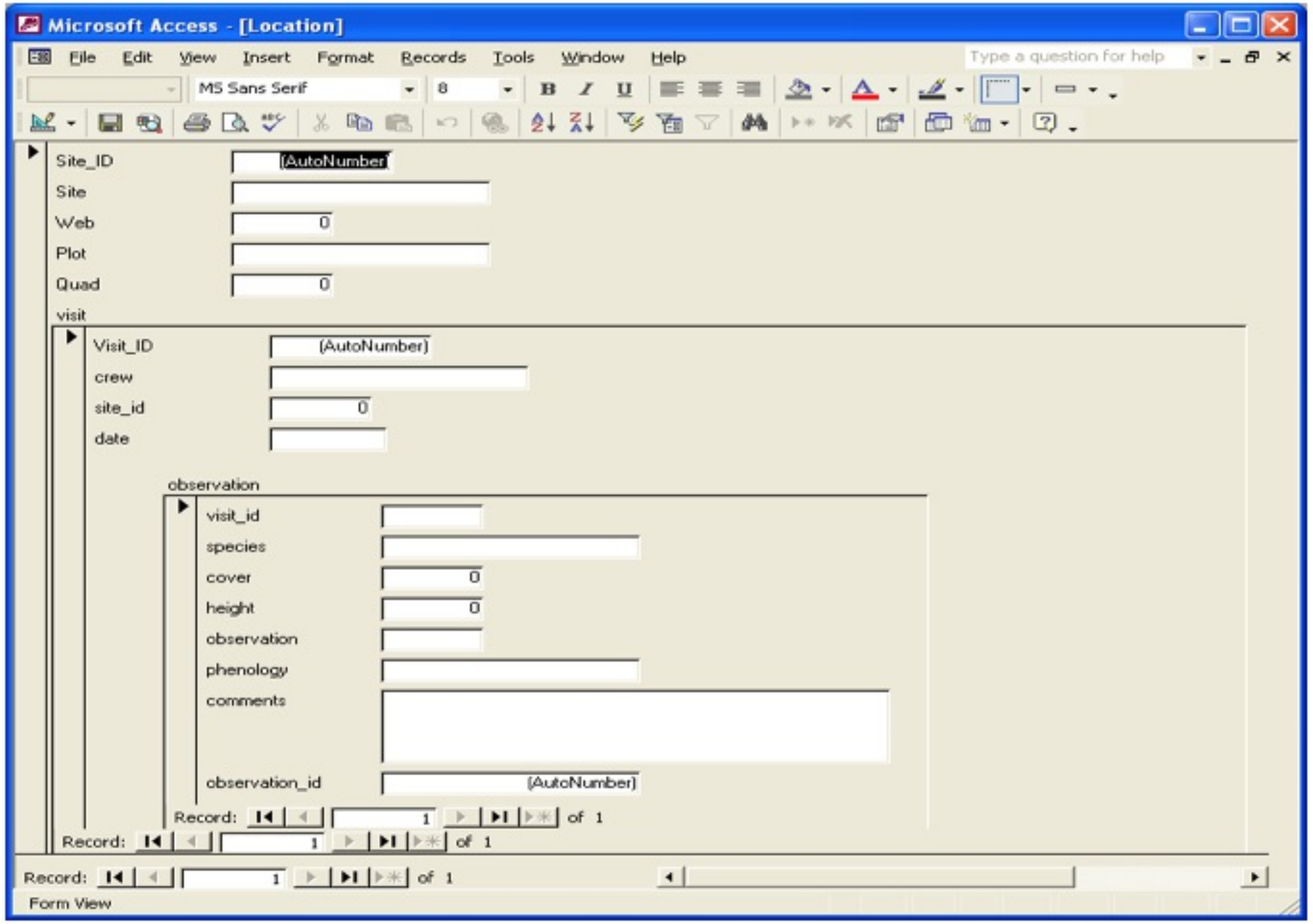
SQL examples: Select Date, Plot, Treatment, SensorDepth, Soil_Temperature from SoilTemp where Date = '2010-02-01'

Date	Plot	Treatment	SensorDepth	Soil_Temperature
2010-02-01	C	R	30	12.8
2010-02-01	B	C	10	13.2

Select * from SoilTemp where Treatment='N' and SensorDepth='0'

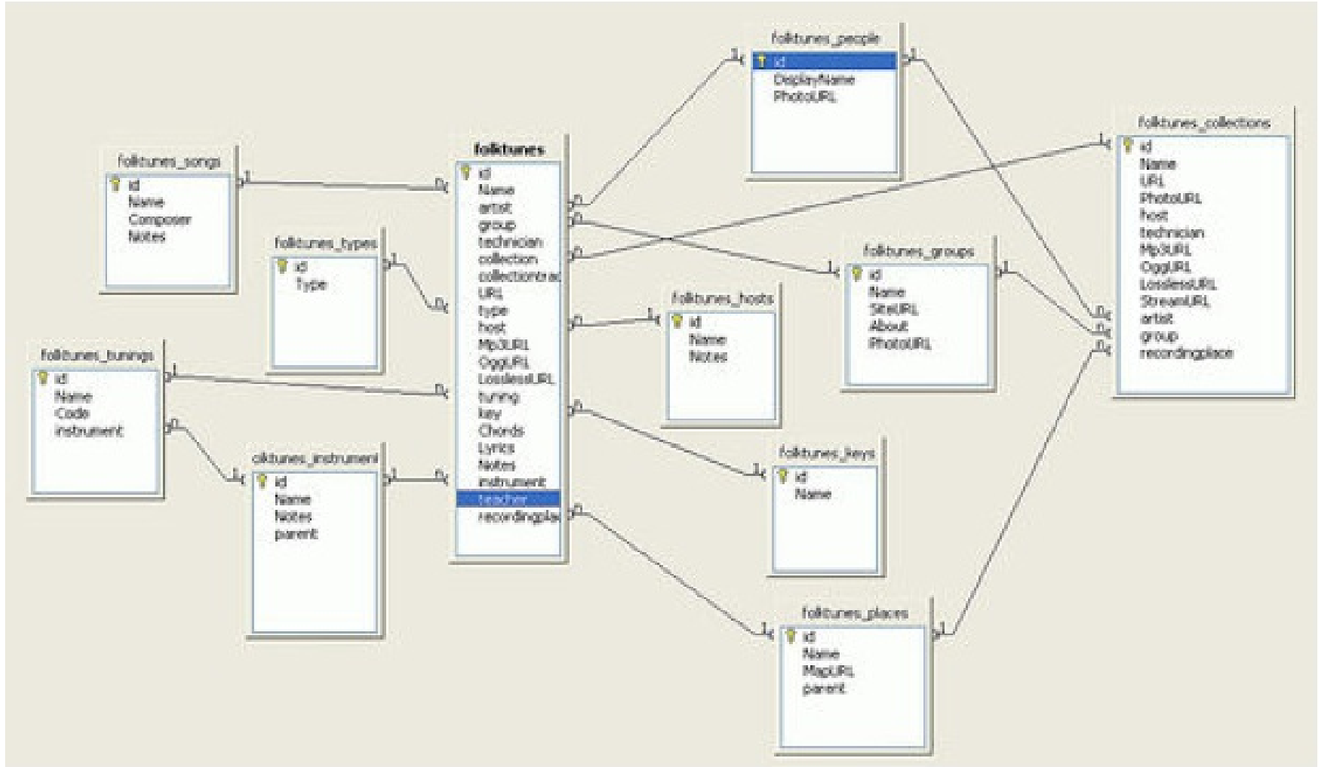
Date	Plot	Treatment	SensorDepth	Soil_Temperature
2010-02-02	A	N	0	15.1

Data entry using a database



Review: Planning for Data Entry

- Be aware of Best Practices in your domain when designing data file structures
- Choose a data entry method that allows some validation of data as it is entered
- Consider investing time in learning how to use a database if datasets are large or complex



If you want to try a database:

Consider trying one of these:

- Personal, single-user databases can be developed in MS Access, which is stored as a file on the user's computer. MS Access comes with easy GUI tools to create databases, run queries, and write reports.
- A more robust database that is free, accommodates multiple users and will run on Windows or Linux is MySQL. GUI interfaces for MySQL include phpMyadmin (free) and Navicat (inexpensive).

To learn more about designing a relational database:

- Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design (2nd Edition). Michael J. Hernandez. Addison-Wesley, 2003.
- Fundamentals of Relational Database Design. Paul Litwin. http://r937.com/relational.html. (Accessed May 12, 2016).

Data Integration Best Practices

- Maintain dataset provenance
  - Document transformations
  - Beware of accidental duplication
- Review metadata for compatibility of context, methods, and meaning
  - For what purpose was the data collected?
  - How was the data collected?
  - Is it sensible to combine these datasets?

Data Integration Best Practices

- Ensure compatibility
  - Convert to common units
  - Choose appropriate numeric precision
  - Evaluate and standardize missing value codes
- Document all assumptions
  - What assumptions underlie the original datasets?
  - What assumptions did you make in combining the datasets?

Data Integration Best Practices

- Recognize that you are creating a new dataset
  - Revisit the data life cycle to ensure the new dataset is properly documented, validated, and preserved
- Use reproducible workflows
  - Enable transparency and reproducibility in the integration process
  - Ensure others understand and can evaluate your decision making process.
  - Automate the integration as much as possible, especially when integrating many or large datasets

Data Integration Best Practices

- Ensure attribution of original dataset owners and respect data usage agreements
- Example resource:

  Jones et al. (2006) The New Bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology and Systematics 37:519-544
- Example citation to the related dataset from the Dryad repository:

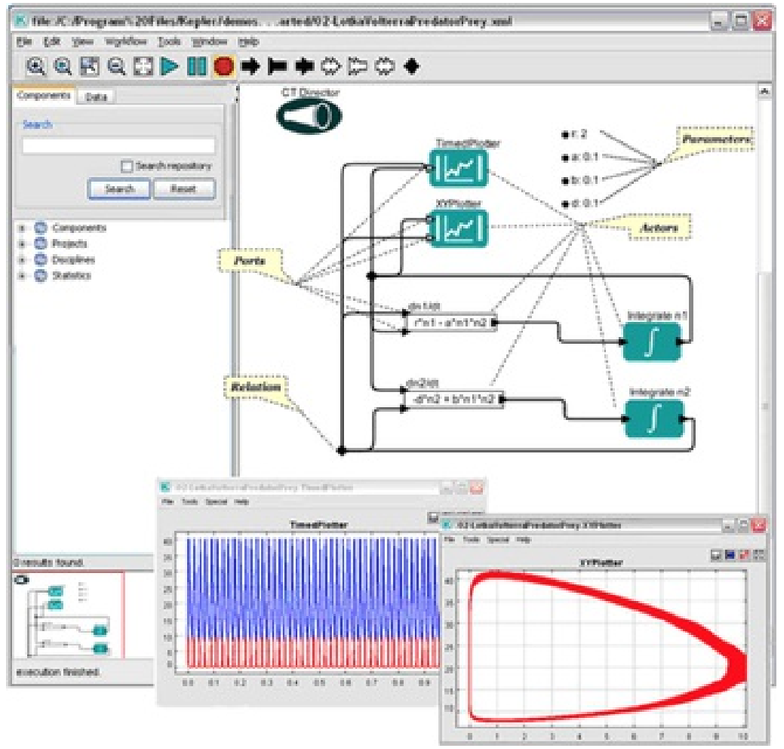
  Jones, Matthew B., Schildhauer, Mark P., Reichman, O. J., and Bowers, Shawn. 2012. Data from "The new bioinformatics: integrating ecological data from the gene to the biosphere." Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.qb0d6?ver=2012-07-16T14:42:48.559-04:00.

Data Manipulation

- Useful for analyzing, subsetting and transforming data
- Can be used to check and assure quality data
- Options include SAS, SPSS, R, and Matlab (not free)
  - SAS: Has comprehensive support
  - SPSS: Has a user-friendly GUI
  - Matlab: Analysis and Visualization platform that has “toolboxes” available for different disciplines, such as modeling or genomic analyses

Using R

- Free (http://www.r-project.org/index.html)
- Produces publication quality graphics
- Lots of forums from which to get help
- Software (such as Kepler for developing workflows) will integrate analytical components written in R



Review: Selecting tools for data storage and use

- Tools such as (but not limited to) spreadsheet tools such as MS Excel and relational databases (MS Access, MySQL, and more) can provide structure, flexibility and potential for working more easily with datasets but also require planning
- Selection of a database or spreadsheet tool depends on the relationships between the data, and how it will be used, as well as other considerations re: time, resources, output.

Review: Data Integration & Manipulation

- Maintaining provenance (a trail of custody and decisions) is important when integrating more than one dataset
- Documenting and understanding context and relationships, as well as changes is crucial when creating a new dataset (any time you combine two or more disparate datasets)
- Create a transparent, reproducible workflow
- Make sure to provide proper attribution and citation to all resources, including the original dataset.
- Tools such as R, Matlab, and others can be useful in establishing workflows and accessing datasets

About

Participate in our GitHub repo: https://dataoneorg.github.io/dataone_lessons/

Suggested citation: DataONE Education Module: Data Management. DataONE. Retrieved November 12, 2016. From https://dataoneorg.github.io/dataone_lessons/