Data Cleaning for Analysis and Publication Using the OpenRefine Software Package

Arctic Data Center, CUAHSI, DataONE, Environmental Data Initiative, GBIF, **iDigBio**, NEON, Neotoma

Jeanette Clark, Deborah Paul

#datahelpdesk
Ecological Society of America 2020
Career Central
August
http://bit.ly/datahelpesa2020

NSF Arctic Data Center

The Arctic Data Center

https://arcticdata.io

NSF NCEAS

Data Archive

Portal for data discovery

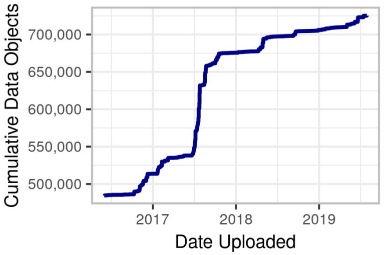
Tools for data and metadata submission

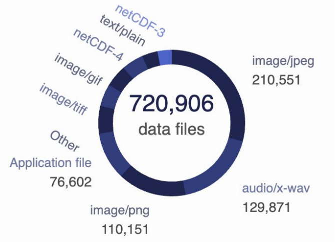
Support services

Training and Outreach

Data Rescue

5,800 DATA SETS	1,500 CREATORS
725K DATA FILES	14,000 USERS
34 TB DATA STORAGE	990K+ FILE DOWNLOADS





@arcticdatactr

NSF Arctic Data Center · CUAHSI Universities Allied for Water Research · DataONE · EDI · ESIP Making Data Matter · GBIF · iDigBio Integrated Digitized Biocollections · NSF · neon Operated by Battelle · Neotoma Paleoecology Database

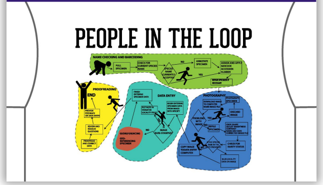
Advancing the Digitization of Biological Collections

iDigBio Hub and Thematic (Museum) Collection Networks

(Meta)data Aggregator

Community Building

Collections Data



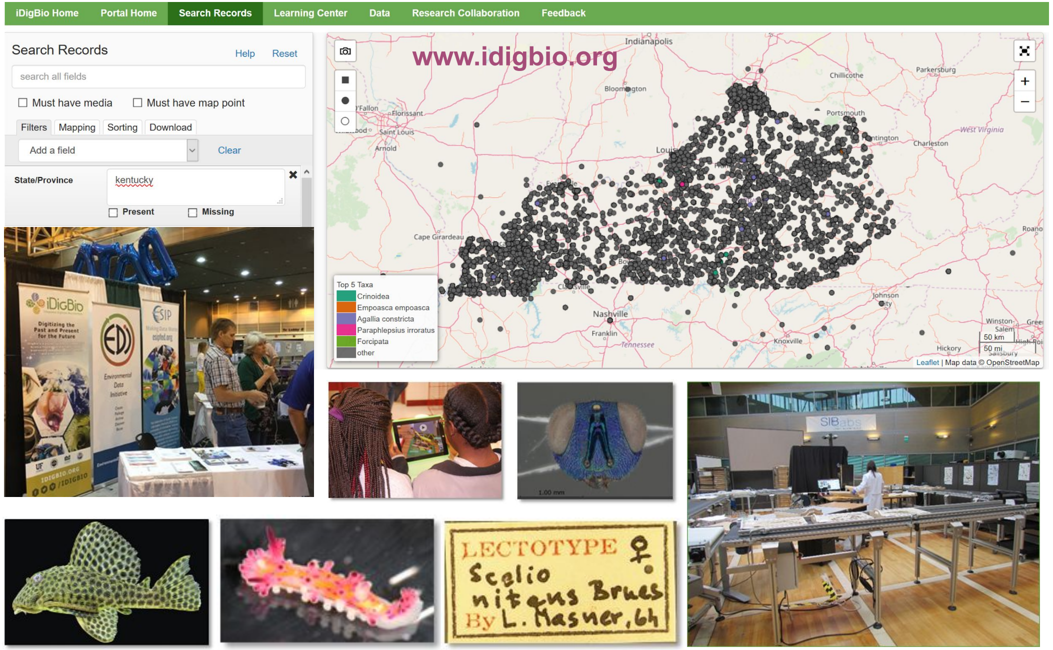
people graphic by Dorothy Allard

Digitization
Workflows & Protocols
Dissemination

Research Use
Cyberinfrastructure
Tool collaboration
Portal development
ENM workshop
Research focus
Data quality
APIs

Training
Biodiversity informatics
Data skills and literacy
Collections software
Imaging
Project Management

total: 121,428,342


iDigBio Home | Portal Home | Search Records | Learning Center | Data | Research Collaboration | Feedback

Search Records — Help — Reset

search all fields

☐ Must have media ☐ Must have map point

Filters | Mapping | Sorting | Download

Add a field — Clear

State/Province: kentucky ✖
☐ Present ☐ Missing

www.idigbio.org

Top 5 Taxa
- 🟩 Crinoidea
- 🟧 Empoasca empoasca
- 🟪 Agallia constricta
- 🟥 Paraphlepsius irroratus
- 🟩 Forcipata
- ⬛ other

Leaflet | Map data © OpenStreetMap

Education Outreach
Citizen Science
K-12 materials
Undergraduate
Fossil Clubs
Mentor teachers

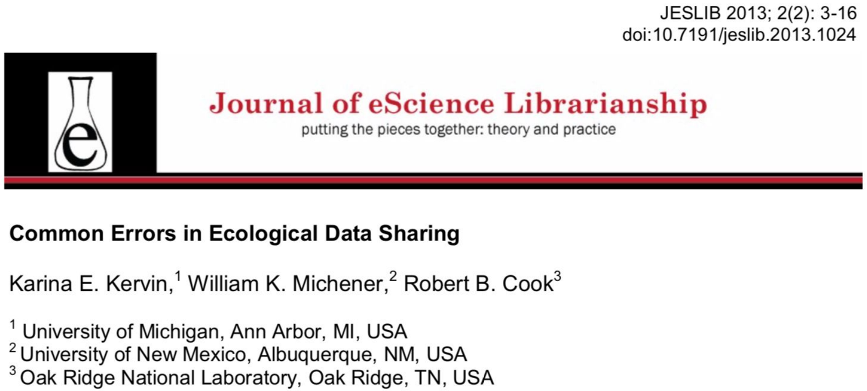
Methods
Workshops
Webinars
Symposia
Conferences
Working Groups
Short Courses
Adobe Connect
Listservs
Publications
Social Media **@idigbio**

NSF Arctic Data Center · CUAHSI Universities Allied for Water Research · DataONE · EDI · ESIP Making Data Matter · GBIF · iDigBio Integrated Digitized Biocollections · NSF · neon Operated by Battelle · Neotoma Paleoecology Database

What do we mean by “Clean” Data?

More Data from More Sources =

- Structural Issues
- Inconsistent/unclear missing values
- Mixed data in single columns
- Mixed data types in single columns
- Ambiguous data values
- Data you can’t use



JESLIB 2013; 2(2): 3-16
doi:10.7191/jeslib.2013.1024

Journal of eScience Librarianship
putting the pieces together: theory and practice

Common Errors in Ecological Data Sharing

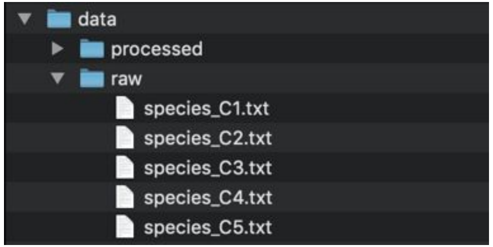
Karina E. Kervin,[1] William K. Michener,[2] Robert B. Cook[3]

[1] University of Michigan, Ann Arbor, MI, USA
[2] University of New Mexico, Albuquerque, NM, USA
[3] Oak Ridge National Laboratory, Oak Ridge, TN, USA

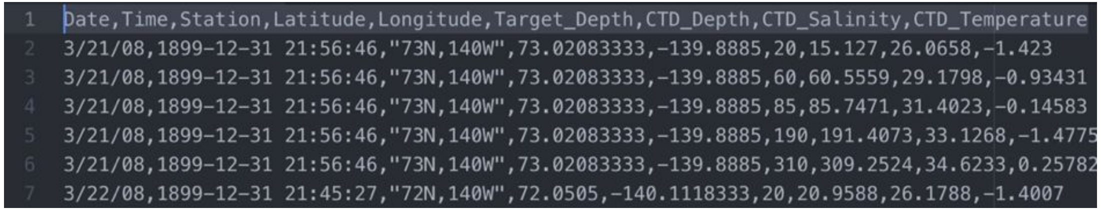
Simple guidelines for data management

- Use a scripted program
- Nonproprietary formats
- Keep a raw version of data
- Descriptive names
- Header line
- Plain ASCII text

.csv, .txt



```
data
  processed
  raw
    species_C1.txt
    species_C2.txt
    species_C3.txt
    species_C4.txt
    species_C5.txt
```



```
Date,Time,Station,Latitude,Longitude,Target_Depth,CTD_Depth,CTD_Salinity,CTD_Temperature
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,20,15.127,26.0658,-1.423
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,60,60.5559,29.1798,-0.93431
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,85,85.7471,31.4023,-0.14583
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,190,191.4073,33.1268,-1.4775
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,310,309.2524,34.6233,0.25782
3/22/08,1899-12-31 21:45:27,"72N,140W",72.0505,-140.1118333,20,20.9588,26.1788,-1.4007
```

Borer, E. T. et al, (2009), Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America

Simple Guidelines for Data Management

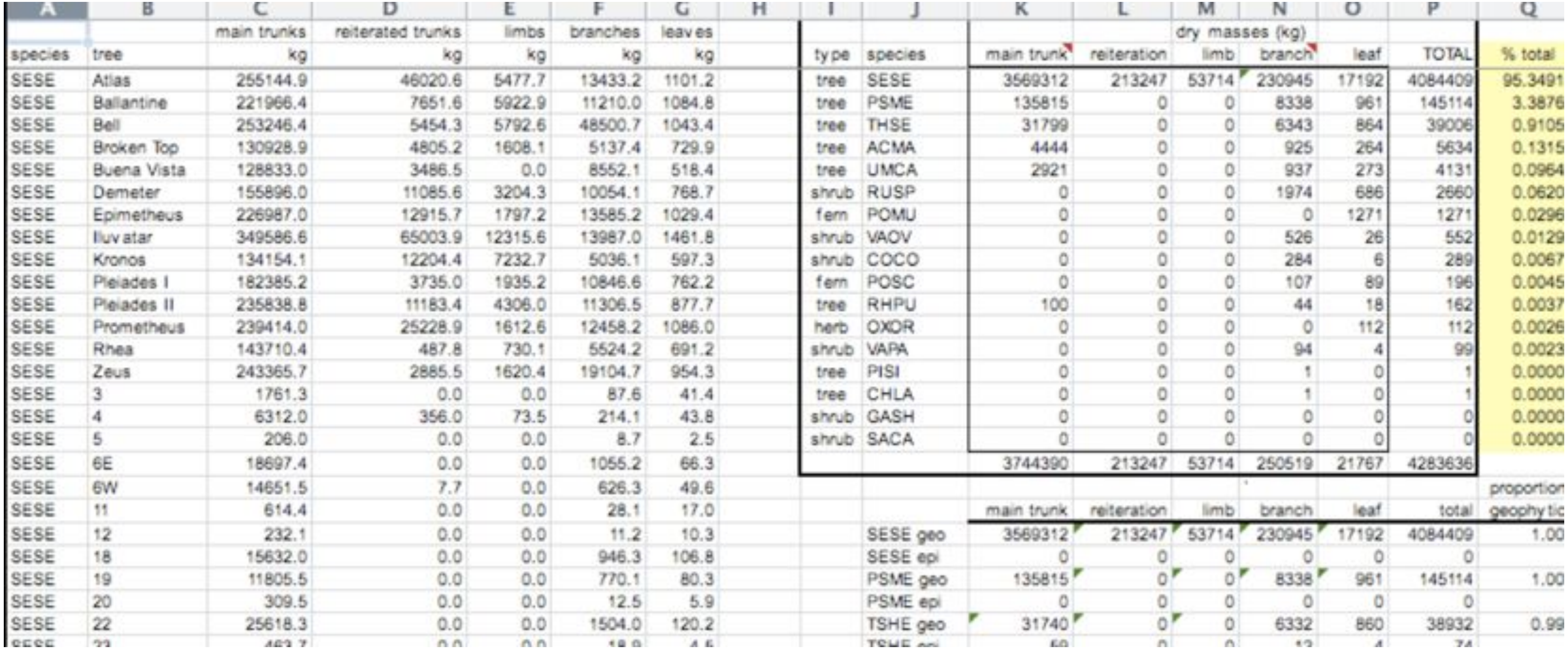
- Design to add rows, not columns

- Each column should contain only one type of information

- Record a single piece of data only once; separate information collected at different scales into different tables. In other words, create a relational database.

Borer, E. T. et al, (2009), Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America

Recognizing untidy data



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leaves						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2		tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8		tree	PSME	135815	0	0	8338	961	145114	3.3876
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4		tree	THSE	31799	0	0	6343	864	39006	0.9105
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9		tree	ACMA	4444	0	0	925	264	5634	0.1315
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4		tree	UMCA	2921	0	0	937	273	4131	0.0964
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7		shrub	RUSP	0	0	0	1974	686	2660	0.0620
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4		fern	POMU	0	0	0	0	1271	1271	0.0296
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8		shrub	VAOV	0	0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3		shrub	COCO	0	0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2		fern	POSC	0	0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7		tree	RHPU	100	0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0		herb	OXOR	0	0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2		shrub	VAPA	0	0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	954.3		tree	PISI	0	0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87.6	41.4		tree	CHLA	0	0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214.1	43.8		shrub	GASH	0	0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8.7	2.5		shrub	SACA	0	0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055.2	66.3				3744390	213247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3	49.6										proportion
SESE	11	614.4	0.0	0.0	28.1	17.0				main trunk	reiteration	limb	branch	leaf	total	geophytic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	69	0	0	13	4	74	

Characteristics of Tidy Data

Observations

- **Separate tables for each entity measured**

Recognizing untidy data

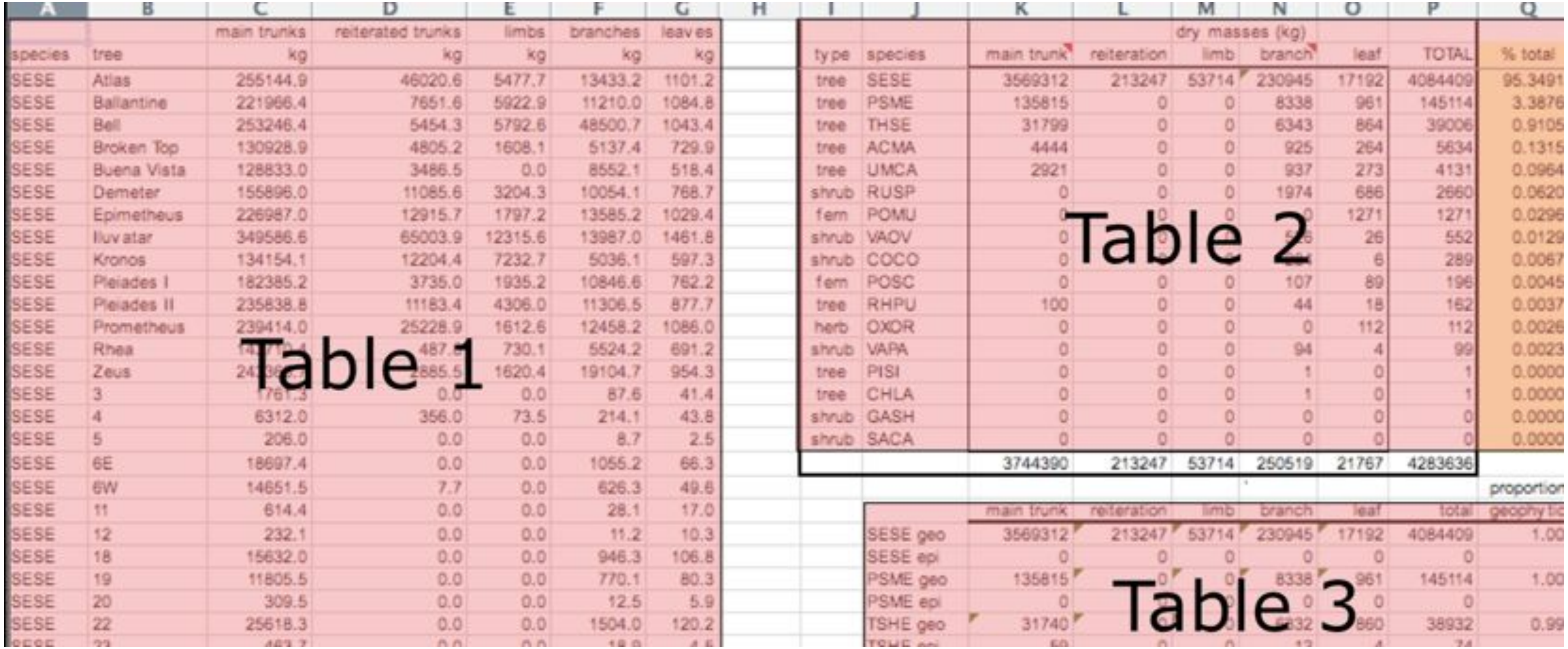


Table 1

		main trunks	reiterated trunks	limbs	branches	leaves
species	tree	kg	kg	kg	kg	kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0
SESE	Rhea	[illegible]	[illegible]	730.1	5524.2	691.2
SESE	Zeus	[illegible]	[illegible]	1620.4	19104.7	954.3
SESE	3	1761.3	0.0	0.0	87.6	41.4
SESE	4	6312.0	356.0	73.5	214.1	43.8
SESE	5	206.0	0.0	0.0	8.7	2.5
SESE	6E	18697.4	0.0	0.0	1055.2	66.3
SESE	6W	14651.5	7.7	0.0	626.3	49.6
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0	0.0	0.0	946.3	106.8
SESE	19	11805.5	0.0	0.0	770.1	80.3
SESE	20	309.5	0.0	0.0	12.5	5.9
SESE	22	25618.3	0.0	0.0	1504.0	120.2
SESE	23	463.7	0.0	0.0	18.9	4.5

Table 2

				dry masses (kg)				
type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
fern	POMU	0	0	0	0	1271	1271	0.0296
shrub	VAOV	0	0	[illegible]	526	26	552	0.0129
shrub	COCO	0	[illegible]	[illegible]	[illegible]	6	289	0.0067
fern	POSC	0	0	0	107	89	196	0.0045
tree	RHPU	100	0	0	44	18	162	0.0037
herb	OXOR	0	0	0	0	112	112	0.0026
shrub	VAPA	0	0	0	94	4	99	0.0023
tree	PISI	0	0	0	1	0	1	0.0000
tree	CHLA	0	0	0	1	0	1	0.0000
shrub	GASH	0	0	0	0	0	0	0.0000
shrub	SACA	0	0	0	0	0	0	0.0000
		3744390	213247	53714	250519	21767	4283636	

Table 3

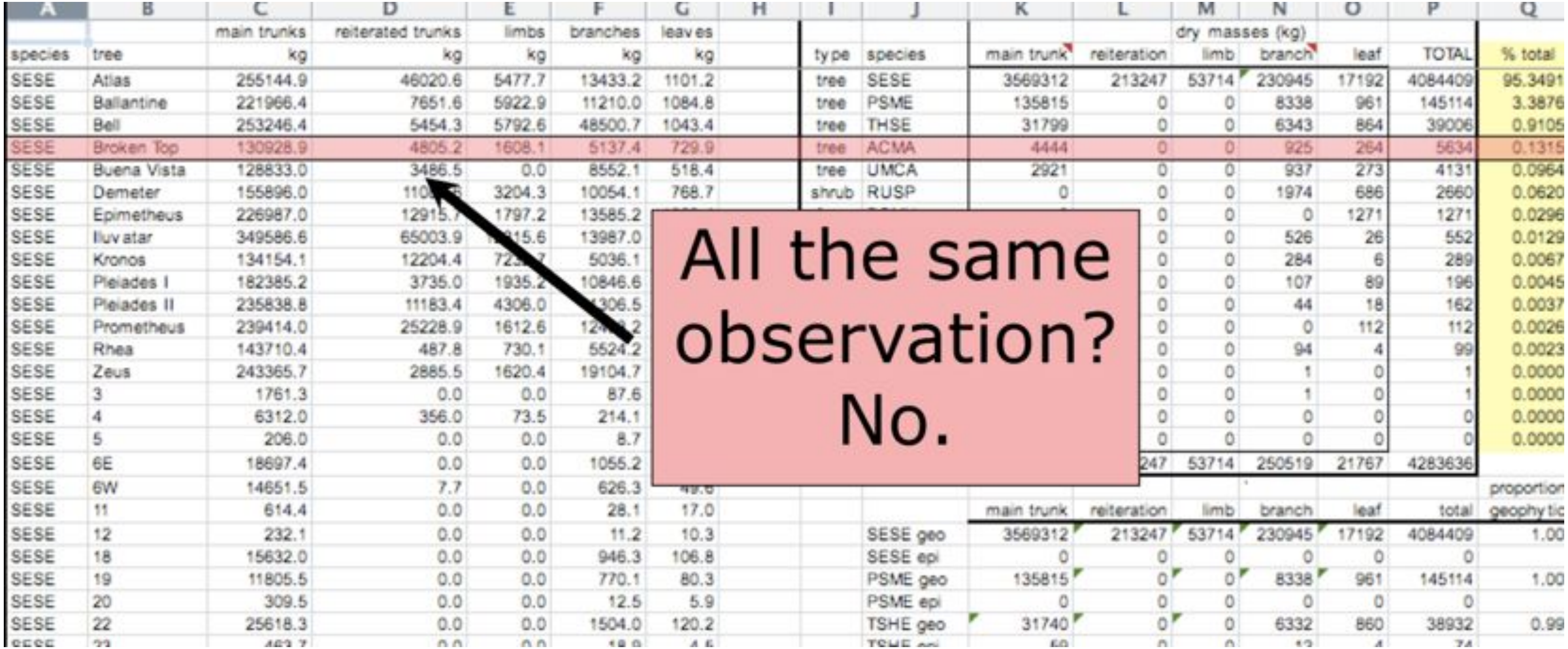
							proportion
	main trunk	reiteration	limb	branch	leaf	total	geophytic
SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE epi	0	0	0	0	0	0	
PSME geo	135815	0	0	8338	961	145114	1.00
PSME epi	0	[illegible]	[illegible]	0	0	0	
TSHE geo	31740	[illegible]	[illegible]	6832	860	38932	0.99
TSHE epi	60	0	0	12	4	74	

Characteristics of Tidy Data

Observations

- Separate tables for each entity measured
- **Each row represents a single observed entity**

Recognizing untidy data



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leav es						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		ty pe	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2		tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8		tree	PSME	135815	0	0	8338	961	145114	3.3876
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4		tree	THSE	31799	0	0	6343	864	39006	0.9105
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9		tree	ACMA	4444	0	0	925	264	5634	0.1315
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4		tree	UMCA	2921	0	0	937	273	4131	0.0964
SESE	Demeter	155896.0	[illegible]	3204.3	10054.1	768.7		shrub	RUSP	0	0	0	1974	686	2660	0.0620
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2						0	0	0	1271	1271	0.0296
SESE	Iluv atar	349586.6	65003.9	[illegible]	13987.0						0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	[illegible]	5036.1						0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6						0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	[illegible]						0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	[illegible]						0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524.2						0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104.7						0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87.6						0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214.1						0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8.7						0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055.2						247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3											proportion
SESE	11	614.4	0.0	0.0	28.1	17.0				main trunk	reiteration	limb	branch	leaf	total	geophy tic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	69	0	0	13	4	74	

All the same observation?

No.

Characteristics of Tidy Data

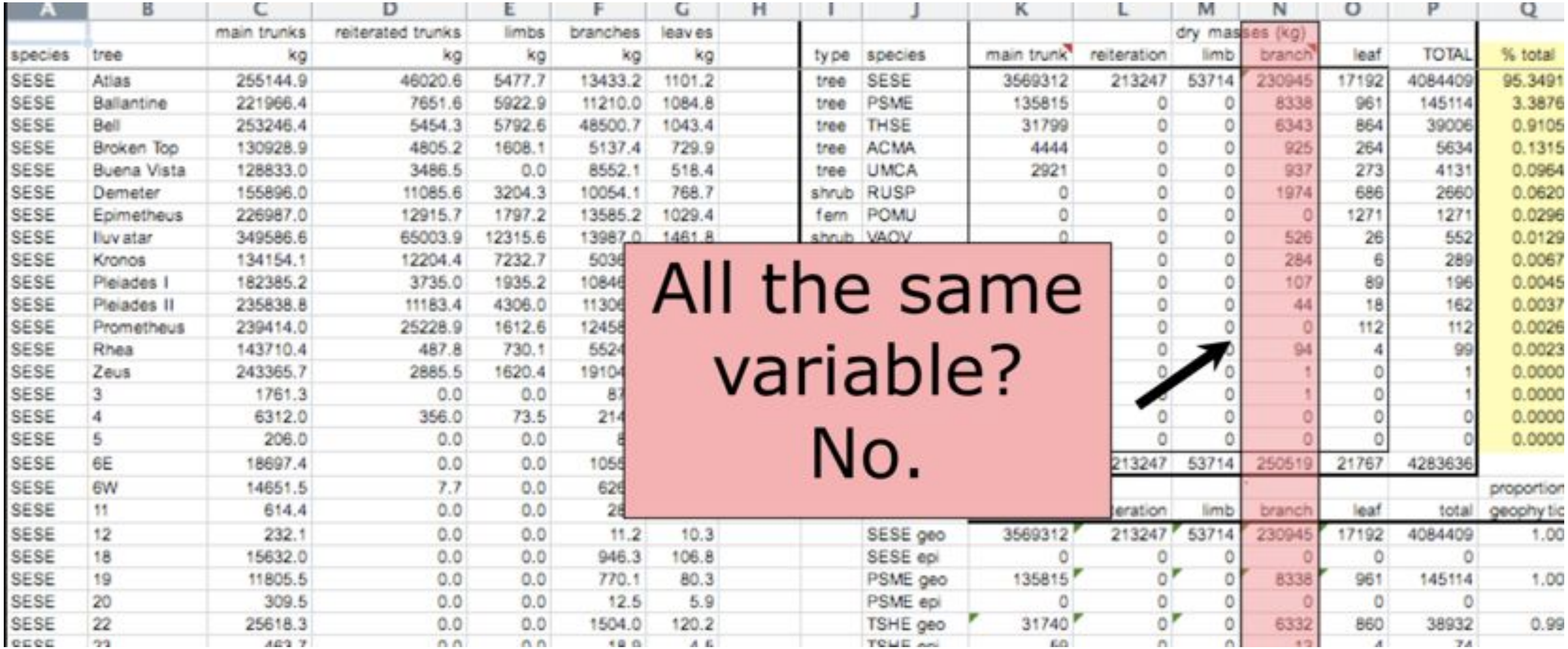
Observations

- Separate tables for each entity measured
- Each row represents a single observed entity

Variables

- **All values in a column are of the same type**

Recognizing untidy data



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leaves						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2		tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8		tree	PSME	135815	0	0	8338	961	145114	3.3876
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4		tree	THSE	31799	0	0	6343	864	39006	0.9105
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9		tree	ACMA	4444	0	0	925	264	5634	0.1315
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4		tree	UMCA	2921	0	0	937	273	4131	0.0964
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7		shrub	RUSP	0	0	0	1974	686	2660	0.0620
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4		fern	POMU	0	0	0	0	1271	1271	0.0296
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8		shrub	VAOV	0	0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	7232.7	5036						0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846						0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	11306						0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	12458						0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524						0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104						0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87						0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214						0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8						0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055						213247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626											proportion
SESE	11	614.4	0.0	0.0	26						eration	limb	branch	leaf	total	geophytic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	59	0	0	13	4	74	

All the same variable?

No.

Characteristics of Tidy Data

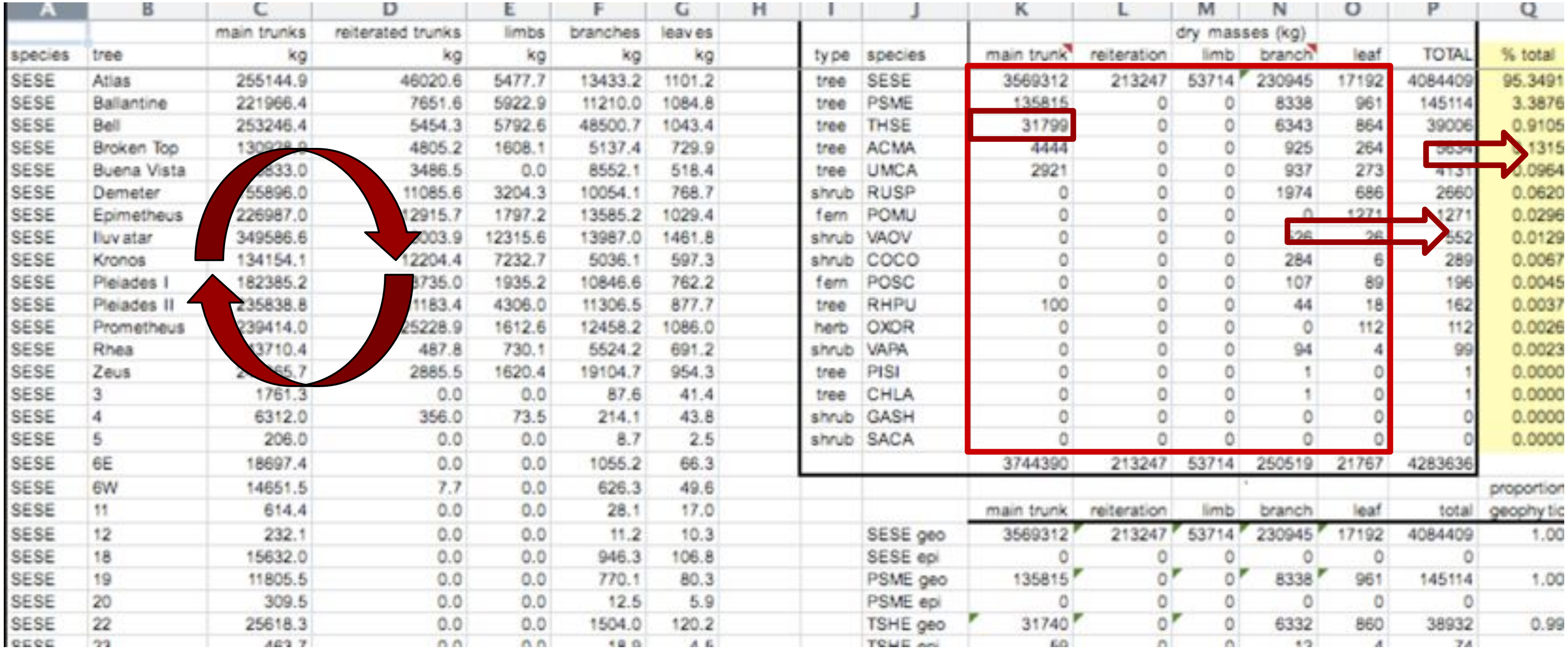
Observations

- Separate tables for each entity measured
- Each row represents a single observed entity
- **Observations (rows) are all unique**

Variables

- All values in a column are of the same type
- **All columns pertain to the same observation (row)**
- **Each column represents either an identifying or measured variable**

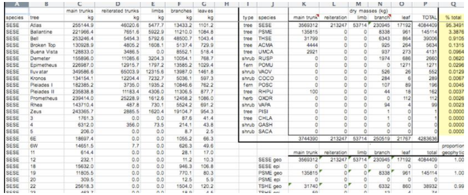
A not-so-reproducible workflow



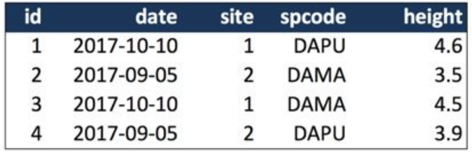
A	B	C	D	E	F	G	H
		main trunks	reiterated trunks	limbs	branches	leaves	
species	tree	kg	kg	kg	kg	kg	
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2	
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8	
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4	
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9	
SESE	Buena Vista	[illegible]833.0	3486.5	0.0	8552.1	518.4	
SESE	Demeter	[illegible]55896.0	11085.6	3204.3	10054.1	768.7	
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4	
SESE	Iluvatar	349586.6	[illegible]003.9	12315.6	13987.0	1461.8	
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3	
SESE	Pleiades I	182385.2	[illegible]8735.0	1935.2	10846.6	762.2	
SESE	Pleiades II	235838.8	[illegible]1183.4	4306.0	11306.5	877.7	
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0	
SESE	Rhea	[illegible]3710.4	487.8	730.1	5524.2	691.2	
SESE	Zeus	2[illegible]65.7	2885.5	1620.4	19104.7	954.3	
SESE	3	1761.3	0.0	0.0	87.6	41.4	
SESE	4	6312.0	356.0	73.5	214.1	43.8	
SESE	5	206.0	0.0	0.0	8.7	2.5	
SESE	6E	18697.4	0.0	0.0	1055.2	66.3	
SESE	6W	14651.5	7.7	0.0	626.3	49.6	
SESE	11	614.4	0.0	0.0	28.1	17.0	
SESE	12	232.1	0.0	0.0	11.2	10.3	
SESE	18	15632.0	0.0	0.0	946.3	106.8	
SESE	19	11805.5	0.0	0.0	770.1	80.3	
SESE	20	309.5	0.0	0.0	12.5	5.9	
SESE	22	25618.3	0.0	0.0	1504.0	120.2	
SESE	23	463.7	0.0	0.0	18.9	4.5	

I	J	K	L	M	N	O	P	Q
				dry masses (kg)				
type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
fern	POMU	0	0	0	0	1271	1271	0.0296
shrub	VAOV	0	0	0	526	26	552	0.0129
shrub	COCO	0	0	0	284	6	289	0.0067
fern	POSC	0	0	0	107	89	196	0.0045
tree	RHPU	100	0	0	44	18	162	0.0037
herb	OXOR	0	0	0	0	112	112	0.0026
shrub	VAPA	0	0	0	94	4	99	0.0023
tree	PISI	0	0	0	1	0	1	0.0000
tree	CHLA	0	0	0	1	0	1	0.0000
shrub	GASH	0	0	0	0	0	0	0.0000
shrub	SACA	0	0	0	0	0	0	0.0000
		3744390	213247	53714	250519	21767	4283636	
								proportion
		main trunk	reiteration	limb	branch	leaf	total	geophytic
	SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
	SESE epi	0	0	0	0	0	0	
	PSME geo	135815	0	0	8338	961	145114	1.00
	PSME epi	0	0	0	0	0	0	
	TSHE geo	31740	0	0	6332	860	38932	0.99
	TSHE epi	60	0	0	13	4	74	

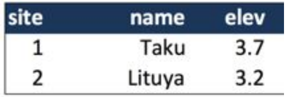
Building a reproducible workflow



Raw, messy data

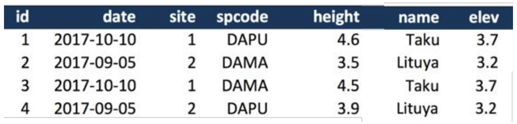


id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9



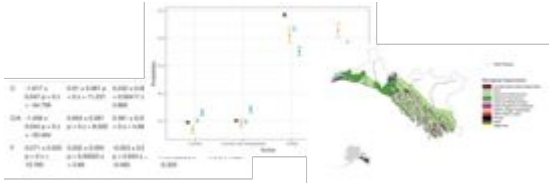
site	name	elev
1	Taku	3.7
2	Lituya	3.2

Clean, raw data



id	date	site	spcode	height	name	elev
1	2017-10-10	1	DAPU	4.6	Taku	3.7
2	2017-09-05	2	DAMA	3.5	Lituya	3.2
3	2017-10-10	1	DAMA	4.5	Taku	3.7
4	2017-09-05	2	DAPU	3.9	Lituya	3.2

Merged/summarized derived data



Figures, tables, maps

Where data come from matters! (a sample)

Issues that may arise from these data sources are our focus today

- Excel
  - Automatic conversion of gene names to dates or floating point numbers*
  - Date values can be converted when transferring data between operating systems and applications
- Text (e.g. CSV) & Excel
  - Free-form structure - lack of enforcement of column-row structure, type consistency
- Text (e.g. CSV)
  - Inconsistent structure - quotes, commas, missing values, spaces
- Database
  - Enforced structure - tables, column typing
  - Specialized methods for interaction (pros and cons to this)

* Ziemann, M., Eren, Y. & El-Osta, A. Gene name errors are widespread in the scientific literature. *Genome Biology* **17**, (2016).

The ESAUSSEE Data Help Desk

who we are and how to find us

Amber Budden, @aebudden, @DataONE_org, aebudden@epscor.unm.edu

Deborah Paul, @idbdeb, @idigbio, dpaul@fsu.edu

Dmitry Schigel, @dschigel, @GBIF, dschigel@gbif.org

Karl Benedict, @kbene, kbene@unm.edu, president@esipfed.org

Kristen Vanderbilt, @vanderbik, @EDIgotdata, krvander@fiu.edu

Kyle Copas, @kylecopas, @GBIF, kcopas@gbif.org

Laura Brenskelle, @lbrensk, @idigbio, lbrensk@ufl.edu

Margaret O'Brien @ , @EDIgotdata, margaret.obrien@ucsb.edu

Megan Jones, @MeganAHJones, @NEON_sci, mjones01@battelleecology.org

Rebekah Wallace, www.eddmaps.org, bekahwal@uga.edu

NSF Arctic Data Center · DataONE · EDI · ESIP MAKING DATA MATTER · GBIF · iDigBio Integrated Digitized Biocollections · NSF · neon Operated by Battelle

Messy data?
Repetitive data tasks?

Increase Reproducibility and Productivity using tools like Open Refine

OpenRefine *A power tool for working with messy data.*

Magic is here.
Ask for it, plan for it.

NSF Arctic Data Center

GBIF

iDigBio Integrated Digitized Biocollections

NSF

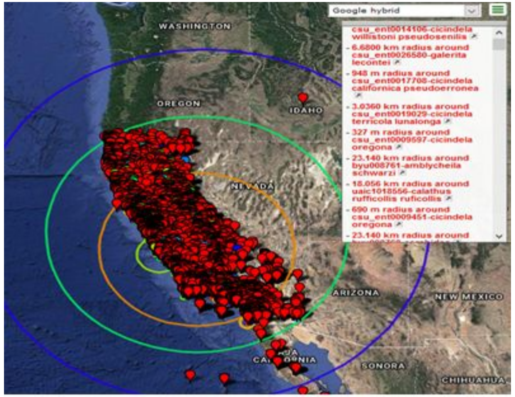
neon Operated by Battelle

Neotoma Paleoecology Database

Data lessons compiled - inspired by workshop

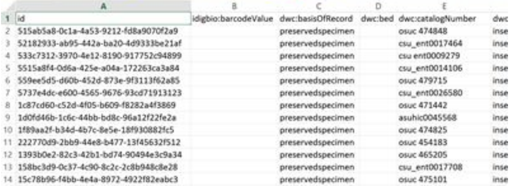
Georeferencing for Research Use of Museum Collections Data

- **Data mapped to standards**
  - supports use and re-use (e.g. Darwin Core DwC, Ecological Metadata Language EML)
  - standards help with data validation and cleaning
- **Data have issues**
  - what are some you have experienced
  - need to be addressed before applying research methods
  - keep raw data raw
  - track your changes
- **Data visualization is key**
  - QGIS lessons
  - Open Refine
  - R, etc.



csu_ent0014106-cicindela willistoni pseudosenilis

- 6.6800 km radius around csu_ent0026580-galerita lecontei
- 948 m radius around csu_ent0017708-cicindela californica pseudoerronea
- 3.0360 km radius around csu_ent0019029-cicindela terricola lunalonga
- 327 m radius around csu_ent0009597-cicindela oregona
- 23.140 km radius around byu008761-amblycheila schwarzi
- 18.056 km radius around uaic1018556-calathus rufficollis ruficollis
- 690 m radius around csu_ent0009451-cicindela oregona
- 23.140 km radius around

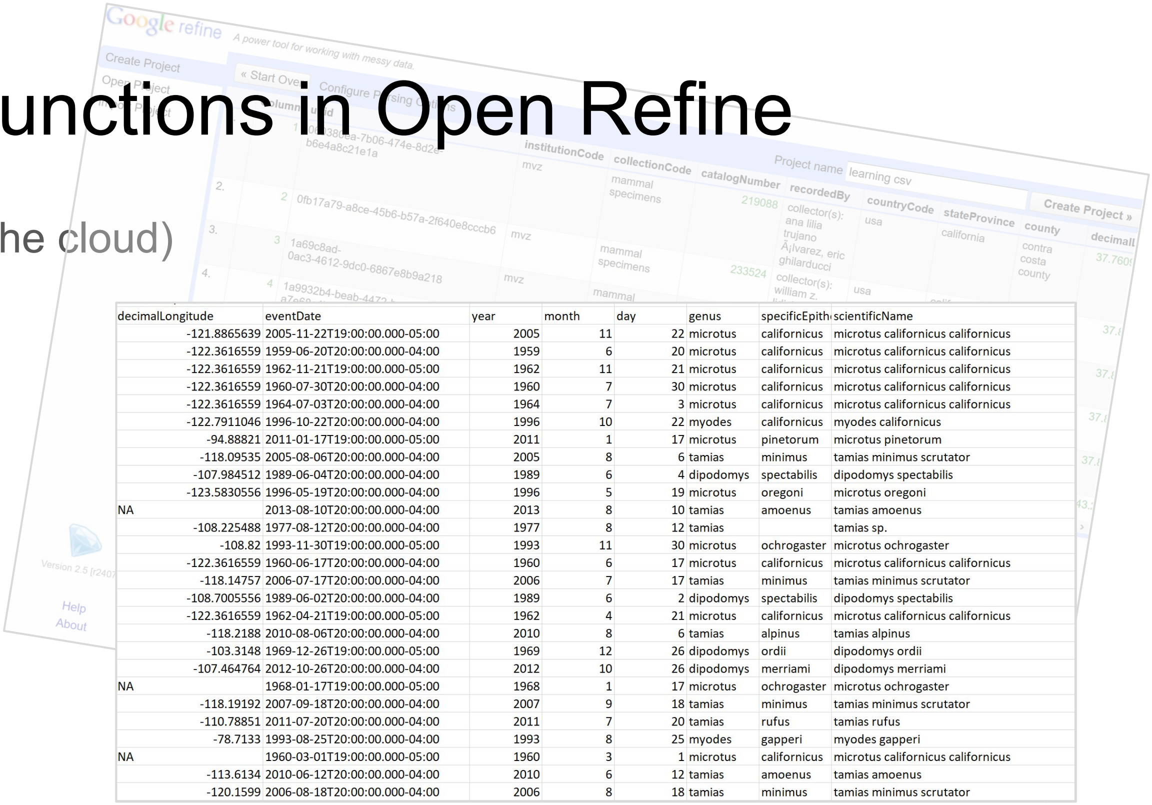


	A	B	C	D	E	
1	id	idigbio:barcodeValue	dwc:basisOfRecord	dwc:bed	dwc:catalogNumber	dwc
2	515ab5a8-0c1a-4a53-9212-fd8a9070f2a9		preservedspecimen		osuc 474848	inse
3	52182933-ab95-442a-ba20-4d9333be21af		preservedspecimen		csu_ent0017464	inse
4	533c7312-3970-4e12-8190-917752c94899		preservedspecimen		csu ent0009279	inse
5	5515a8f4-0d6a-425e-a04a-172263ca3a84		preservedspecimen		csu_ent0014106	inse
6	559ee5d5-d60b-452d-873e-9f3113f62a85		preservedspecimen		osuc 479715	inse
7	5737e4dc-e600-4565-9676-93cd71913123		preservedspecimen		csu_ent0026580	inse
8	1c87cd60-c52d-4f05-b609-f8282a4f3869		preservedspecimen		osuc 471442	inse
9	1d0fd46b-1c6c-44bb-bd8c-96a12f22fe2a		preservedspecimen		asuhic0045568	inse
10	1f89aa2f-b34d-4b7c-8e5e-18f930882fc5		preservedspecimen		osuc 474825	inse
11	222770d9-2bb9-44e8-b477-13f45632f512		preservedspecimen		osuc 454183	inse
12	1393b0e2-82c3-42b1-bd74-90494e3c9a34		preservedspecimen		osuc 465205	inse
13	158bc3d9-0c37-4c90-8c2c-2c8b948c8e28		preservedspecimen		csu_ent0017708	inse
14	15c78b96-f4bb-4e4a-8972-4922f82eabc3		preservedspecimen		osuc 475101	inse

Carabidae (beetles) of California

(Fun!) features and functions in Open Refine

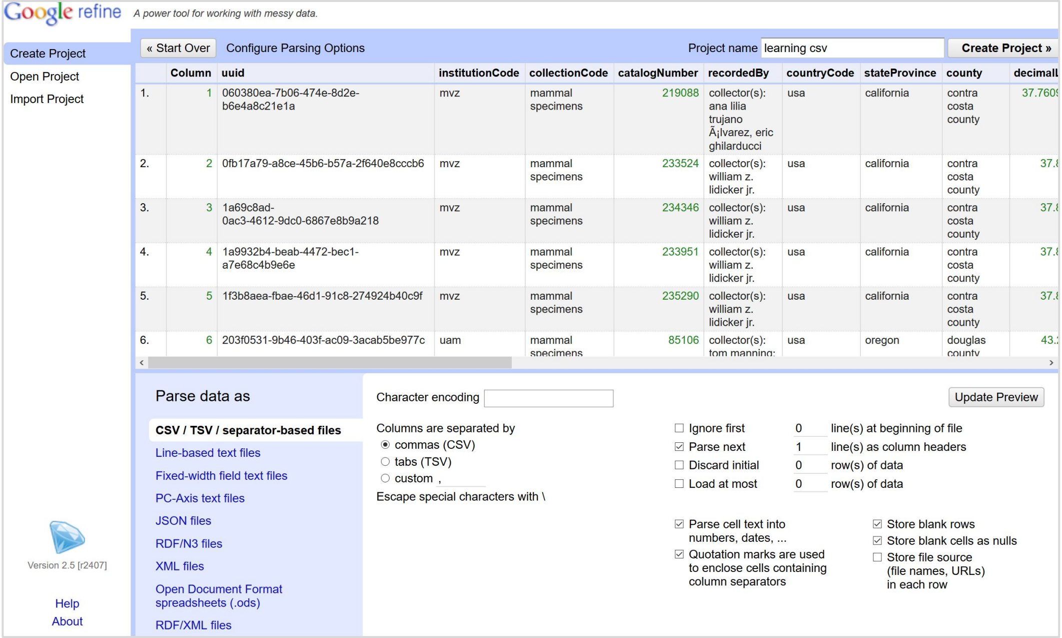
- runs on your computer (not in the cloud)
- data formats supported
- raw data
- column manipulation
- text facet
- routine cleaning (white space)
- clustering
- step-wise editable task script
- APIs
- regular expressions
- export
- share project files



decimalLongitude	eventDate	year	month	day	genus	specificEpith	scientificName
-121.8865639	2005-11-22T19:00:00.000-05:00	2005	11	22	microtus	californicus	microtus californicus californicus
-122.3616559	1959-06-20T20:00:00.000-04:00	1959	6	20	microtus	californicus	microtus californicus californicus
-122.3616559	1962-11-21T19:00:00.000-05:00	1962	11	21	microtus	californicus	microtus californicus californicus
-122.3616559	1960-07-30T20:00:00.000-04:00	1960	7	30	microtus	californicus	microtus californicus californicus
-122.3616559	1964-07-03T20:00:00.000-04:00	1964	7	3	microtus	californicus	microtus californicus californicus
-122.7911046	1996-10-22T20:00:00.000-04:00	1996	10	22	myodes	californicus	myodes californicus
-94.88821	2011-01-17T19:00:00.000-05:00	2011	1	17	microtus	pinetorum	microtus pinetorum
-118.09535	2005-08-06T20:00:00.000-04:00	2005	8	6	tamias	minimus	tamias minimus scrutator
-107.984512	1989-06-04T20:00:00.000-04:00	1989	6	4	dipodomys	spectabilis	dipodomys spectabilis
-123.5830556	1996-05-19T20:00:00.000-04:00	1996	5	19	microtus	oregoni	microtus oregoni
NA	2013-08-10T20:00:00.000-04:00	2013	8	10	tamias	amoenus	tamias amoenus
-108.225488	1977-08-12T20:00:00.000-04:00	1977	8	12	tamias		tamias sp.
-108.82	1993-11-30T19:00:00.000-05:00	1993	11	30	microtus	ochrogaster	microtus ochrogaster
-122.3616559	1960-06-17T20:00:00.000-04:00	1960	6	17	microtus	californicus	microtus californicus californicus
-118.14757	2006-07-17T20:00:00.000-04:00	2006	7	17	tamias	minimus	tamias minimus scrutator
-108.7005556	1989-06-02T20:00:00.000-04:00	1989	6	2	dipodomys	spectabilis	dipodomys spectabilis
-122.3616559	1962-04-21T19:00:00.000-05:00	1962	4	21	microtus	californicus	microtus californicus californicus
-118.2188	2010-08-06T20:00:00.000-04:00	2010	8	6	tamias	alpinus	tamias alpinus
-103.3148	1969-12-26T19:00:00.000-05:00	1969	12	26	dipodomys	ordii	dipodomys ordii
-107.464764	2012-10-26T20:00:00.000-04:00	2012	10	26	dipodomys	merriami	dipodomys merriami
NA	1968-01-17T19:00:00.000-05:00	1968	1	17	microtus	ochrogaster	microtus ochrogaster
-118.19192	2007-09-18T20:00:00.000-04:00	2007	9	18	tamias	minimus	tamias minimus scrutator
-110.78851	2011-07-20T20:00:00.000-04:00	2011	7	20	tamias	rufus	tamias rufus
-78.7133	1993-08-25T20:00:00.000-04:00	1993	8	25	myodes	gapperi	myodes gapperi
NA	1960-03-01T19:00:00.000-05:00	1960	3	1	microtus	californicus	microtus californicus californicus
-113.6134	2010-06-12T20:00:00.000-04:00	2010	6	12	tamias	amoenus	tamias amoenus
-120.1599	2006-08-18T20:00:00.000-04:00	2006	8	18	tamias	minimus	tamias minimus scrutator

Open Refine - getting started is quick and easy

- download and install
- launch
- import your data
- your raw data is NOT touched
- supported data formats
- subset data



Google refine *A power tool for working with messy data.*

Project name learning csv

	Column	uuid	institutionCode	collectionCode	catalogNumber	recordedBy	countryCode	stateProvince	county	decimalL
1.	1	060380ea-7b06-474e-8d2e-b6e4a8c21e1a	mvz	mammal specimens	219088	collector(s): ana lilia trujano Ã¡lvarez, eric ghilarducci	usa	california	contra costa county	37.760
2.	2	0fb17a79-a8ce-45b6-b57a-2f640e8cccb6	mvz	mammal specimens	233524	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.
3.	3	1a69c8ad-0ac3-4612-9dc0-6867e8b9a218	mvz	mammal specimens	234346	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.
4.	4	1a9932b4-beab-4472-bec1-a7e68c4b9e6e	mvz	mammal specimens	233951	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.
5.	5	1f3b8aea-fbae-46d1-91c8-274924b40c9f	mvz	mammal specimens	235290	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.
6.	6	203f0531-9b46-403f-ac09-3acab5be977c	uam	mammal specimens	85106	collector(s): tom manning;	usa	oregon	douglas county	43.

Parse data as

- **CSV / TSV / separator-based files**
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- RDF/N3 files
- XML files
- Open Document Format spreadsheets (.ods)
- RDF/XML files

Character encoding

Columns are separated by

- ● commas (CSV)
- ○ tabs (TSV)
- ○ custom ,

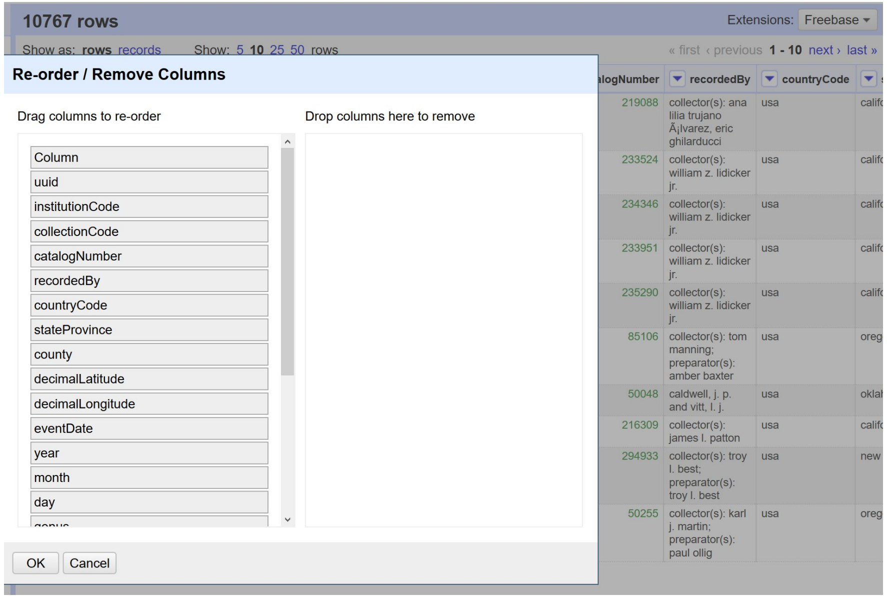
Escape special characters with \

- □ Ignore first 0 line(s) at beginning of file
- ☑ Parse next 1 line(s) as column headers
- □ Discard initial 0 row(s) of data
- □ Load at most 0 row(s) of data

- ☑ Parse cell text into numbers, dates, ...
- ☑ Quotation marks are used to enclose cells containing column separators
- ☑ Store blank rows
- ☑ Store blank cells as nulls
- □ Store file source (file names, URLs) in each row

Open Refine - managing columns

- reorganize columns easily



10767 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows

« first ‹ previous 1 - 10 next › last »

Re-order / Remove Columns

Drag columns to re-order

Drop columns here to remove

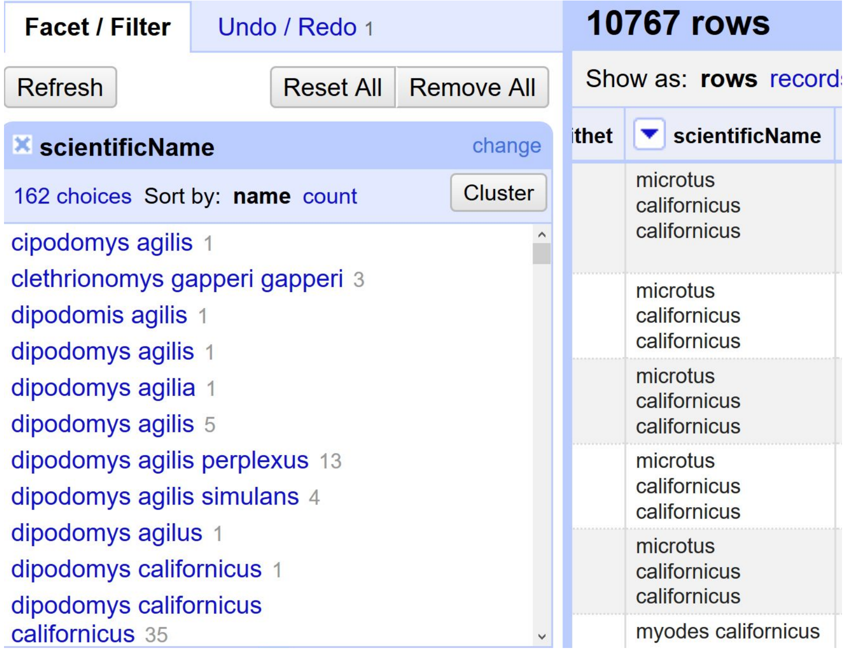
- Column
- uuid
- institutionCode
- collectionCode
- catalogNumber
- recordedBy
- countryCode
- stateProvince
- county
- decimalLatitude
- decimalLongitude
- eventDate
- year
- month
- day
- genus

OK Cancel

alogNumber	recordedBy	countryCode	
219088	collector(s): ana lilia trujano Ã¡lvarez, eric ghilarducci	usa	calif
233524	collector(s): william z. lidicker jr.	usa	calif
234346	collector(s): william z. lidicker jr.	usa	calif
233951	collector(s): william z. lidicker jr.	usa	calif
235290	collector(s): william z. lidicker jr.	usa	calif
85106	collector(s): tom manning; preparator(s): amber baxter	usa	oreg
50048	caldwell, j. p. and vitt, l. j.	usa	oklal
216309	collector(s): james l. patton	usa	calif
294933	collector(s): troy l. best; preparator(s): troy l. best	usa	new
50255	collector(s): karl j. martin; preparator(s): paul ollig	usa	oreg

Open refine - text facet

lists and counts the distinct values in a column



Facet / Filter

Undo / Redo 1

Refresh

Reset All

Remove All

scientificName

change

162 choices Sort by: **name** count

Cluster

cipodomys agilis	1
clethrionomys gapperi gapperi	3
dipodomis agilis	1
dipodomys agilis	1
dipodomys agilia	1
dipodomys agilis	5
dipodomys agilis perplexus	13
dipodomys agilis simulans	4
dipodomys agilus	1
dipodomys californicus	1
dipodomys californicus californicus	35

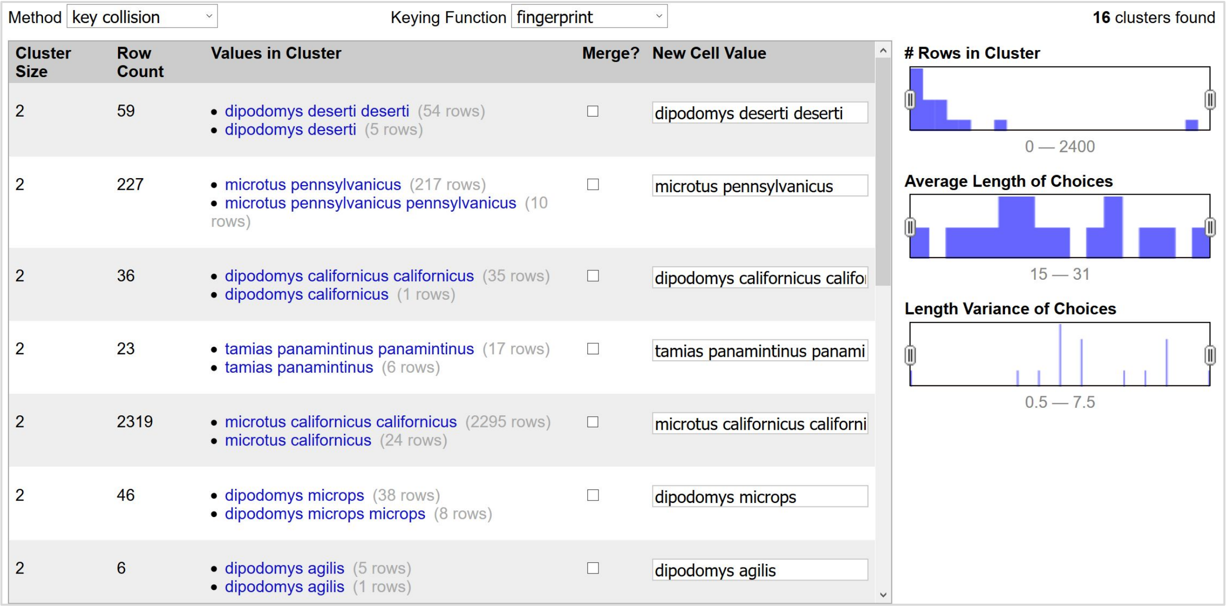
10767 rows

Show as: **rows** record

ithet	scientificName
	microtus californicus californicus
	microtus californicus californicus
	microtus californicus californicus
	microtus californicus californicus
	microtus californicus californicus
	myodes californicus

Open Refine - the magic of clustering algorithms

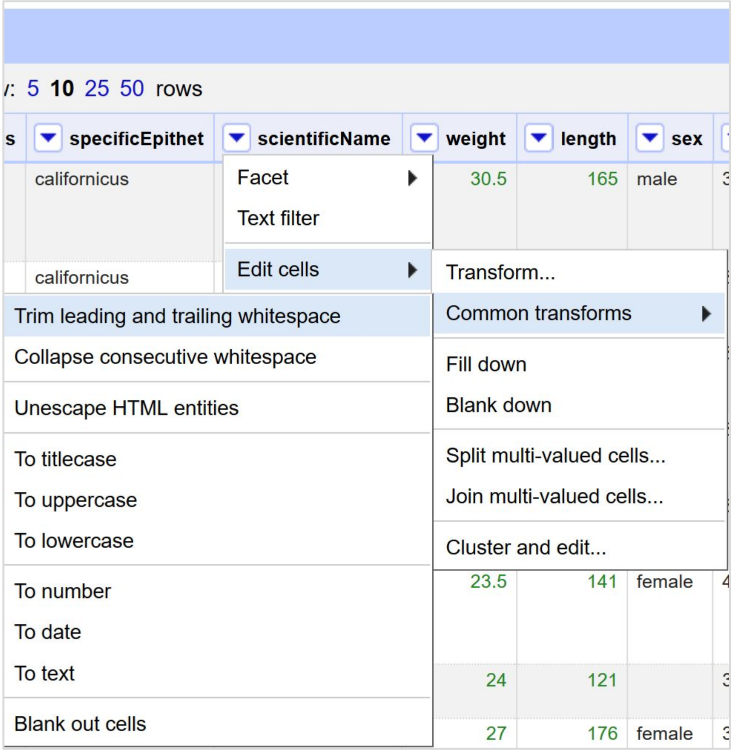
or how to find issues that abc sort won’t
and fix them all at once - no hunting



Method key collision ⌄ Keying Function fingerprint ⌄ **16** clusters found

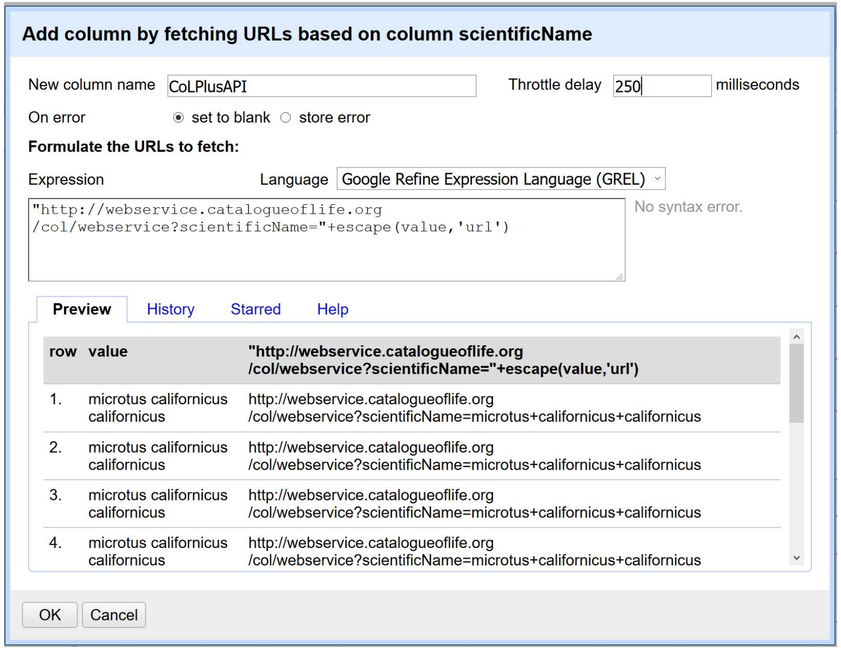
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	59	• dipodomys deserti deserti (54 rows) • dipodomys deserti (5 rows)	□	dipodomys deserti deserti
2	227	• microtus pennsylvanicus (217 rows) • microtus pennsylvanicus pennsylvanicus (10 rows)	□	microtus pennsylvanicus
2	36	• dipodomys californicus californicus (35 rows) • dipodomys californicus (1 rows)	□	dipodomys californicus califo
2	23	• tamias panamintinus panamintinus (17 rows) • tamias panamintinus (6 rows)	□	tamias panamintinus panami
2	2319	• microtus californicus californicus (2295 rows) • microtus californicus (24 rows)	□	microtus californicus californi
2	46	• dipodomys microps (38 rows) • dipodomys microps microps (8 rows)	□	dipodomys microps
2	6	• dipodomys agilis (5 rows) • dipodomys agilis (1 rows)	□	dipodomys agilis

Open Refine - manages pesky white spaces



Open Refine - add data to your data using APIs

that's application programming interface



Add column by fetching URLs based on column scientificName

New column name: CoLPlusAPI

Throttle delay: 250 milliseconds

On error: ● set to blank ○ store error

Formulate the URLs to fetch:

Expression

Language: Google Refine Expression Language (GREL)

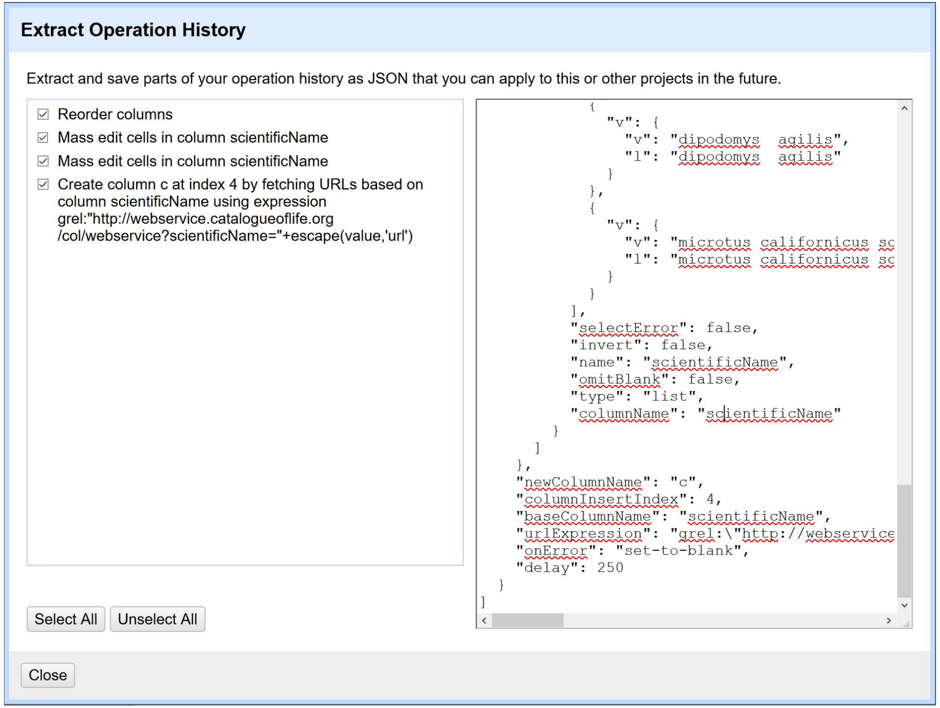
```
"http://webservice.cataloguelife.org
/col/webservice?scientificName="+escape(value,'url')
```

No syntax error.

row	value	"http://webservice.cataloguelife.org /col/webservice?scientificName="+escape(value,'url')
1.	microtus californicus californicus	http://webservice.cataloguelife.org /col/webservice?scientificName=microtus+californicus+californicus
2.	microtus californicus californicus	http://webservice.cataloguelife.org /col/webservice?scientificName=microtus+californicus+californicus
3.	microtus californicus californicus	http://webservice.cataloguelife.org /col/webservice?scientificName=microtus+californicus+californicus
4.	microtus californicus californicus	http://webservice.cataloguelife.org /col/webservice?scientificName=microtus+californicus+californicus

Open Refine - saves your steps

- ❑ *supports reproducibility*
- ❑ *tracks your work for you*
- ❑ *easy to go back to earlier steps with confidence*



Extract Operation History

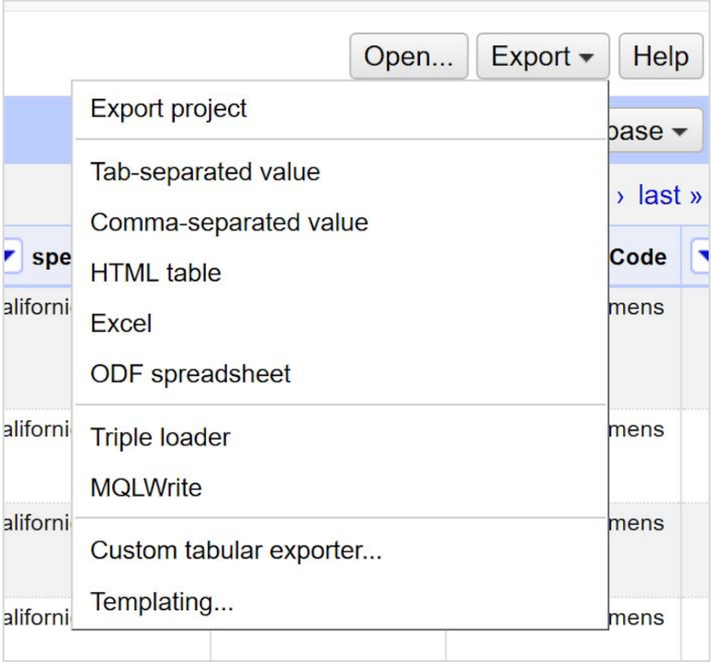
Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

☑ Reorder columns
☑ Mass edit cells in column scientificName
☑ Mass edit cells in column scientificName
☑ Create column c at index 4 by fetching URLs based on column scientificName using expression grel:"http://webservice.catalogueoflife.org/col/webservice?scientificName="+escape(value,'url')

```
            {
              "v": {
                "v": "dipodomys  agilis",
                "l": "dipodomys  agilis"
              }
            },
            {
              "v": {
                "v": "microtus californicus sc
                "l": "microtus californicus sc
              }
            }
          ],
          "selectError": false,
          "invert": false,
          "name": "scientificName",
          "omitBlank": false,
          "type": "list",
          "columnName": "scientificName"
        }
      ]
    },
    "newColumnName": "c",
    "columnInsertIndex": 4,
    "baseColumnName": "scientificName",
    "urlExpression": "grel:\"http://webservice
    "onError": "set-to-blank",
    "delay": 250
  }
]
```

Open Refine - export your data, share project files

select the format
export subsets too
and project files



Export project

Tab-separated value

Comma-separated value

HTML table

Excel

ODF spreadsheet

Triple loader

MQLWrite

Custom tabular exporter...

Templating...

Open Refine - make some friends

- share this tool with students, friends, families, colleagues
- **imagine future tools, think beyond spreadsheets**

Increase Reproducibility and Productivity
using tools like Open Refine

Magic is here.
Ask for it, plan for it.

Looking for next steps now?
R, Open Refine, and Data Management resources

- The #datahelpdesk is ready to offer data assistance!
- #CareerCentral Q and A: Wednesday, August 5th, 9:30-10:30 PDT (12:30-1:00 EDT)
- Data Help Desk Wiki **https://bit.ly/datahelpesa2020**
- Data Carpentry lessons
- on Twitter #ESA2020 #datahelpdesk