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

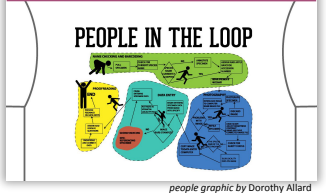
#datahelpdesk

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

Advancing the Digitization of Biological Collections

iDigBio Hub and Thematic (Museum) Collection Networks

(Meta)data Aggregator
Community Building
Collections Data



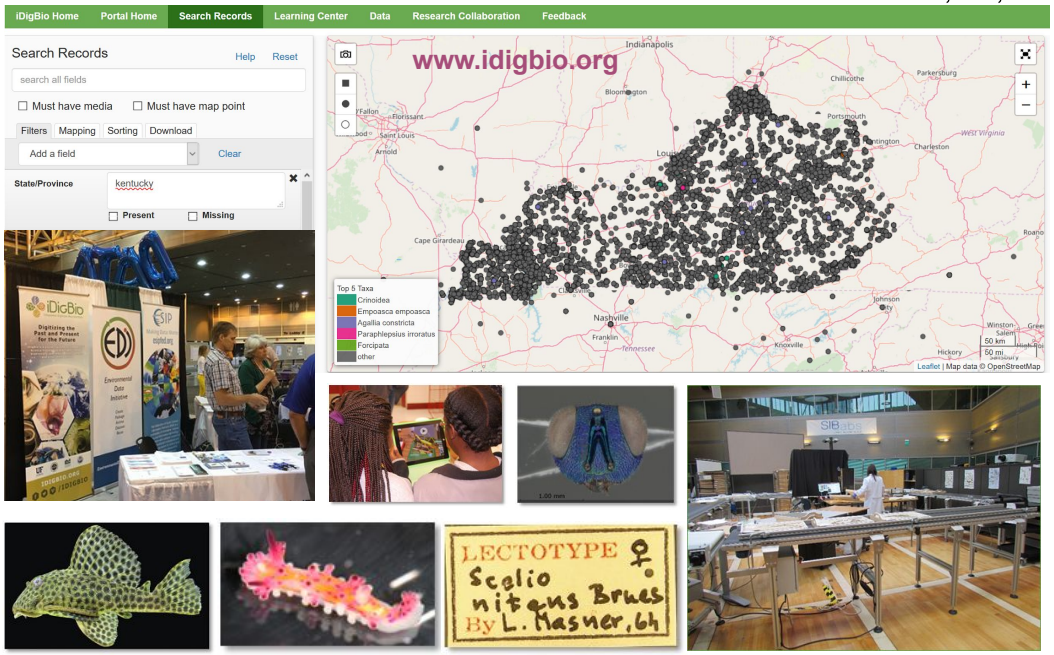
people graphic by Dorothy Allard

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



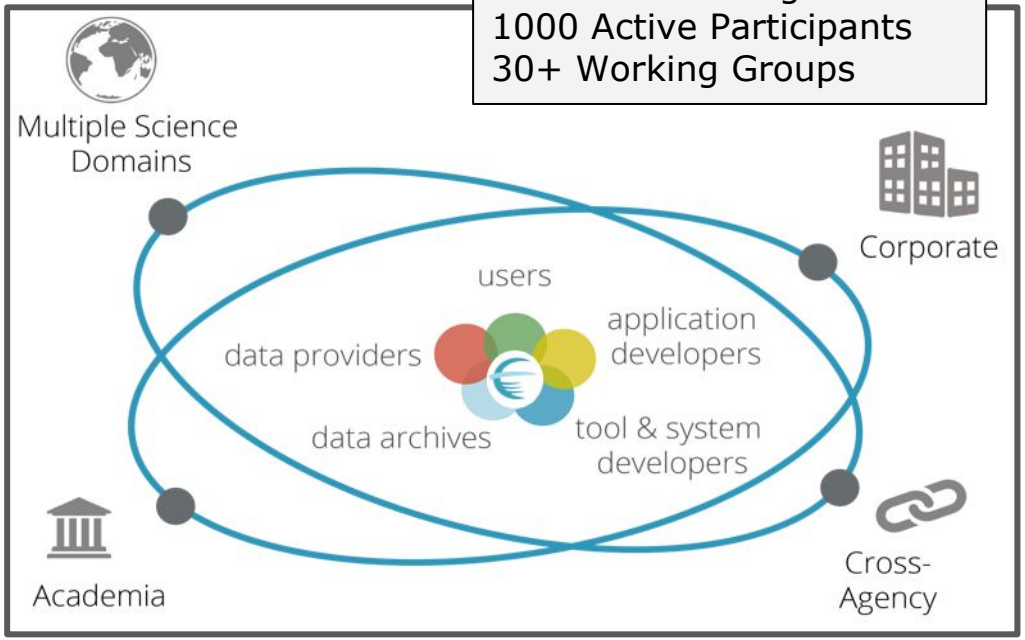
Education Outreach
Citizen Science
K-12 materials
Undergraduate
Fossil Clubs
Mentor teachers

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

Earth Science Information Partners

Supported by:

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 setor, and academia.

https://esipfed.org/

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,[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

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

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)

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).

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, @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

Messy data?
Repetitive data tasks?

Increase Reproducibility and Productivity using tools like Open Refine

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

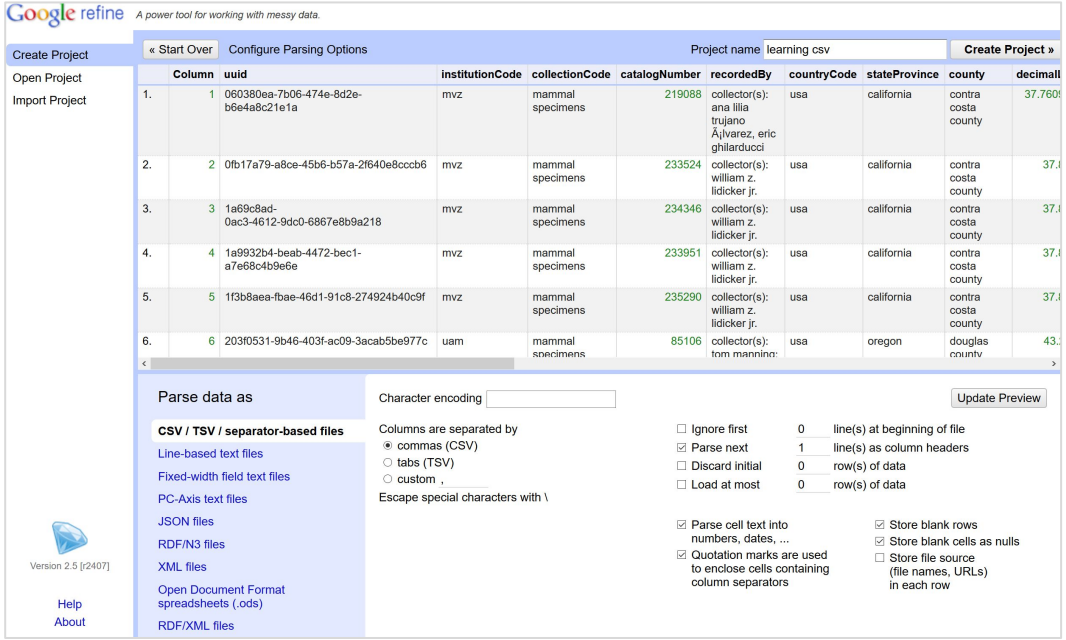
(Fun!) features and functions in Open Refine

- 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

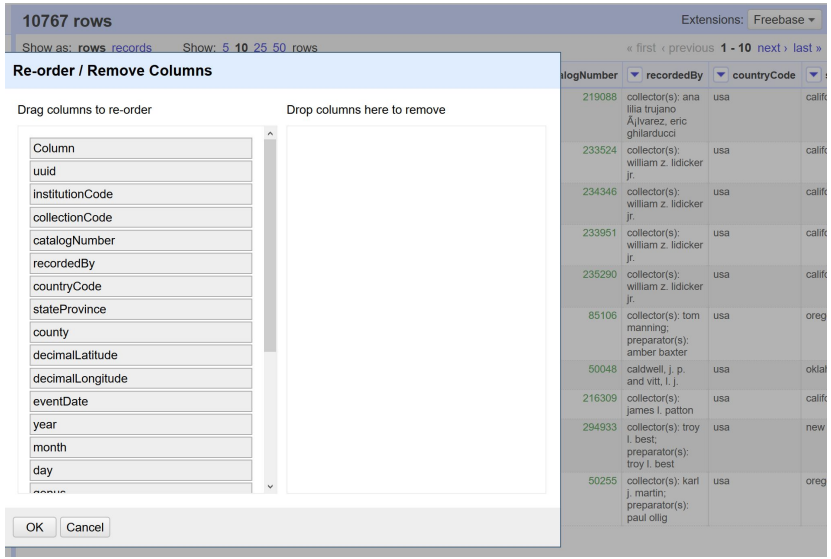
Open Refine - getting started is quick and easy

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



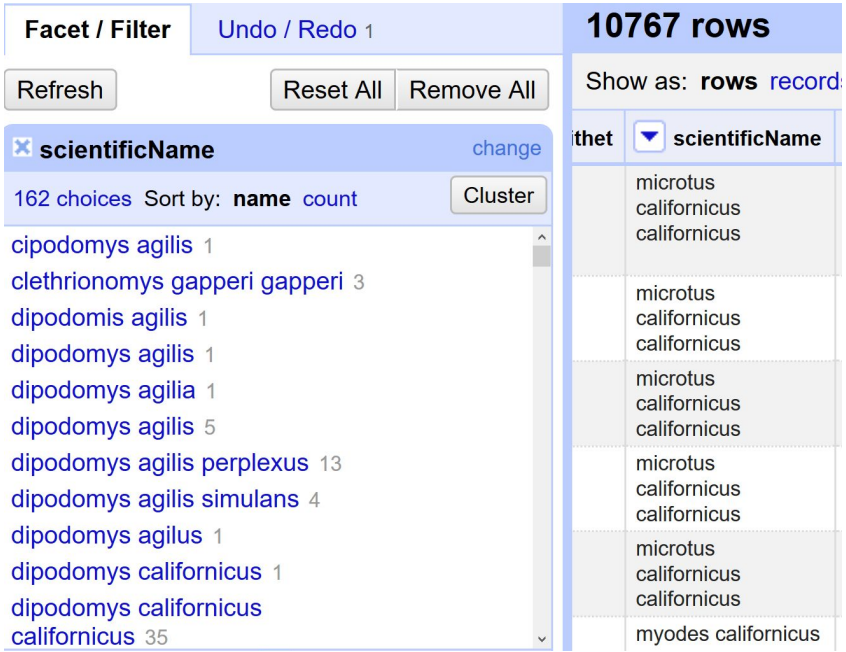
Open Refine - managing columns

- reorganize columns easily



Open refine - text facet

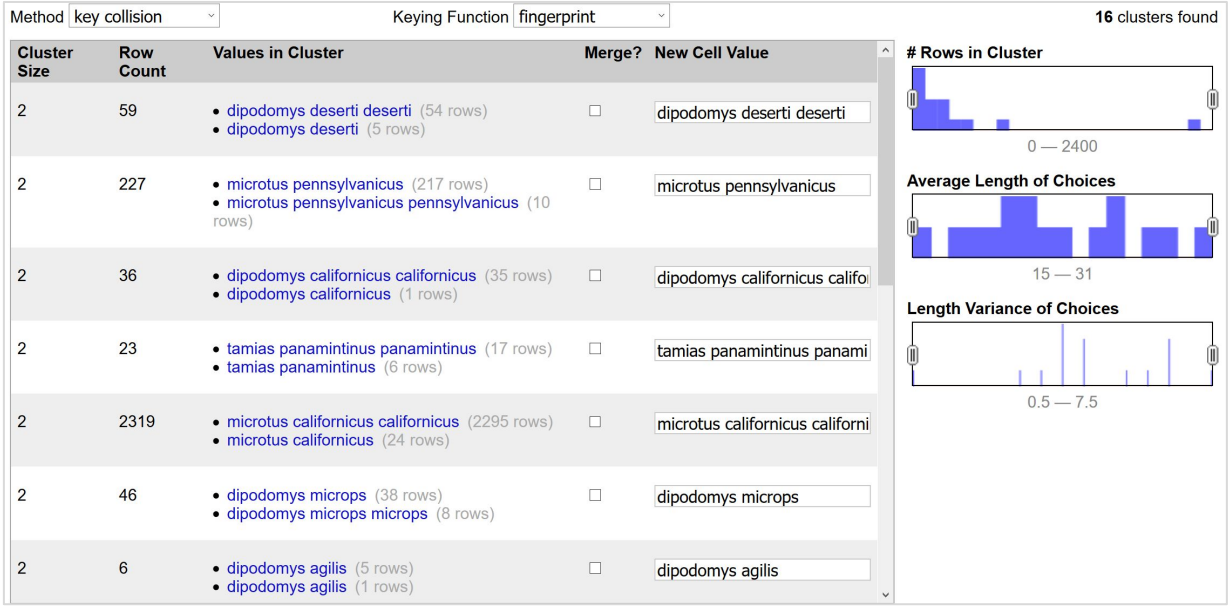
lists and counts the distinct values in a column



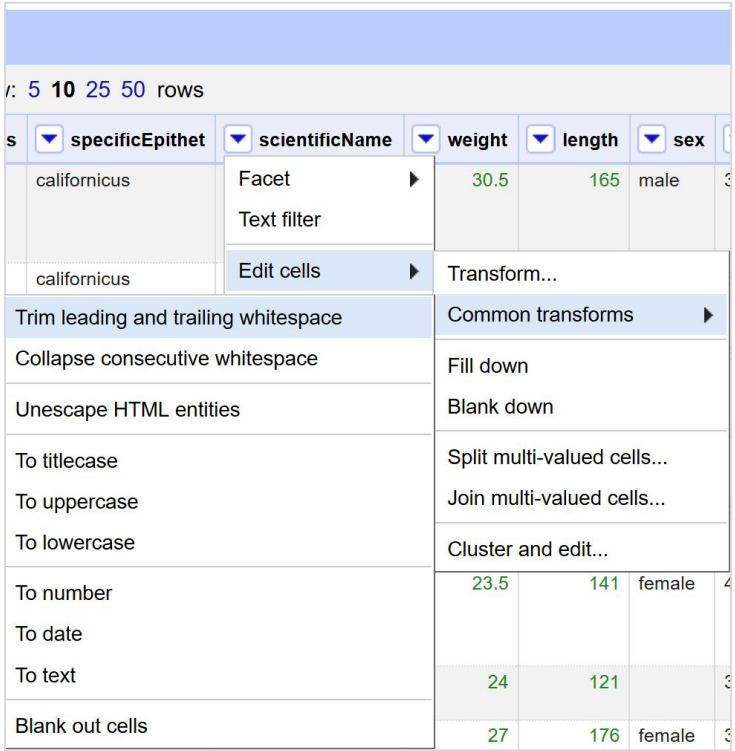
Open Refine - the magic of clustering algorithms

or how to find issues that abc sort won't

and fix them all at once - no hunting

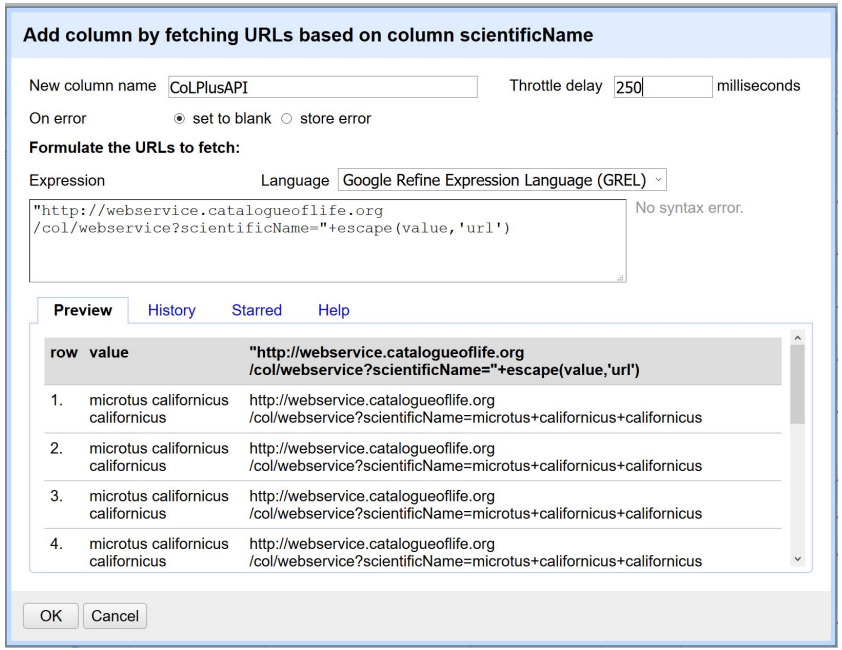


Open Refine - manages pesky white spaces



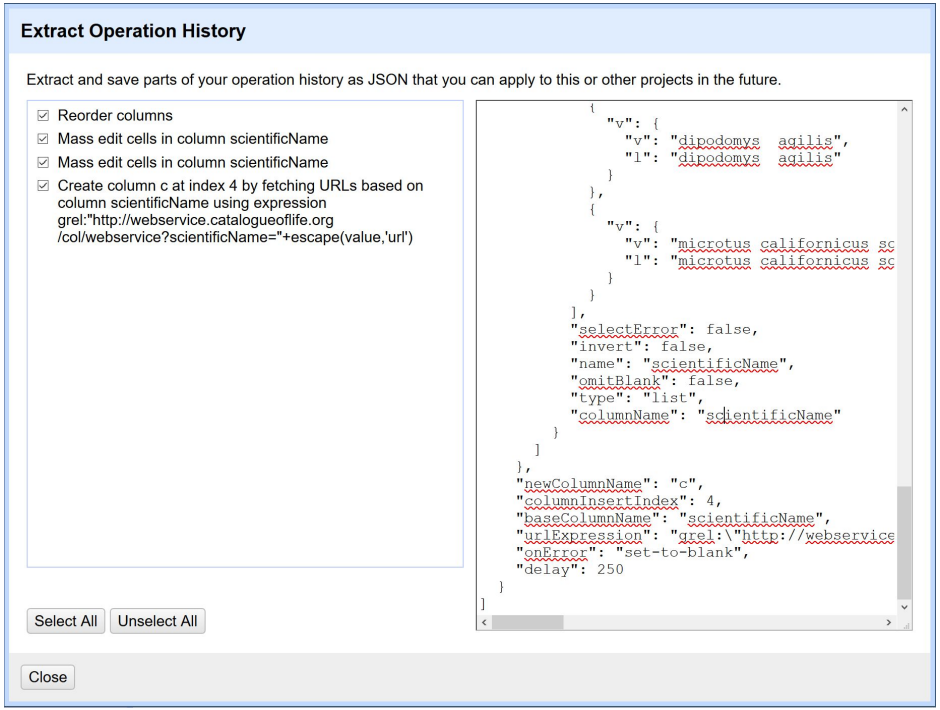
Open Refine - add data to your data using APIs

that's application programming interface



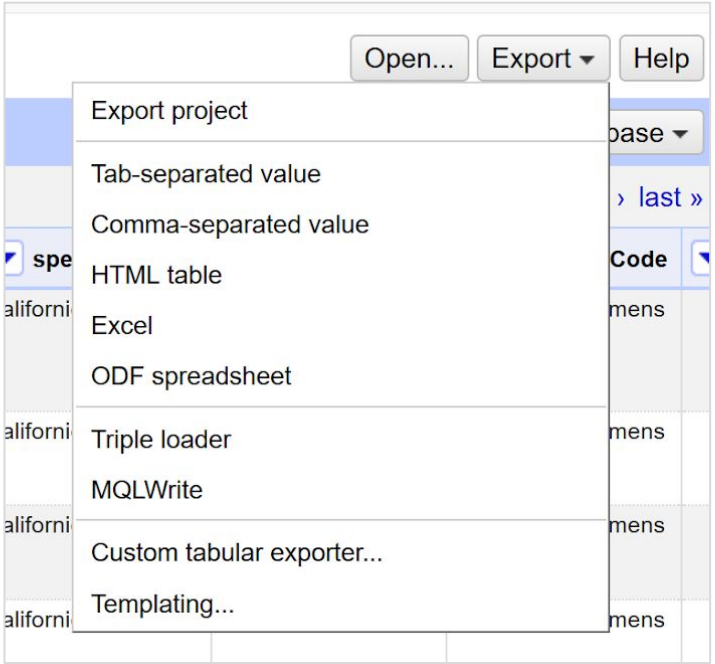
Open Refine - saves your steps

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



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.

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