Introduction to OpenRefine: Using Open Software to Weed and Manage a Government Documents Collection

ADAPTED FROM "WORKING WITH MESSY DATA IN OPENREFINE,"
IASSIST 2018 CONFERENCE, LEANNE TRIMBLE AND KELLY SCHULTZ,
CONCORDIA UNIVERSITY, CANADA

Eimmy Solis
University of Southern California
Social Sciences Data Librarian
eimmysol@usc.edu

tinyurl.com/FDLC2019OPENREFINE

Agenda

- Background
- What is OpenRefine?
- OpenRefine Setup
- Demonstrations and Hands-on Practice
- Additional Helpful Resources

Learning Objectives

Participants will be able to use OpenRefine to:

- Search, sort, and filter data in a variety of ways
- Restructure and manipulate a dataset
- Perform basic data cleanup

Background

-NEW GOV DOCS LIBRARIAN
-NO PRIOR WEEDING EXPERIENCE

ALL SLIDES, HANDOUTS AND DATASET HERE:

tinyurl.com/FDLC2019OPENREFINE

Installing OpenRefine

OpenRefine is installed locally on your computer, even though it uses a web browser as the user interface.

A copy of your data files are saved locally to your computer.

What is Messy and Clean Data?

	A	B
1	Customer Name	
3	John K. Doe Jr.	Doe, John
4	Mr. Doe, John	Doe, John
5	Jane A. Smith	Smith, Jane
6	MS. Jane Smith	Smith, Jane
7	Smith, Jane	Smith, Jane
8	Dr Anthony R Von Fange III	Von Fange, Anthony
9	Peter Tyson	Tyson, Peter
10	Dan E. Williams	Williams, Dan
11	James Davis Sr.	Davis, James
12	James J. Davis	Davis, James
13	Mr. Donald Edward Miller	Miller, Donald
14	Miller, Donald	Miller, Donald
15	Rajesh Krishnan	Krishnan, Rajesh
16	Daniel Chen	Chen, Daniel

What is OpenRefine?

Open source tool for working with messy data to clean and transform it from one format to another.

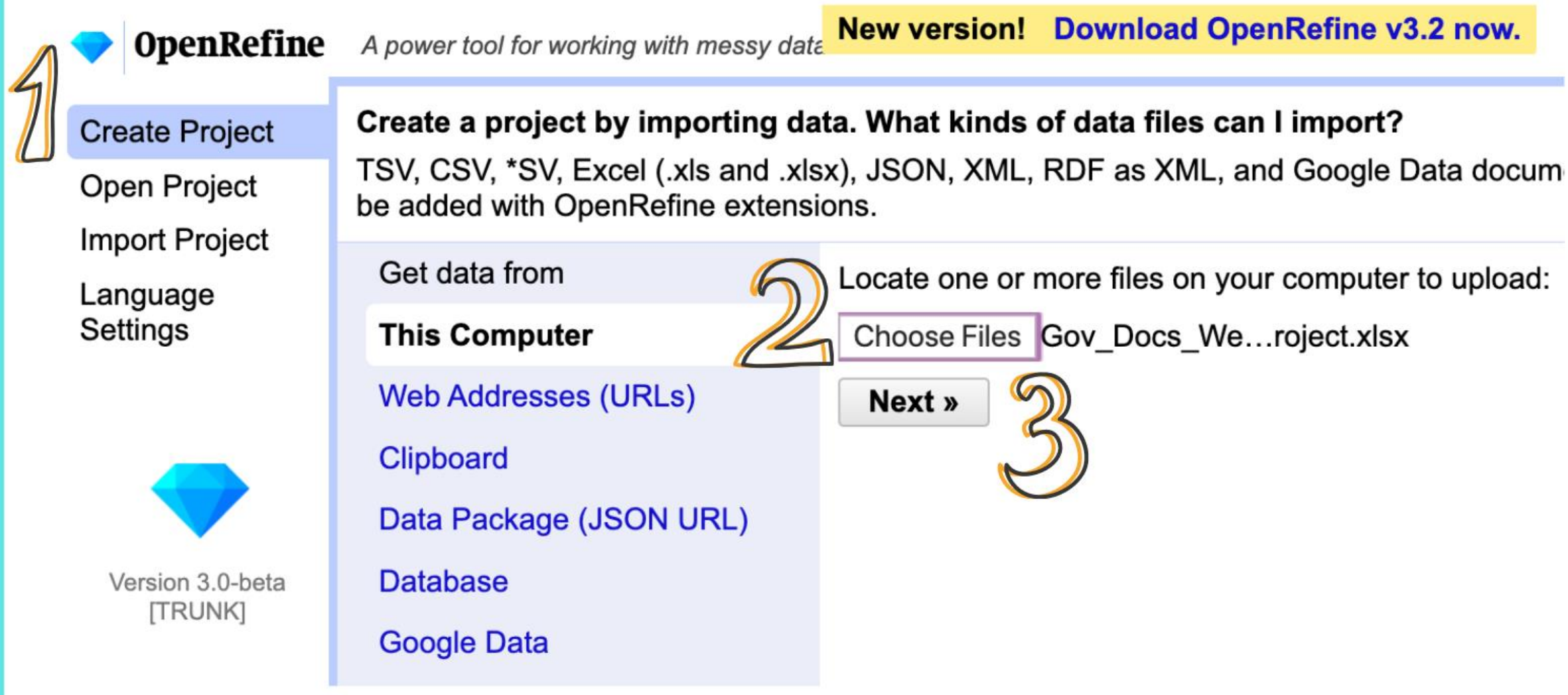
Why OpenRefine?

VS

Demonstrations & Hands-on Practice

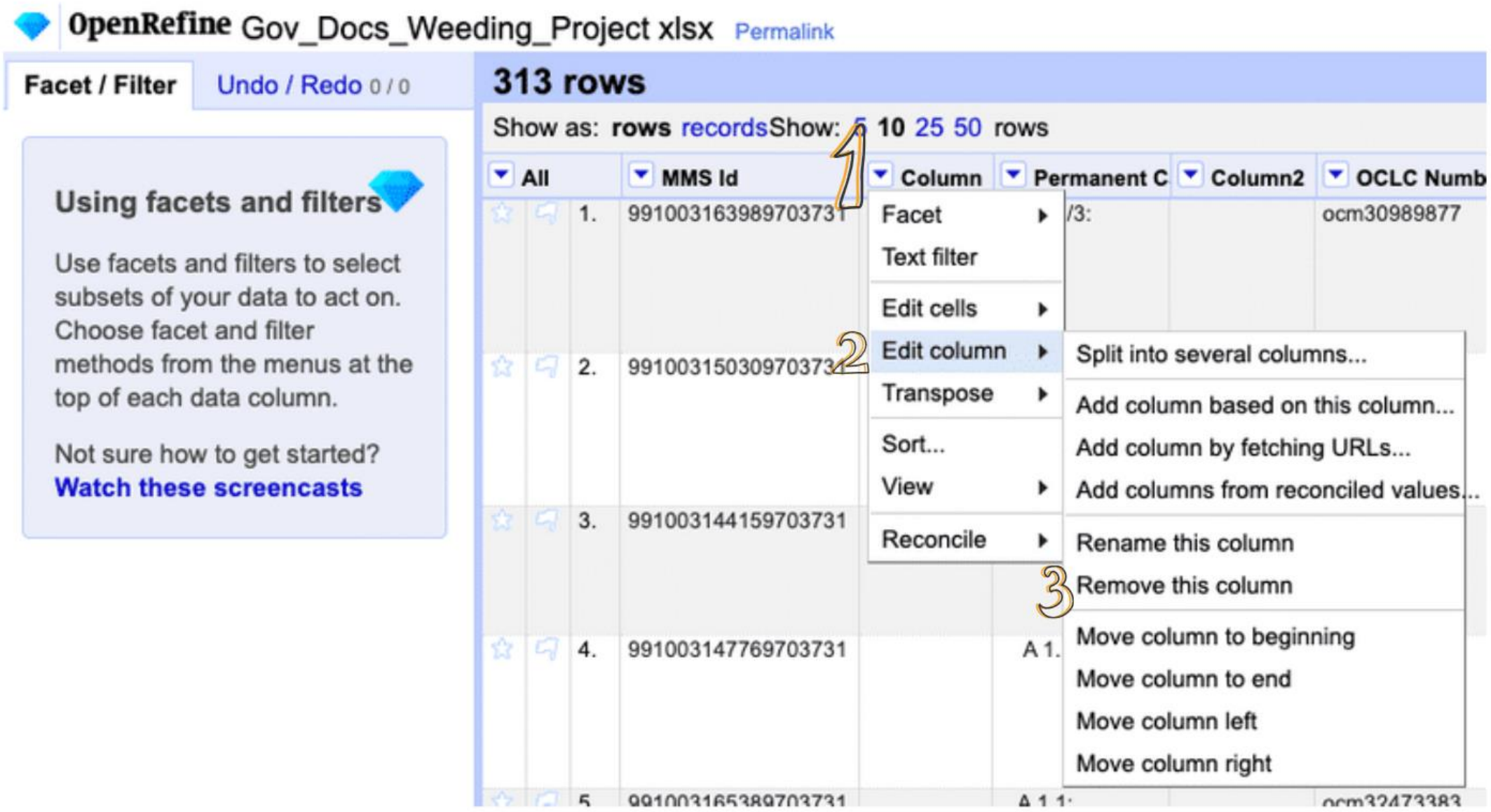
tinyurl.com/FDLC2019OPENREFINE

IMPORTING A DATASET INTO OPENREFINE

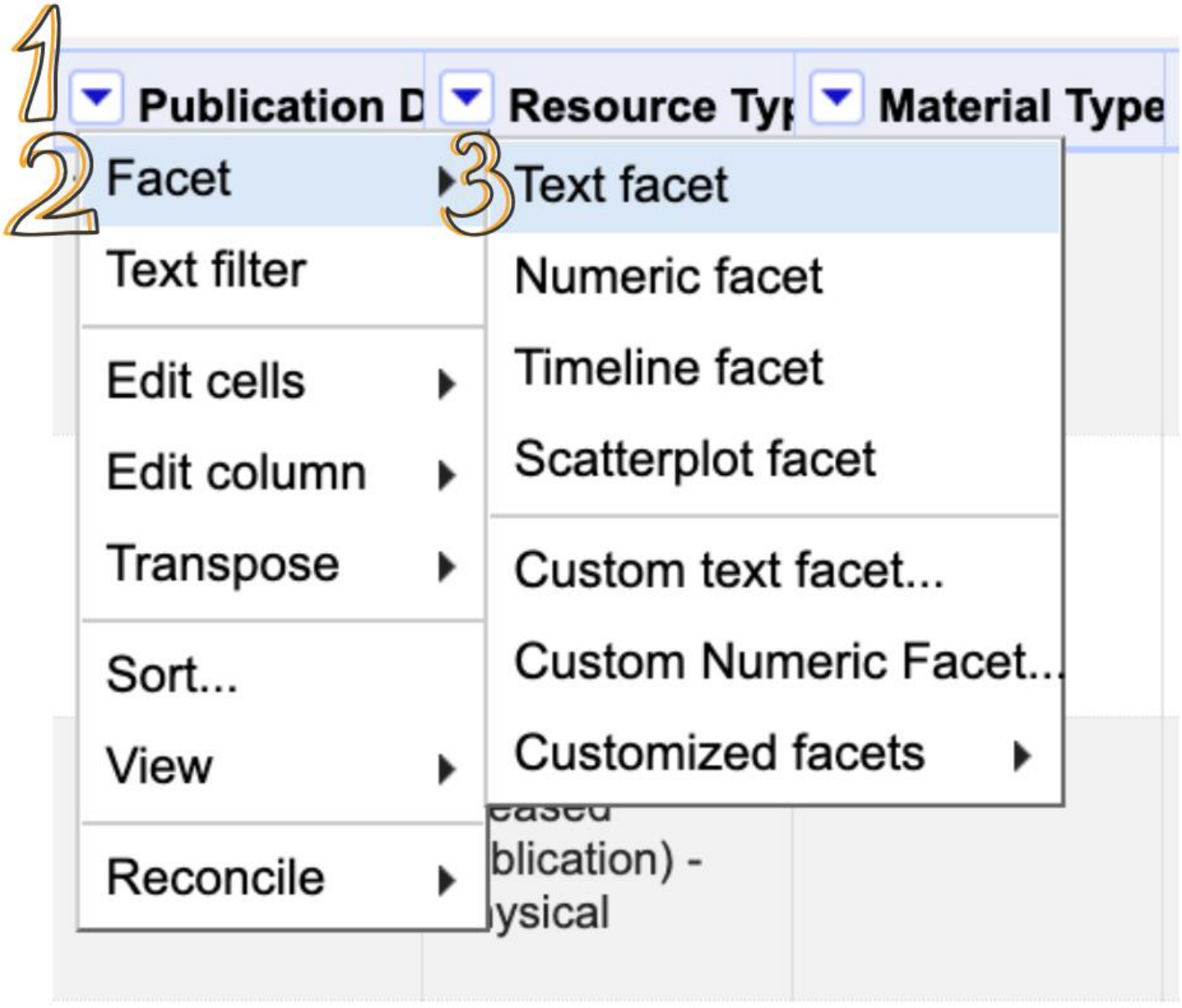


IMPORTING A DATASET INTO OPENREFINE

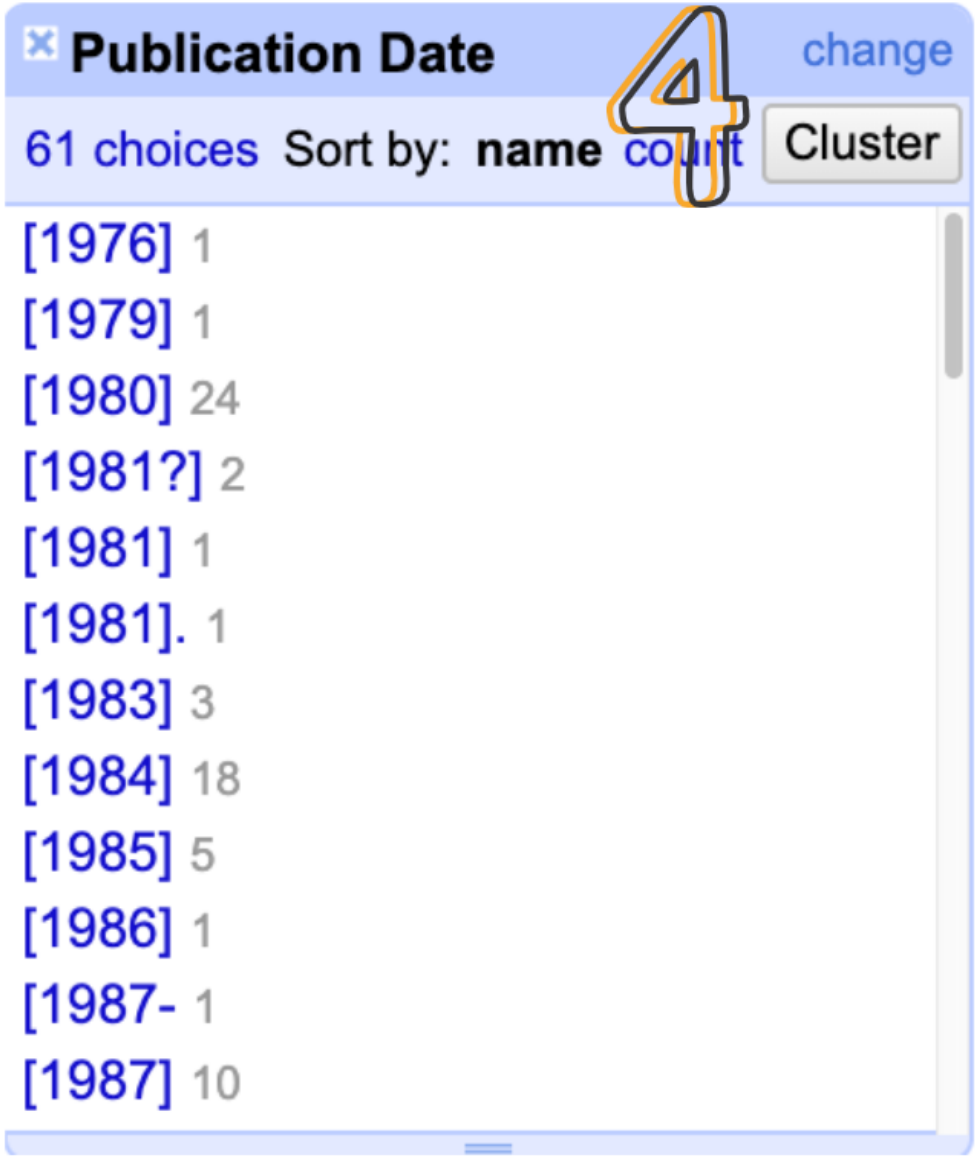
REMOVING A COLUMN



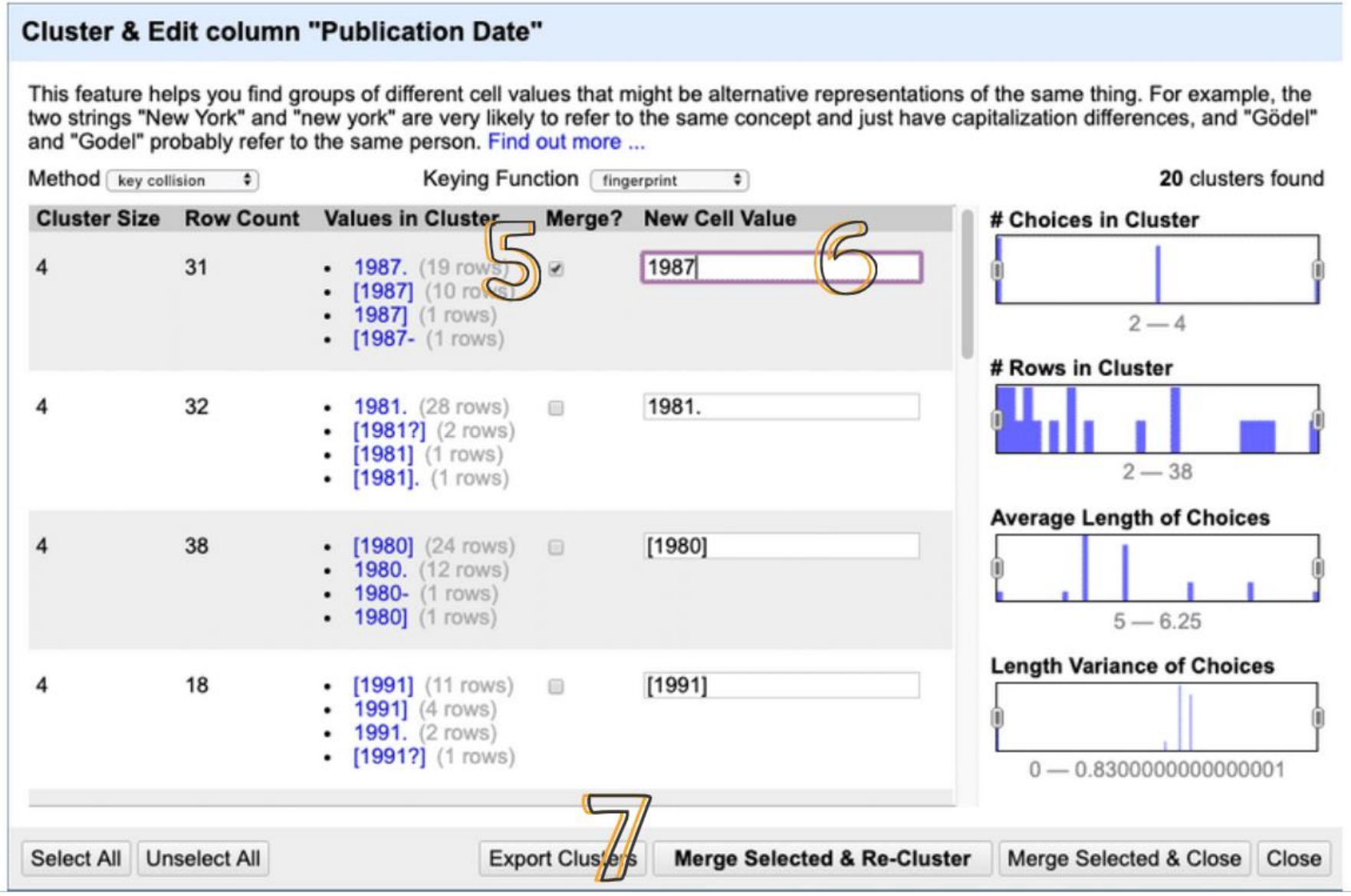
CLUSTERING


1
2
3
Publication D
Resource Typ
Material Type
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom Numeric Facet...
Customized facets

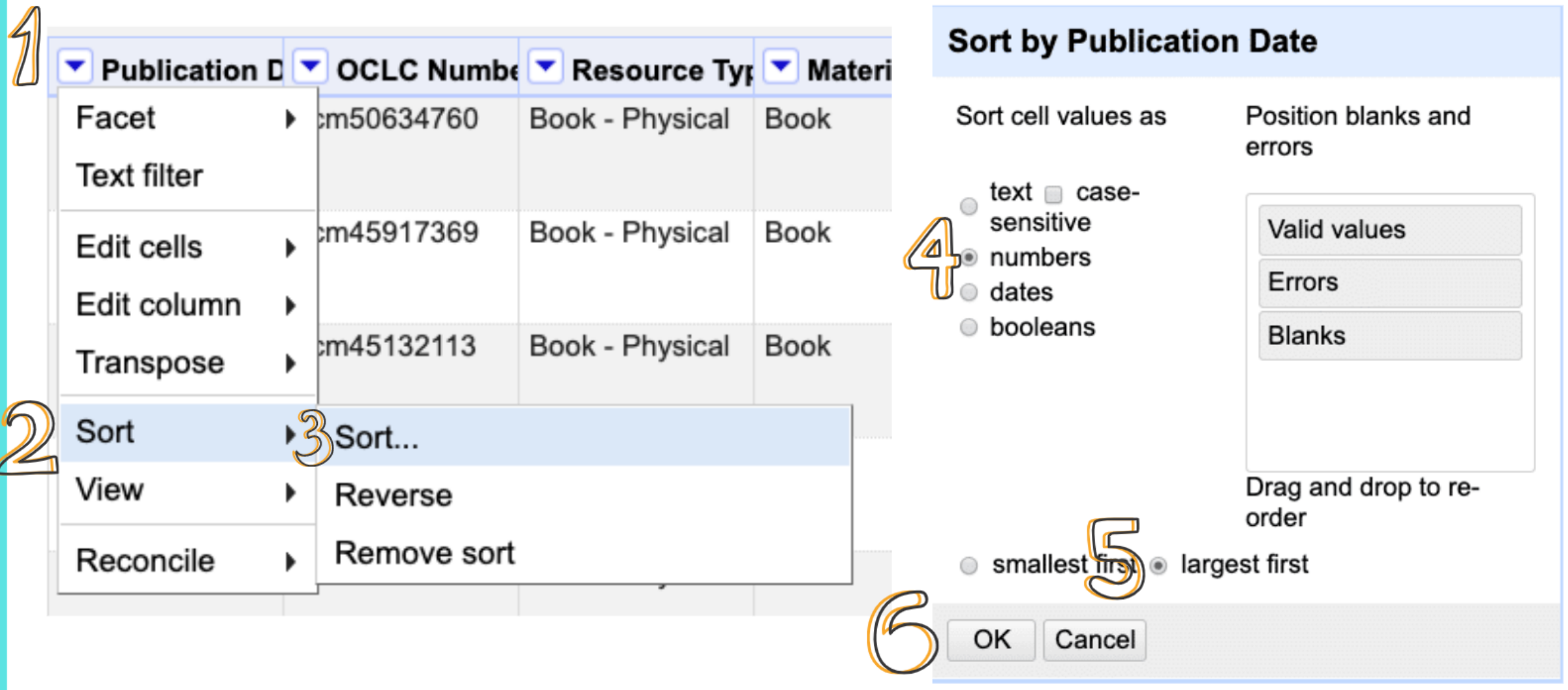
CLUSTERING


Publication Date
change
61 choices
Sort by: name count
Cluster
[1976] 1
[1979] 1
[1980] 24
[1981?] 2
[1981] 1
[1981]. 1
[1983] 3
[1984] 18
[1985] 5
[1986] 1
[1987- 1
[1987] 10

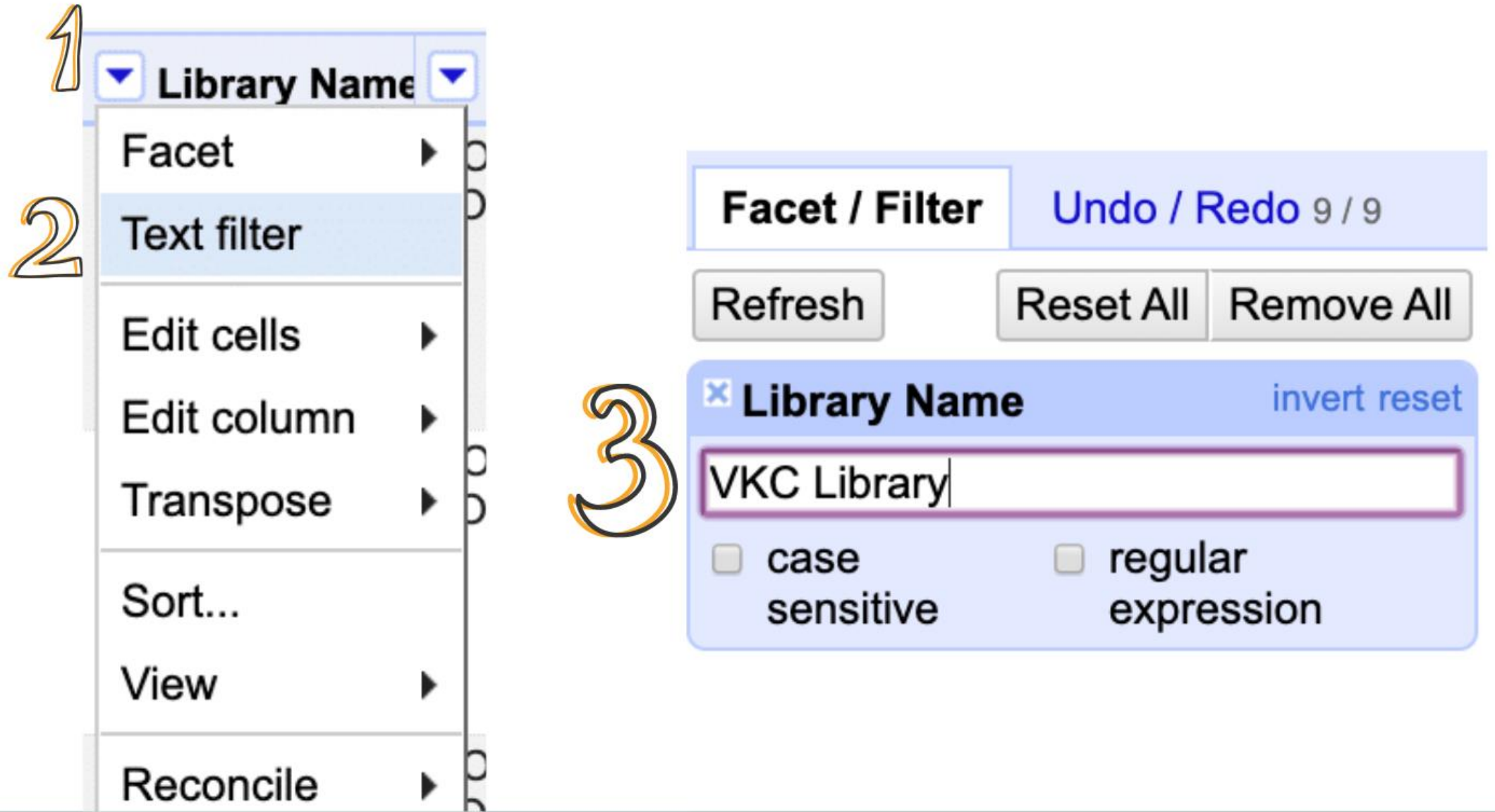
CLUSTERING


Cluster & Edit column "Publication Date"
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...
Method key collision
Keying Function fingerprint
20 clusters found
Cluster Size
Row Count
Values in Cluster
Merge?
New Cell Value
4
31
1987. (19 rows)
[1987] (10 rows)
1987] (1 rows)
[1987- (1 rows)
1987
4
32
1981. (28 rows)
[1981?] (2 rows)
[1981] (1 rows)
[1981]. (1 rows)
1981.
4
38
[1980] (24 rows)
1980. (12 rows)
1980- (1 rows)
1980] (1 rows)
[1980]
4
18
[1991] (11 rows)
1991] (4 rows)
1991. (2 rows)
[1991?] (1 rows)
[1991]
5
6
7
Choices in Cluster
2 — 4
Rows in Cluster
2 — 38
Average Length of Choices
5 — 6.25
Length Variance of Choices
0 — 0.8300000000000001
Select All
Unselect All
Export Clusters
Merge Selected & Re-Cluster
Merge Selected & Close
Close

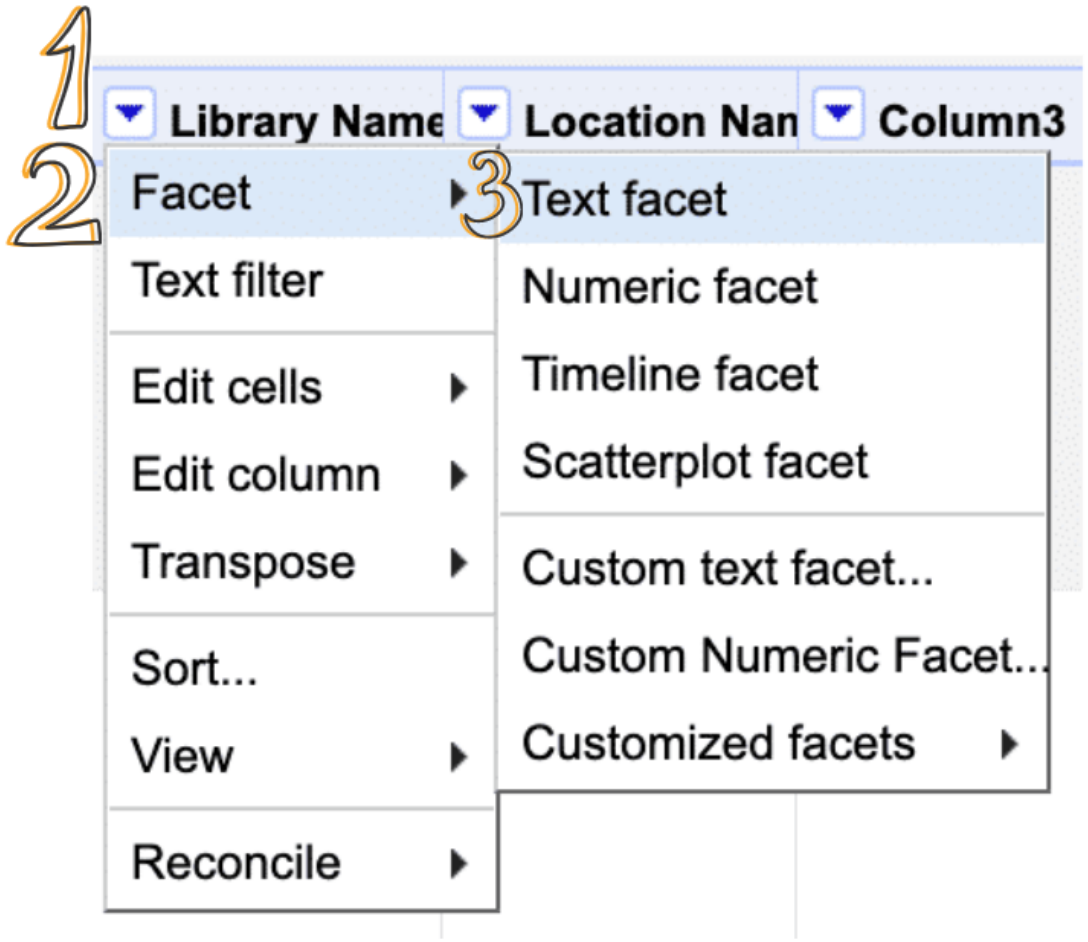
Sort

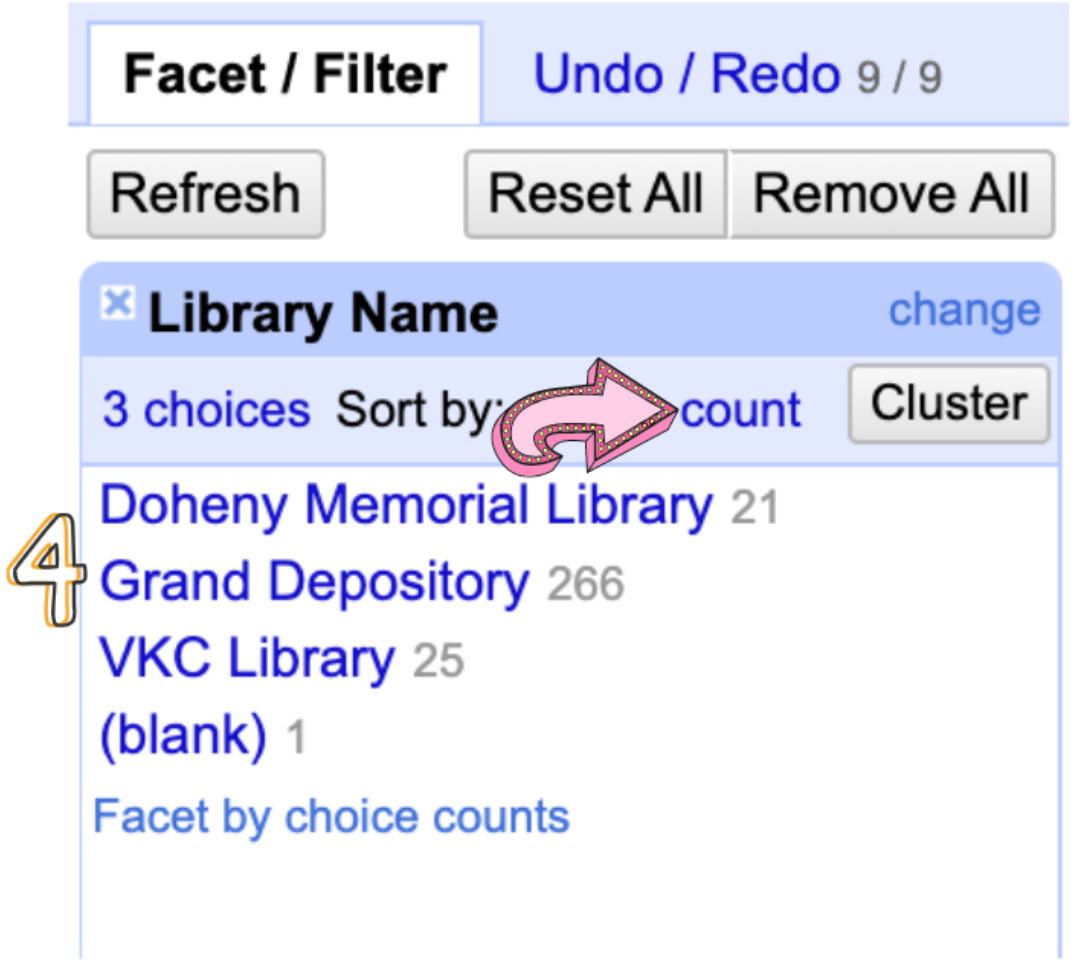

1
Publication D
OCLC Numbe
Resource Typ
Materi
cm50634760
Book - Physical
Book
cm45917369
Book - Physical
Book
cm45132113
Book - Physical
Book
Facet
Text filter
Edit cells
Edit column
Transpose
2
Sort
View
Reconcile
3
Sort...
Reverse
Remove sort
Sort by Publication Date
Sort cell values as
4
text
case-sensitive
numbers
dates
booleans
Position blanks and errors
Valid values
Errors
Blanks
Drag and drop to re-order
5
smallest first
largest first
6
OK
Cancel

FILTER

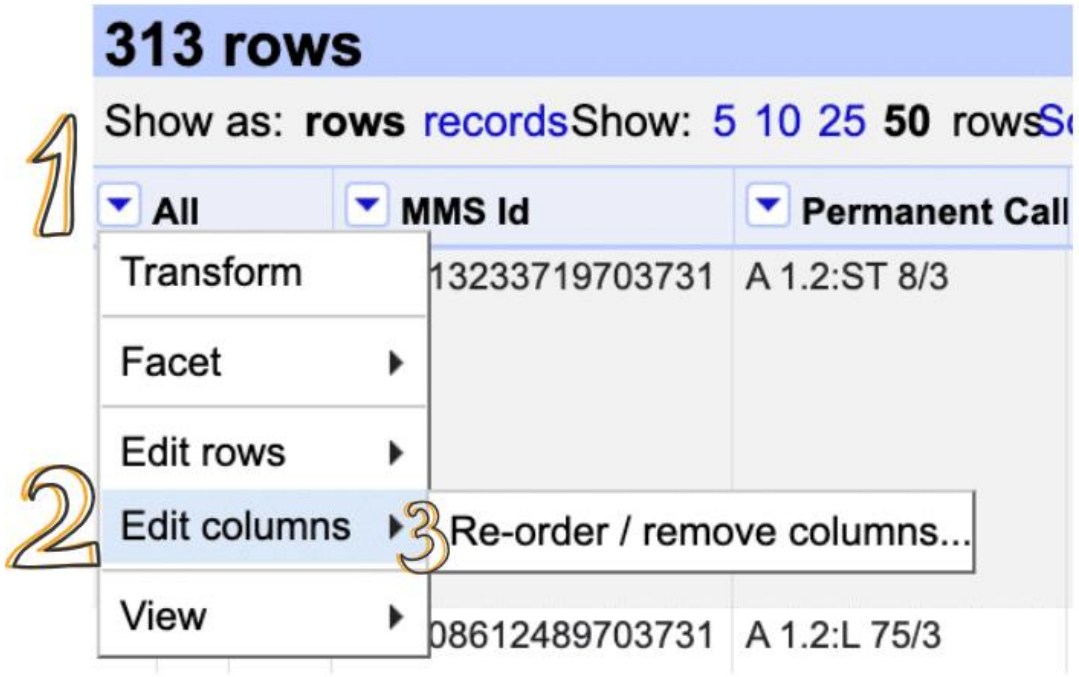

1
Library Name
Facet
2
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Facet / Filter
Undo / Redo 9 / 9
Refresh
Reset All
Remove All
3
Library Name
invert reset
VKC Library
case sensitive
regular expression

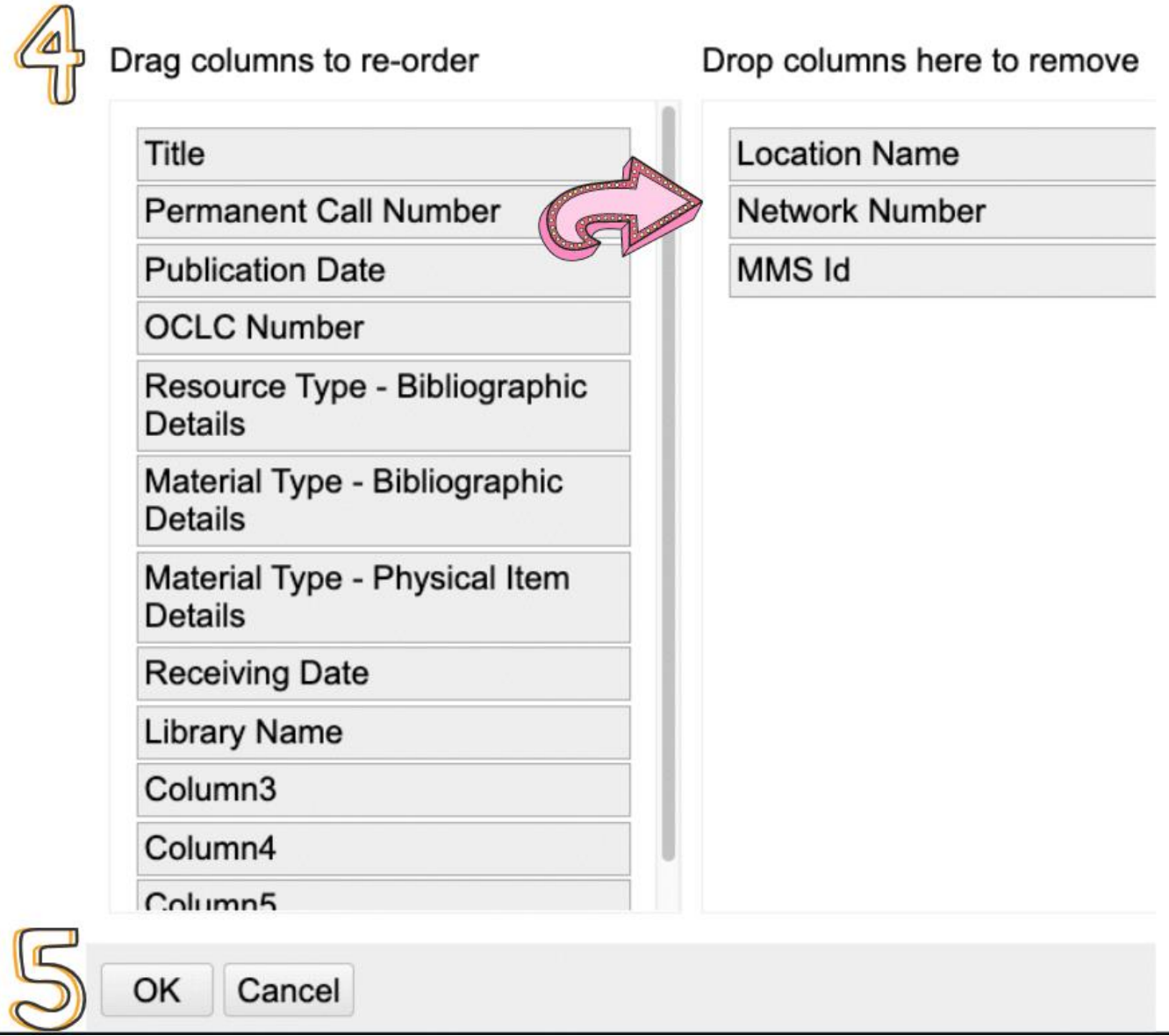
FACET


1
2
3
Library Name
Location Nan
Column3
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom Numeric Facet...
Customized facets


Facet / Filter
Undo / Redo 9 / 9
Refresh
Reset All
Remove All
Library Name
change
3 choices
Sort by:
count
Cluster
4
Doheny Memorial Library 21
Grand Depository 266
VKC Library 25
(blank) 1
Facet by choice counts

RE-ORDER / REMOVE COLUMNS





Closing OpenRefine

- Click on OpenRefine icon and type Command- Q.
- Wait until there’s a message that says the shutdown is complete.

Helpful Resoures

- OpenRefine documentation wiki: https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- OpenRefine Tutorial from John Little (Duke University): https://libjohn.github.io/openrefine/index.html
- Software Carpentry OpenRefine Workshop: https://data-lessons.github.io/library-openrefine/
- Cleaning Data with OpenRefine from the Programming Historian: https://programminghistorian.org/lessons/cleaning-data-with-openrefine
- Fetching and Parsing Data from the Web with OpenRefine from the Programming Historian: https://programminghistorian.org/lessons/fetch-and-parse-data-with-openrefine
- Regex Cheat Sheet: http://www.rexegg.com/regex-quickstart.html

Questions?

Eimmy Solis
eimmysol@usc.edu
tinyurl.com/FDLC2019OPENREFINE