

# Data Cleaning for Analysis and Publication Using R and OpenRefine Software Packages

*Presenter* : Arctic Data Center, DataONE, Environmental Data Initiative, **ESIP**, GBIF, **iDigBio**, NEON

**Karl Benedict, Deborah Paul**

**#datahelpdesk**

Ecological Society of America 2019 ESAUSSEE

Career Fair Center in the Exhibit Hall <https://esa.org/louisville/career-fair/>

Wednesday 14 August 11:30 AM - 1:15 PM



**GBIF**



**neon**  
Operated by Battelle



# Advancing the Digitization of Biological Collections

## iDigBio Hub and Thematic (Museum) Collection Networks

total: 119,768,942

### Digitization

Workflows & Protocols  
Dissemination

### Research Use

Cyberinfrastructure  
Tool collaboration  
Portal development  
ENM workshop  
Research Spotlight  
Data quality  
APIs

### Training

Biodiversity informatics  
Data skills and literacy  
Collections software  
Imaging  
Project Management

www.idigbio.org

Search Records

search all fields

Must have media Must have map point

Filters Mapping Sorting Download

Add a field

State/Province: kentucky

Present Missing

Top 6 Taxa

- Cnidaria
- Empoasca empoasca
- Agallia constricta
- Paraphlepsius irroratus
- Fungaria
- other

LECTOTYPE ♀  
Scelio niteus Brues  
By L. Masner, 64

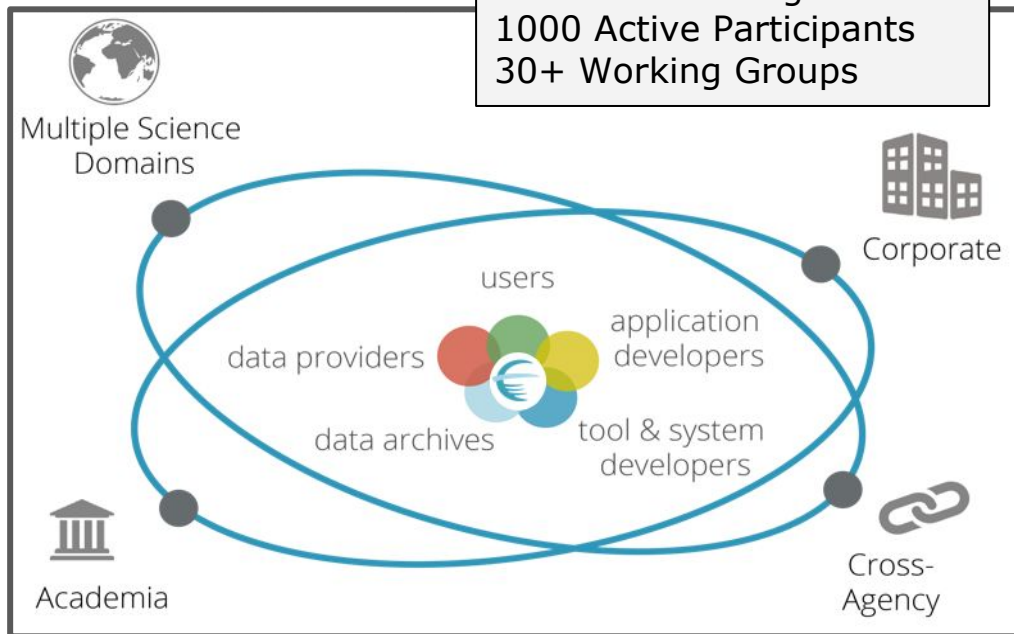
### Education Outreach

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

120+ Partner Orgs  
1000 Active Participants  
30+ Working Groups



## DISCOVER

Find people and tools to make your data findable, accessible, interoperable, and reusable.



## COLLABORATE

Join-in or create a new collaboration area around your Earth science data challenges.



## INNOVATE

Utilize small-grant funding to build or expand Earth data technologies.



## NETWORK

Extend your network. Build connections across federal agencies, the private sector, and academia.

# 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

Clean data don't have these issues (and others)

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,<sup>1</sup> William K. Michener,<sup>2</sup> Robert B. Cook<sup>3</sup>

<sup>1</sup> University of Michigan, Ann Arbor, MI, USA

<sup>2</sup> University of New Mexico, Albuquerque, NM, USA

<sup>3</sup> Oak Ridge National Laboratory, Oak Ridge, TN, USA

# Why do we care about “Clean” data

- Many analyses can't be performed until the data have been cleaned, e.g.
  - Location-based analysis based on separate Latitude-Longitude values
  - Temporal analysis based upon date/timestamps in different formats/timezones
  - Reconciliation of changing relationships between scientific names and taxonomic syntax\*
- Erroneous results may be generated when using data with data entry errors\*\*
- Integration of disparate data requires standardized common values - e.g. using relational database models for integration

\*Remsen, D. The use and limits of scientific names in biological informatics. *Zookeys*207–223 (2016). doi:[10.3897/zookeys.550.9546](https://doi.org/10.3897/zookeys.550.9546)

\*\*Barchard, K. A. & Pace, L. A. Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior* 27, 1834–1839 (2011).

# Core concepts when creating / using tabular data

- Variables in Columns, Observations in Rows ⇒ **follow standards**
- 1 Column / Variable, 1 Row / Observation ⇒ **clear structure**
- Short, descriptive column names without spaces ⇒ **portability**
- Single data type (e.g. numeric, text, date) in each column ⇒ **efficient processing**
- Explicitly & consistently represent (encode) missing values ⇒ **reduce ambiguity**
- Document your structure & processes ⇒ **discoverability, replicability**
- **Do Not** modify your raw data



Following these rules makes life easier for **you, your collaborators**, and **future users** of your data. Ignore them and you create *Technical Debt*

# Where data come from matters! (a sample)

- **Excel** Issues that may arise from these data sources are our focus today
  - 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).

# Data for today's demonstrations

Download from iDigBio.

This dataset contains specimen records for rodent specimens in natural science collections. Each record is about a single rodent and contains such information as where it was collected, by whom, when, and what taxon name was applied and a georeference for where the specimen was collected.



# Specific cleanup tasks to be demonstrated

- Data Import and checking for errors
- Inspection of a sample of items
  - Summary statistics
  - Visual inspection
- Managing missing values
- Column splitting
- Column type conversion
- Selectively replacing values

# Today's Demonstrations

- Cleanup in R
- Cleanup in Open Refine

# 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, @EDlgotdata, krvander@fiu.edu

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

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

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

Megan Jones, @MeganAHJones, @NEON\_sci, mjones01@battelleecology.org

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



# Messy data? Repetitive data tasks?

Increase Reproducibility and Productivity using tools  
like Open Refine

*Magic is coming.  
Ask for it, plan for it.*

# (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	decimal
1.	060380ea-7b06-474e-8d2e-b6e4a8c21e1a	mvz	mammal specimens	219088	collector(s): ana lilia trujano Ajlvarez, eric ghilarducci	usa	california	contra costa county	37.760
2.	0fb17a79-a8ce-45b6-b57a-2f640e8cccb6	mvz	mammal specimens	233524	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.1
3.	1a69c8ad-0ac3-4612-9dc0-6867e8b9a218	mvz	mammal specimens	234346	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.1
4.	1a9932b4-beab-4472-bec1-a7e68c4b9e6e	mvz	mammal specimens	233951	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.1
5.	1f3b8aea-fbae-46d1-91c8-274924b40c9f	mvz	mammal specimens	235290	collector(s): william z. lidicker jr.	usa	california	contra costa county	37.1
6.	203f0531-9b46-403f-ac09-3acab5be977c	uam	mammal specimens	85106	collector(s): tom manning	usa	oregon	douglas county	43.1

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, ...

Store blank rows

Quotation marks are used to enclose cells containing column separators

Store blank cells as nulls

Store file source (file names, URLs) in each row

Version 2.5 [r2407]

Help About

# 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	logNumber	recordedBy	countryCode
uuid	219088	collector(s): ana lilia trujano	usa
institutionCode	233524	Álvarez, eric ghilarducci	usa
collectionCode	234346	collector(s): william z. lidicker jr.	usa
catalogNumber	233951	collector(s): william z. lidicker jr.	usa
recordedBy	233951	collector(s): william z. lidicker jr.	usa
countryCode	235290	collector(s): william z. lidicker jr.	usa
stateProvince	85106	collector(s): tom manning; preparator(s): amber baxter	usa
county	50048	collector(s): caldwell, j. p. and vitt, l. j.	usa
decimalLatitude	216309	collector(s): james l. patton	usa
decimalLongitude	294933	collector(s): troy l. best; preparator(s): troy l. best	usa
eventDate	50255	collector(s): karl j. martin; preparator(s): paul ollig	usa
year			
month			
day			

OK Cancel

# Open refine - text facet

*lists and counts the distinct values in a column*

Facet / Filter    Undo / Redo 1    **10767 rows**

Refresh    Reset All    Remove All    Show as: rows records

**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

thet	scientificName
	microtus californicus californicus
	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	<ul style="list-style-type: none"><li>dipodomys deserti deserti (54 rows)</li><li>dipodomys deserti (5 rows)</li></ul>	<input type="checkbox"/>	dipodomys deserti deserti
2	227	<ul style="list-style-type: none"><li>microtus pennsylvanicus (217 rows)</li><li>microtus pennsylvanicus pennsylvanicus (10 rows)</li></ul>	<input type="checkbox"/>	microtus pennsylvanicus
2	36	<ul style="list-style-type: none"><li>dipodomys californicus californicus (35 rows)</li><li>dipodomys californicus (1 rows)</li></ul>	<input type="checkbox"/>	dipodomys californicus califoi
2	23	<ul style="list-style-type: none"><li>tamias panamintinus panamintinus (17 rows)</li><li>tamias panamintinus (6 rows)</li></ul>	<input type="checkbox"/>	tamias panamintinus panami
2	2319	<ul style="list-style-type: none"><li>microtus californicus californicus (2295 rows)</li><li>microtus californicus (24 rows)</li></ul>	<input type="checkbox"/>	microtus californicus californi
2	46	<ul style="list-style-type: none"><li>dipodomys microps (38 rows)</li><li>dipodomys microps microps (8 rows)</li></ul>	<input type="checkbox"/>	dipodomys microps
2	6	<ul style="list-style-type: none"><li>dipodomys agilis (5 rows)</li><li>dipodomys agilis (1 rows)</li></ul>	<input type="checkbox"/>	dipodomys agilis

The histograms provide statistical insights into the clustering process. The first histogram, '# Rows in Cluster', shows a distribution of cluster sizes with a peak at 2 rows. The second histogram, 'Average Length of Choices', shows the average length of the strings used for clustering, with a peak around 20. The third histogram, 'Length Variance of Choices', shows the variance of the string lengths, with a peak around 1.5.

# Open Refine - manages pesky white spaces

5 10 25 50 rows

specificEpithet	scientificName	weight	length	sex
californicus		30.5	165	male
californicus				
		23.5	141	female
		24	121	
		27	176	female

Facet

Text filter

Edit cells

Transform...

Common transforms

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- Blank out cells

Fill down

Blank down

Split multi-valued cells...

Join multi-valued cells...

Cluster and edit...

# 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  Throttle delay  milliseconds

On error  set to blank  store error

**Formulate the URLs to fetch:**

Expression  Language  No syntax error.

**Preview** History Starred Help

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

OK Cancel

# 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')`

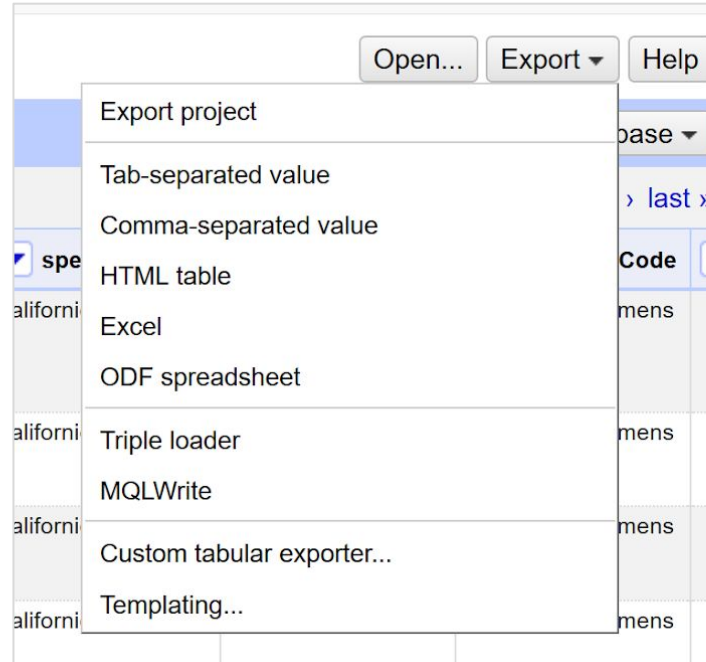
```
{
  "v": {
    "v": "dipodomys agilis",
    "l": "dipodomys agilis"
  },
  "v": {
    "v": "microtus californicus sc",
    "l": "microtus californicus sc"
  }
},
"selectError": false,
"invert": false,
"name": "scientificName",
"omitBlank": false,
"type": "list",
"columnName": "scientificName"
}
],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

Select All Unselect All

Close

# Open Refine - export your data, share project files

*select the **format**  
export **subsets too**  
and **project files***



# 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 coming.  
Ask for it, plan for it.



NSF  
Arctic  
Data  
Center



MAKING DATA MATTER



Operated by Battelle

# Looking for next steps now?

## *R, Open Refine, and Data Management resources*

- find us in the Data Help Desk booth
- Data Help Desk Wiki <https://bit.ly/datahelpesa2019>
- Data Carpentry lessons
- Help us help you - take our survey
  - <http://dhdsurveyesa2019>