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:

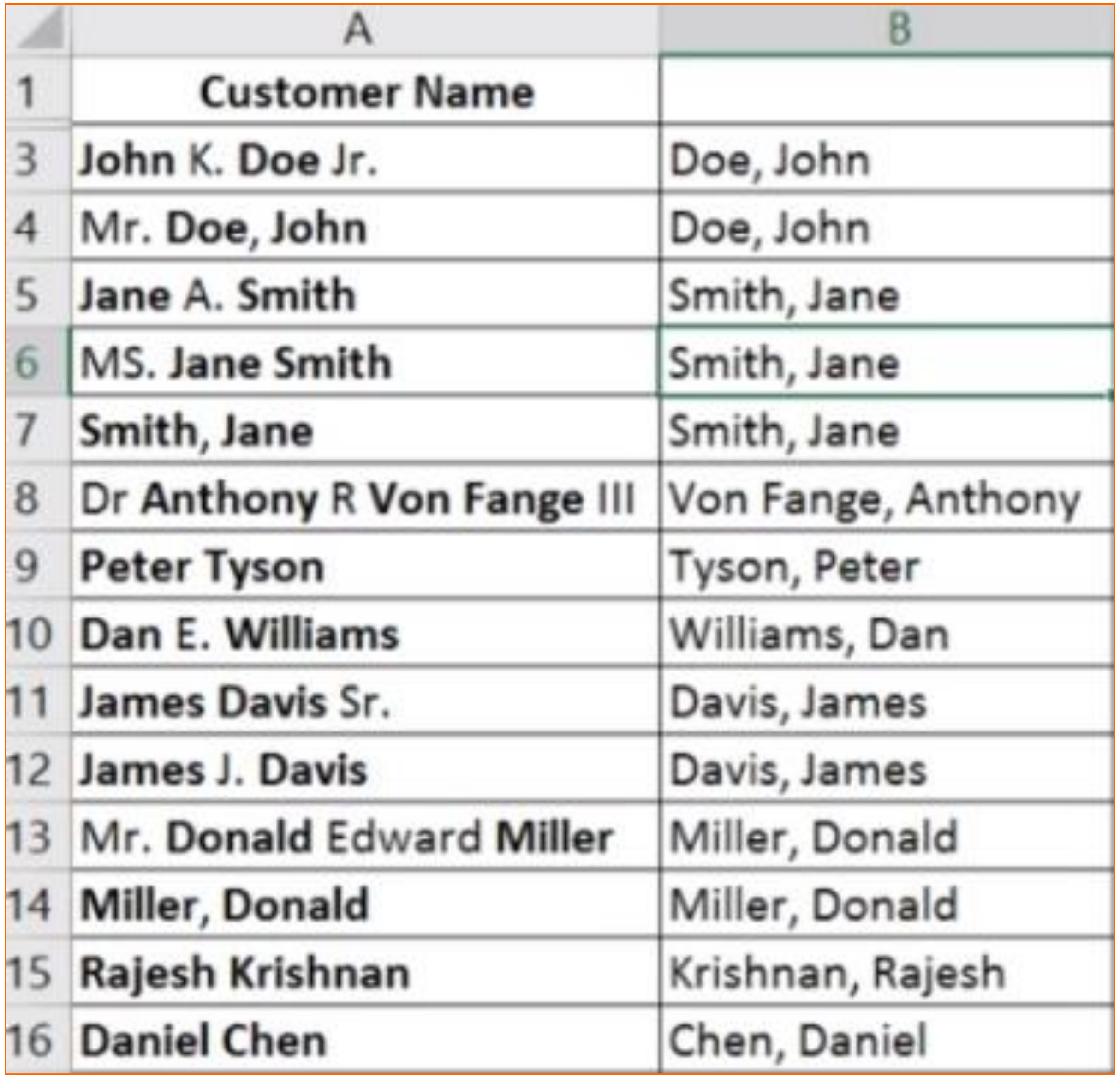
tinyurl.com/FDLC2019OPENREFINE

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?

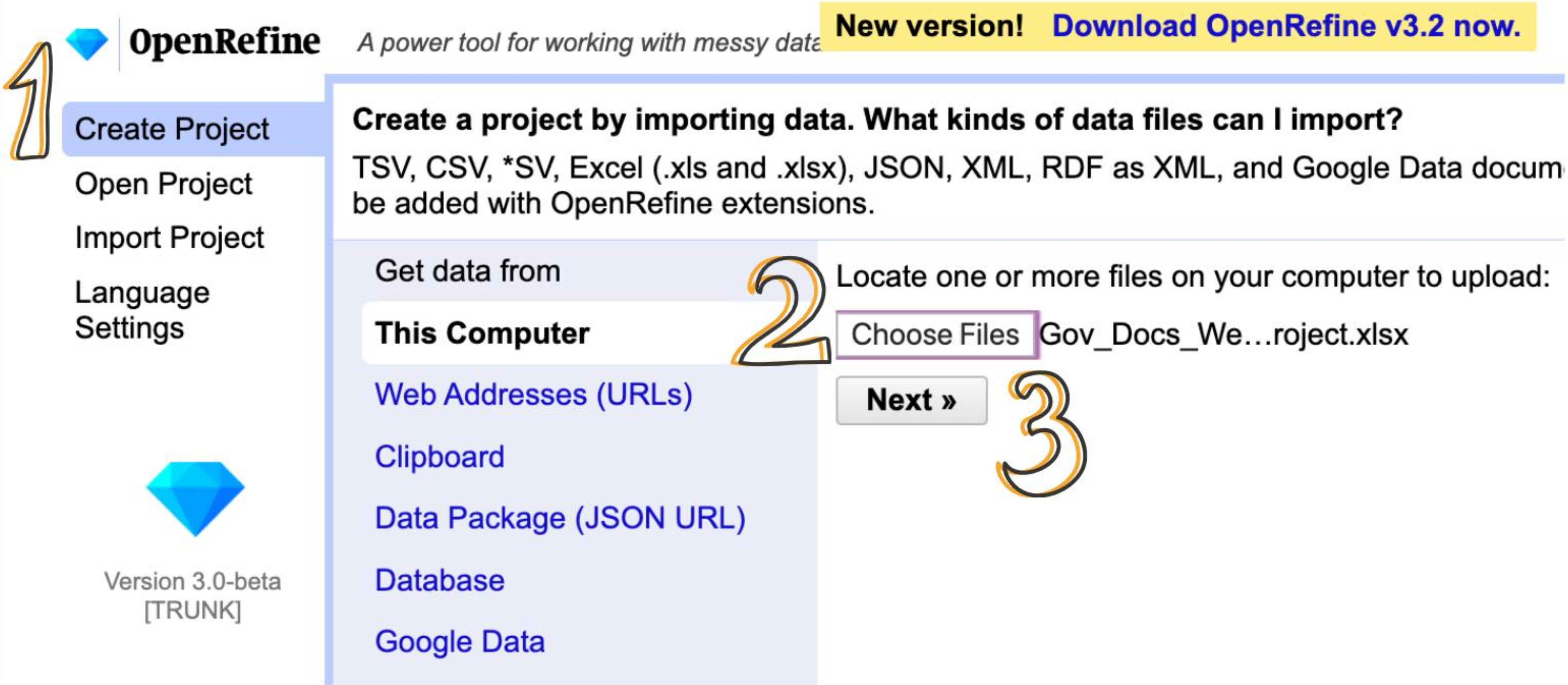
OPEN
Refine

vs

Demonstrations & Hands-on Practice

tinyurl.com/FDLC2019OPENREFINE

IMPORTING A DATASET INTO OPENREFINE



OpenRefine *A power tool for working with messy data*

New version! Download OpenRefine v3.2 now.

Create a project by importing data. What kinds of data files can I import?

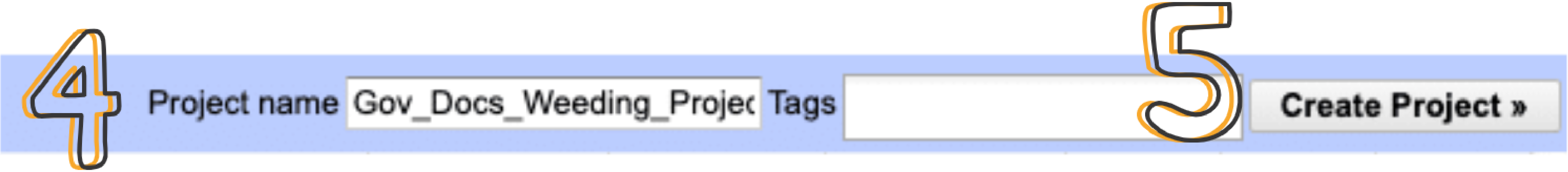
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data docum be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

Gov_Docs_We…roject.xlsx

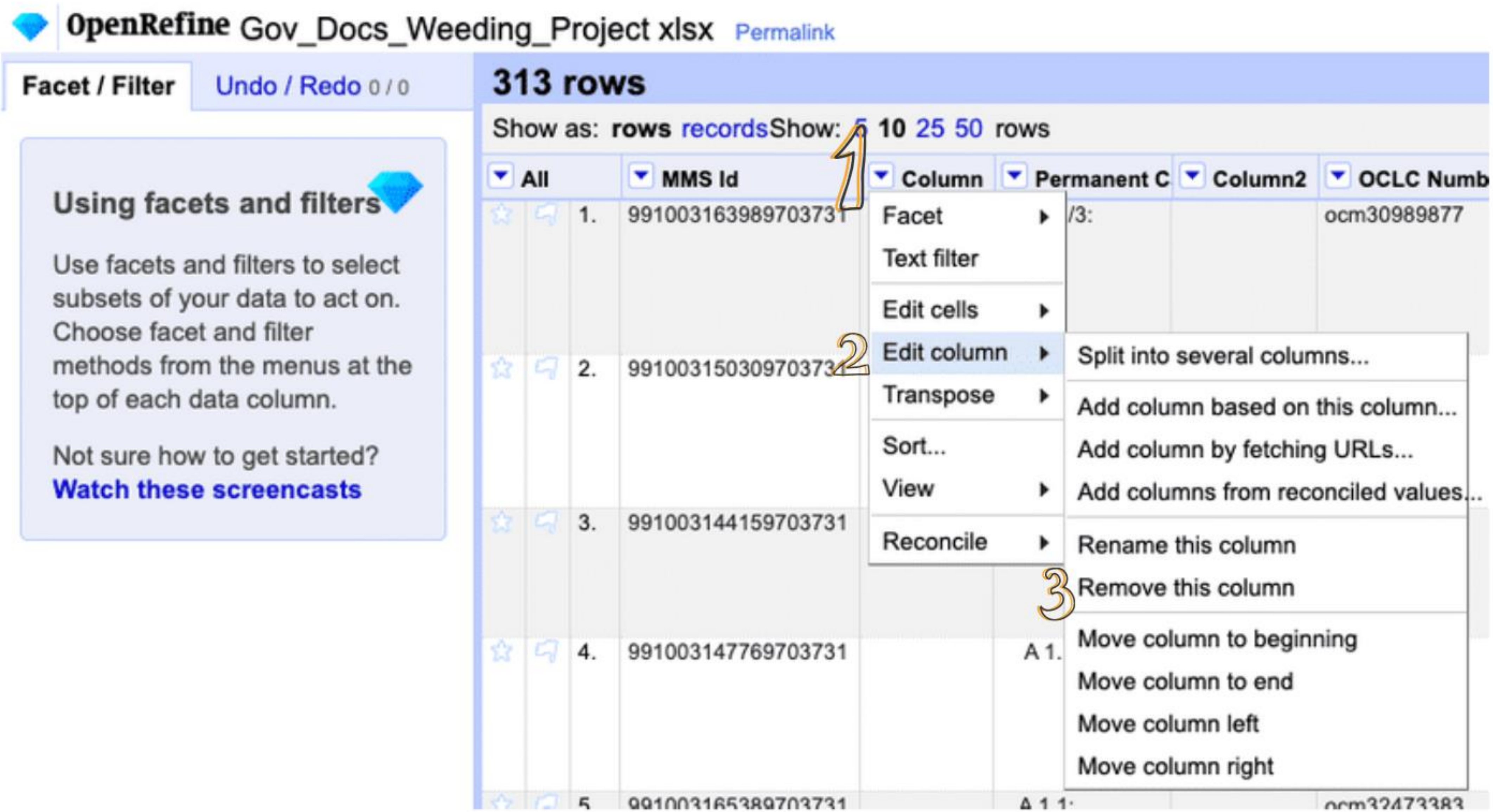
Version 3.0-beta [TRUNK]

IMPORTING A DATASET INTO OPENREFINE



Project name Gov_Docs_Weeding_Projec Tags

REMOVING A COLUMN

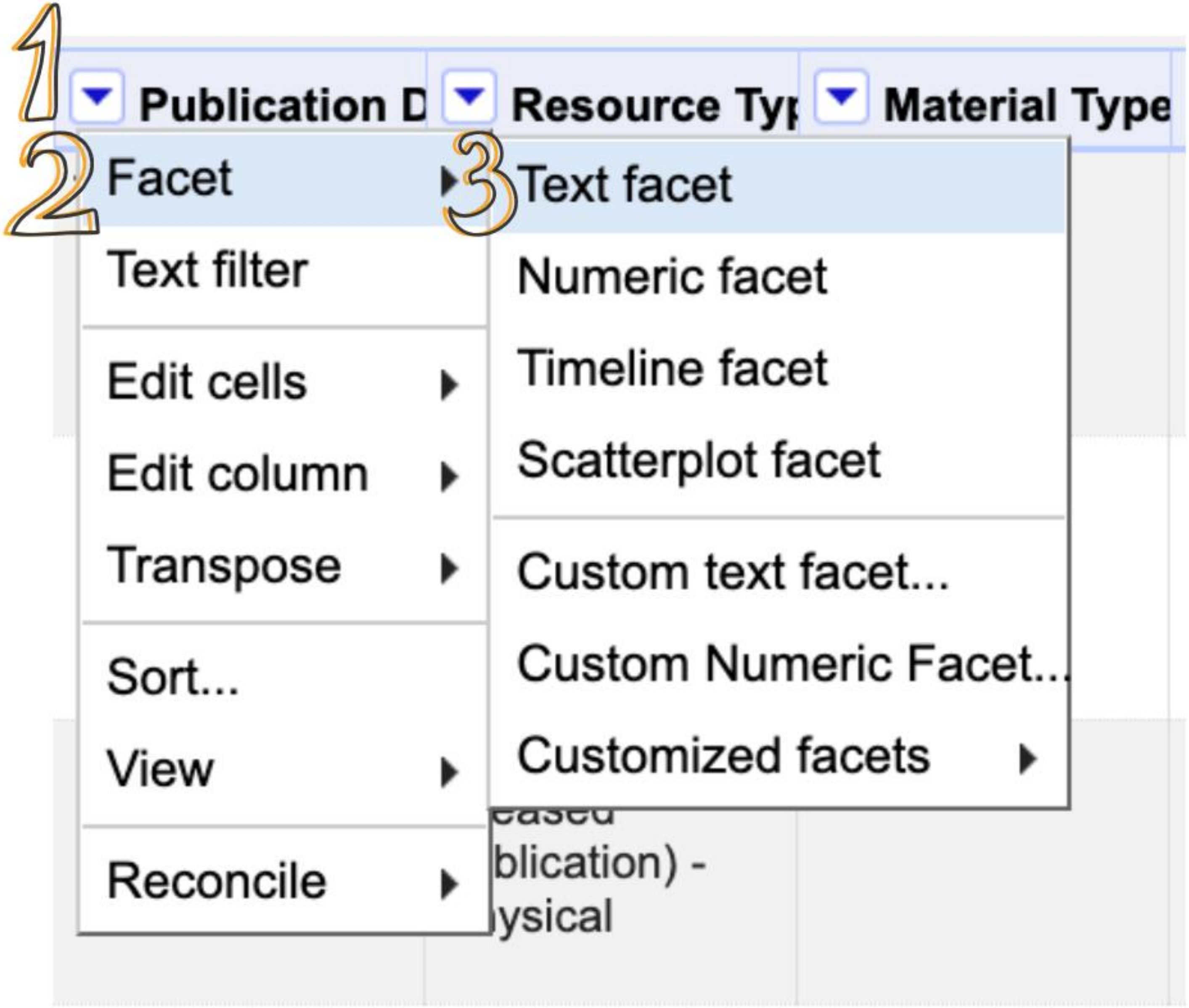


Using facets and filters

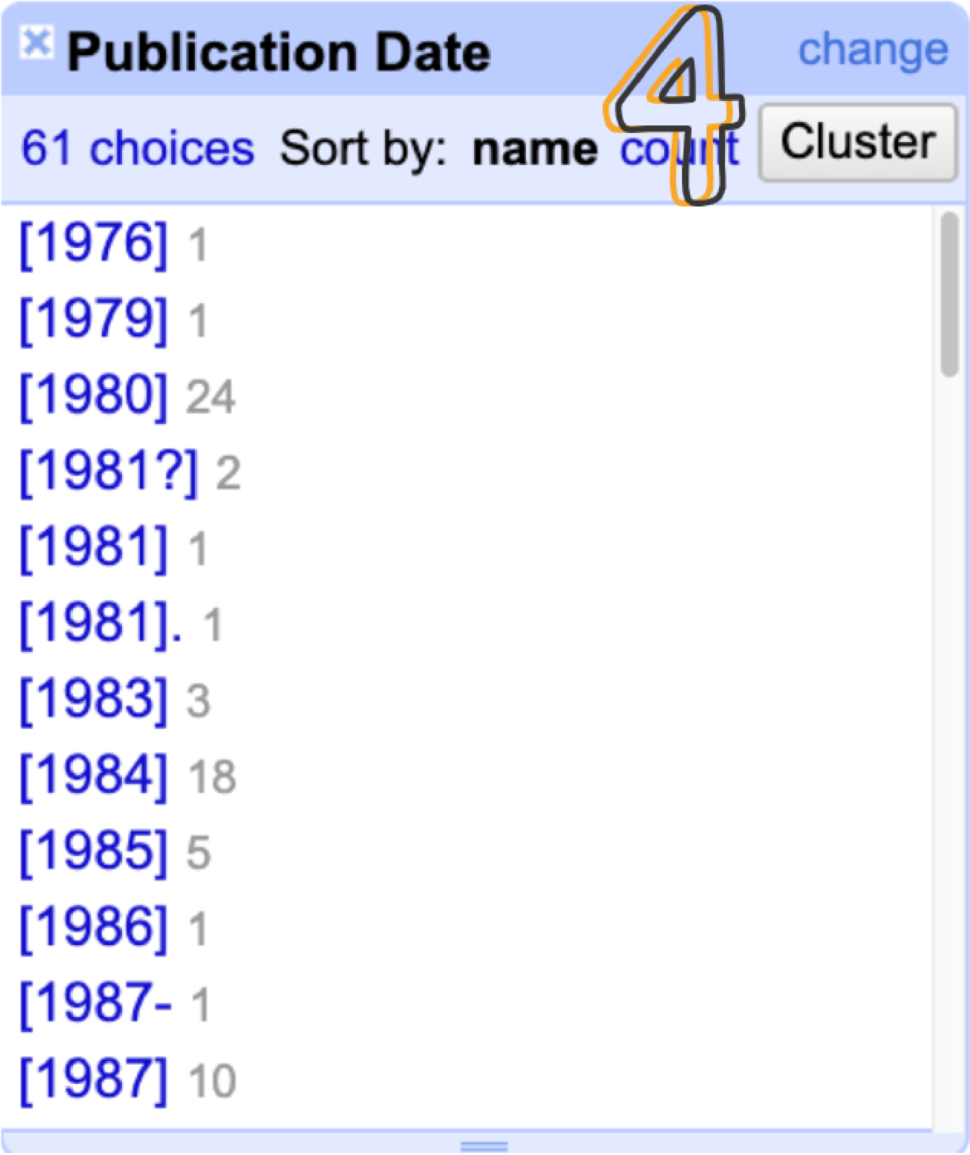
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? **Watch these screencasts**

CLUSTERING


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

CLUSTERING

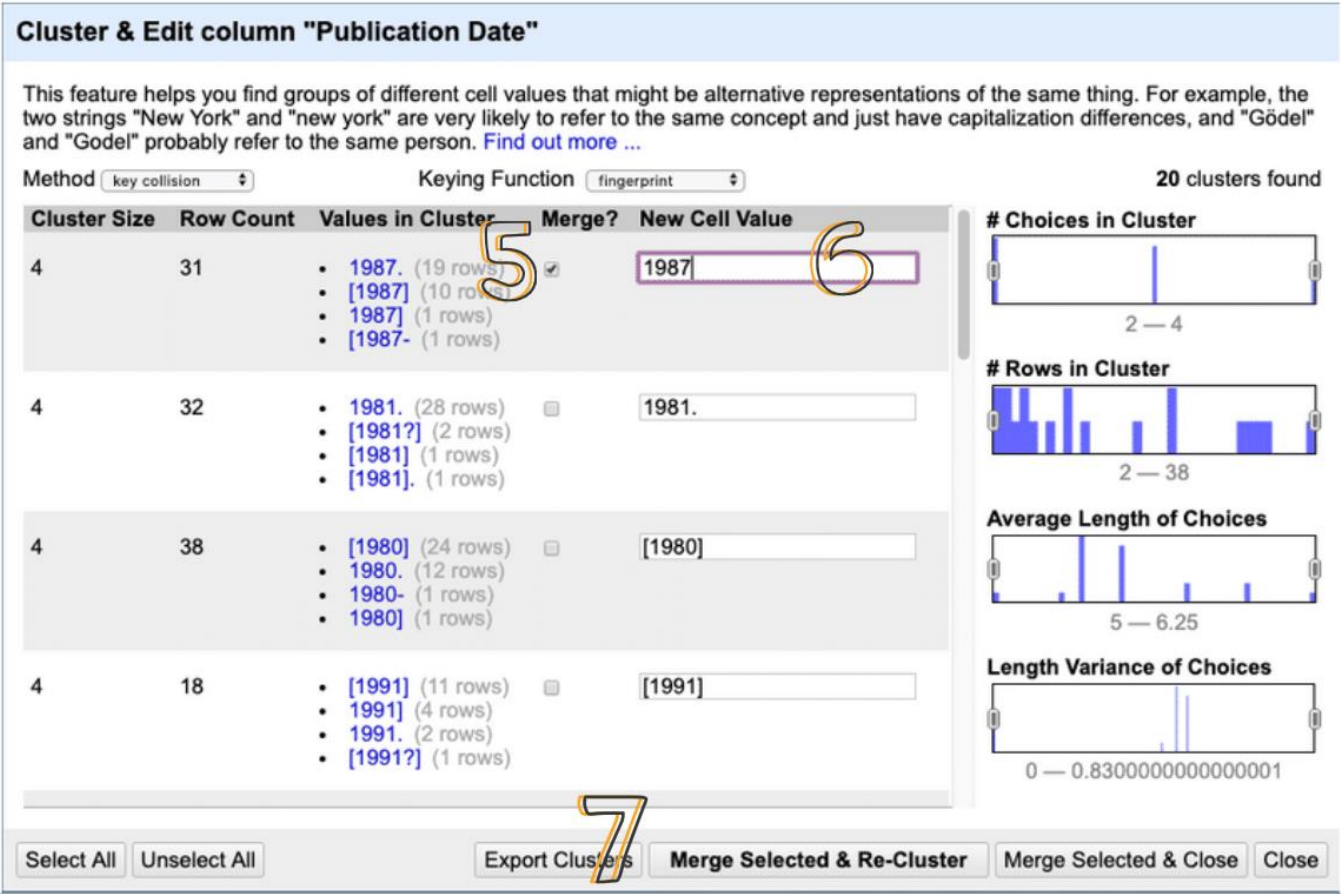


Publication Date

61 choices Sort by: **name** count

[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 ...

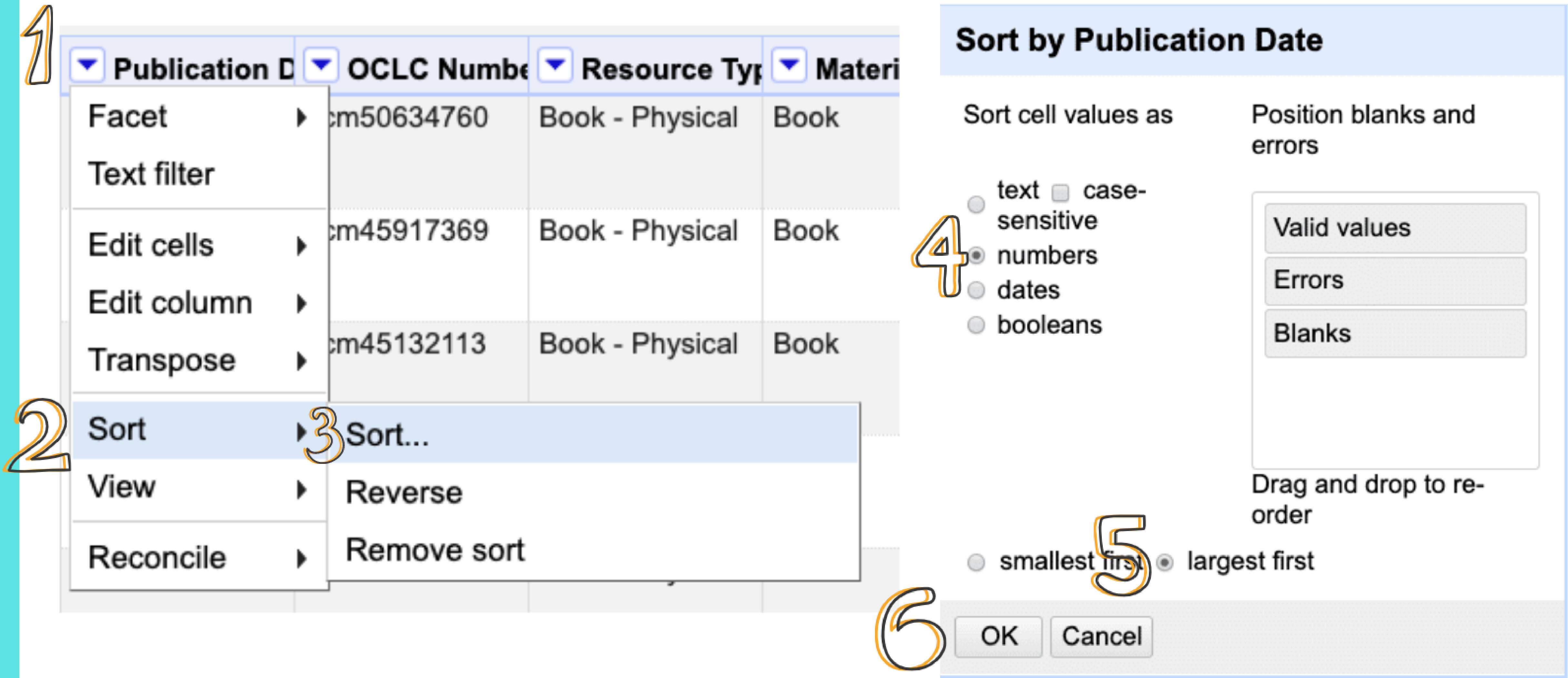
Method: key collision

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]

Sort



Sort by Publication Date

Sort cell values as

4
○ text □ case-sensitive
● numbers
○ dates
○ booleans

Position blanks and errors

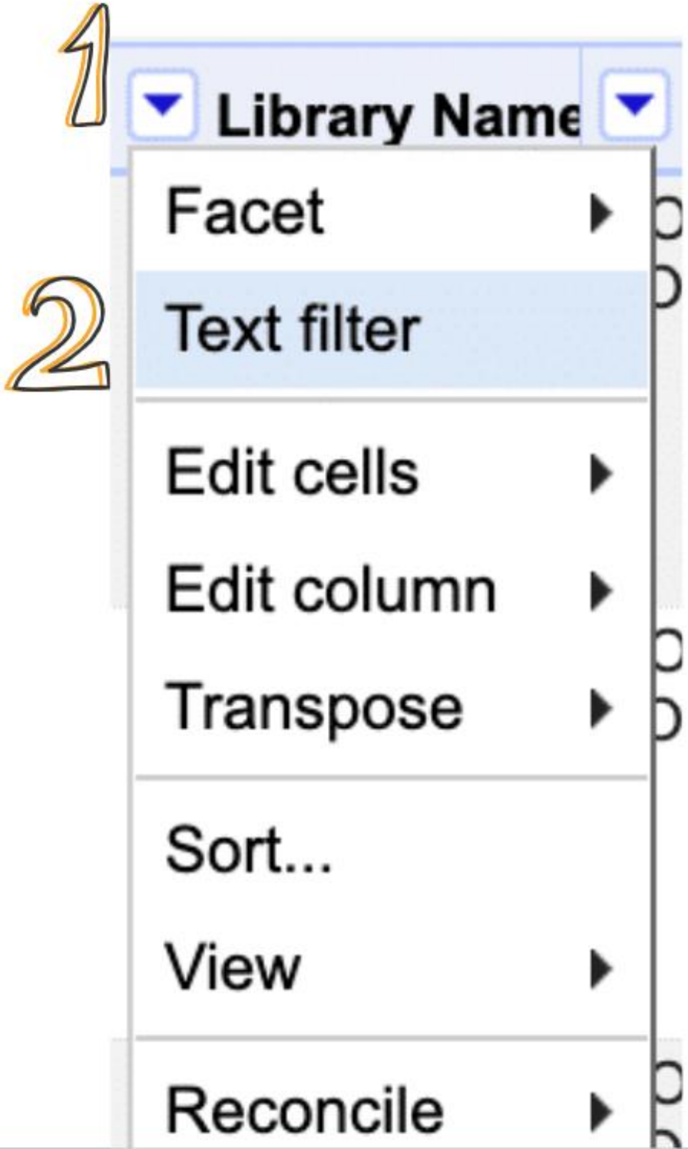
Valid values
Errors
Blanks

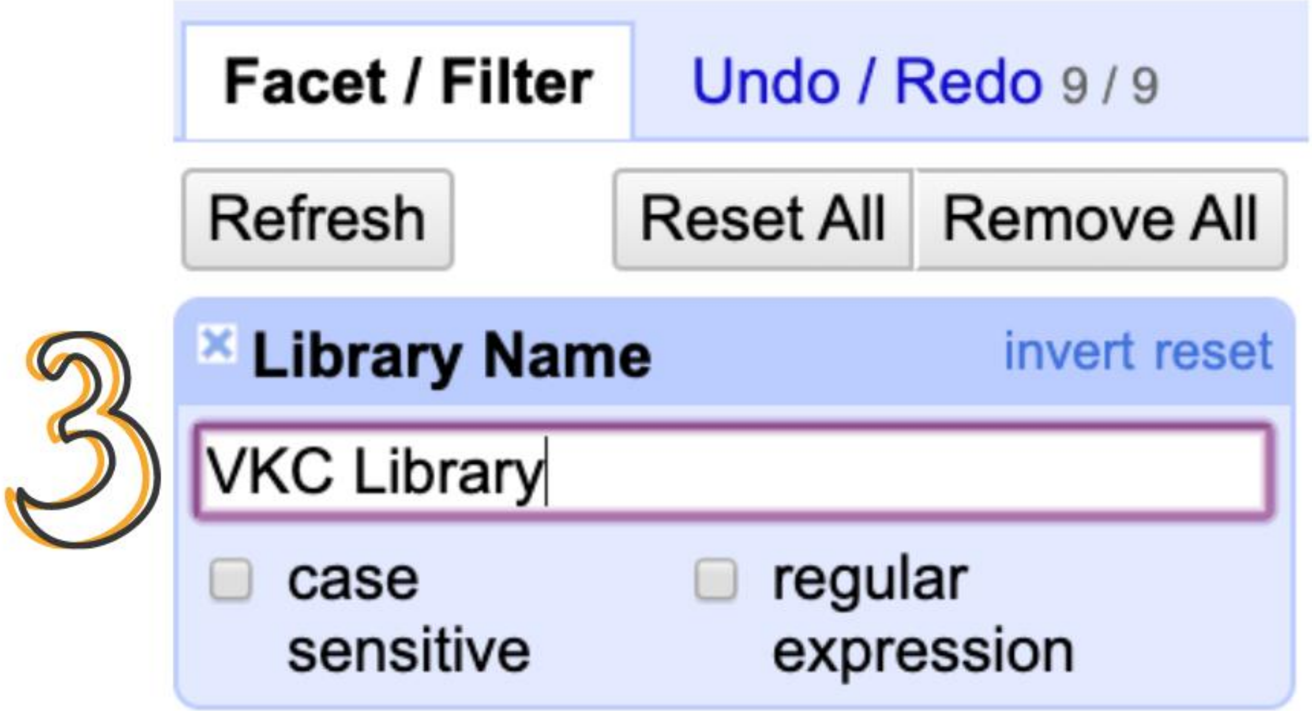
Drag and drop to re-order

5
○ smallest first ● largest first

6

FILTER

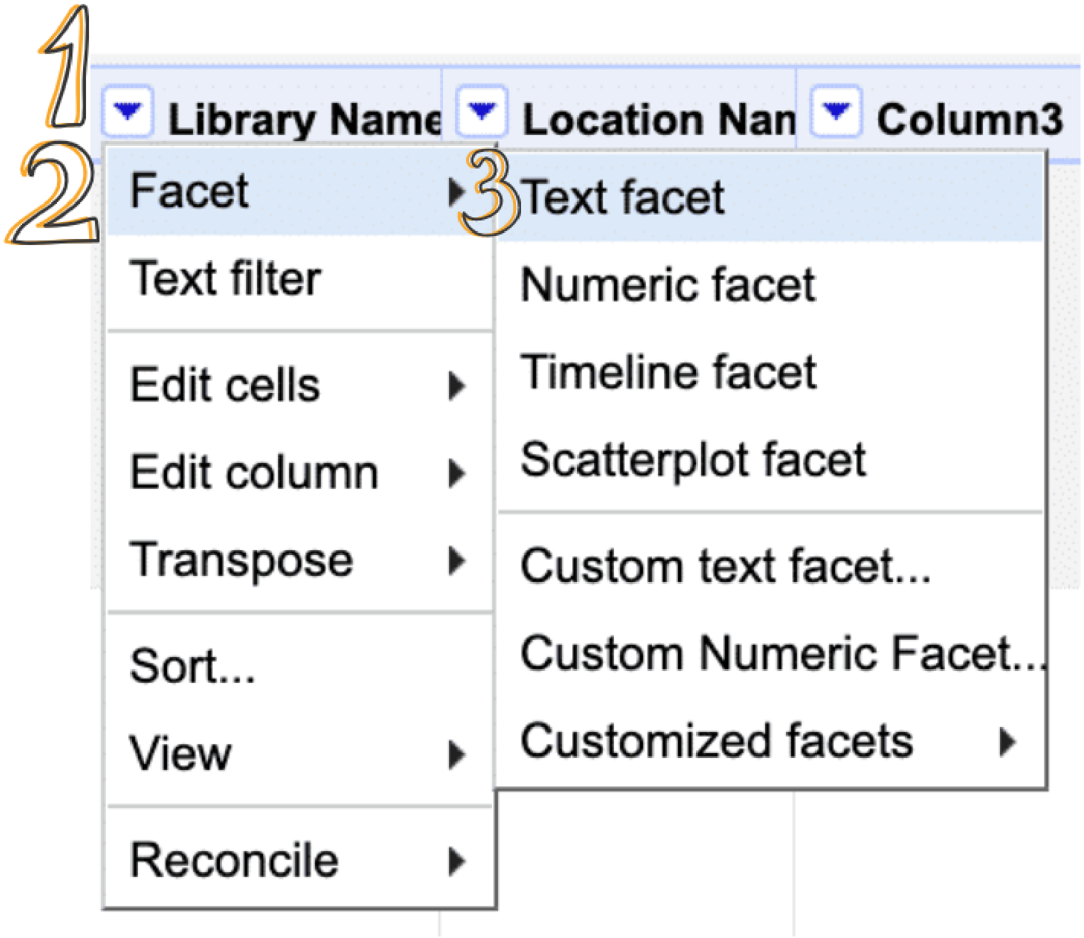


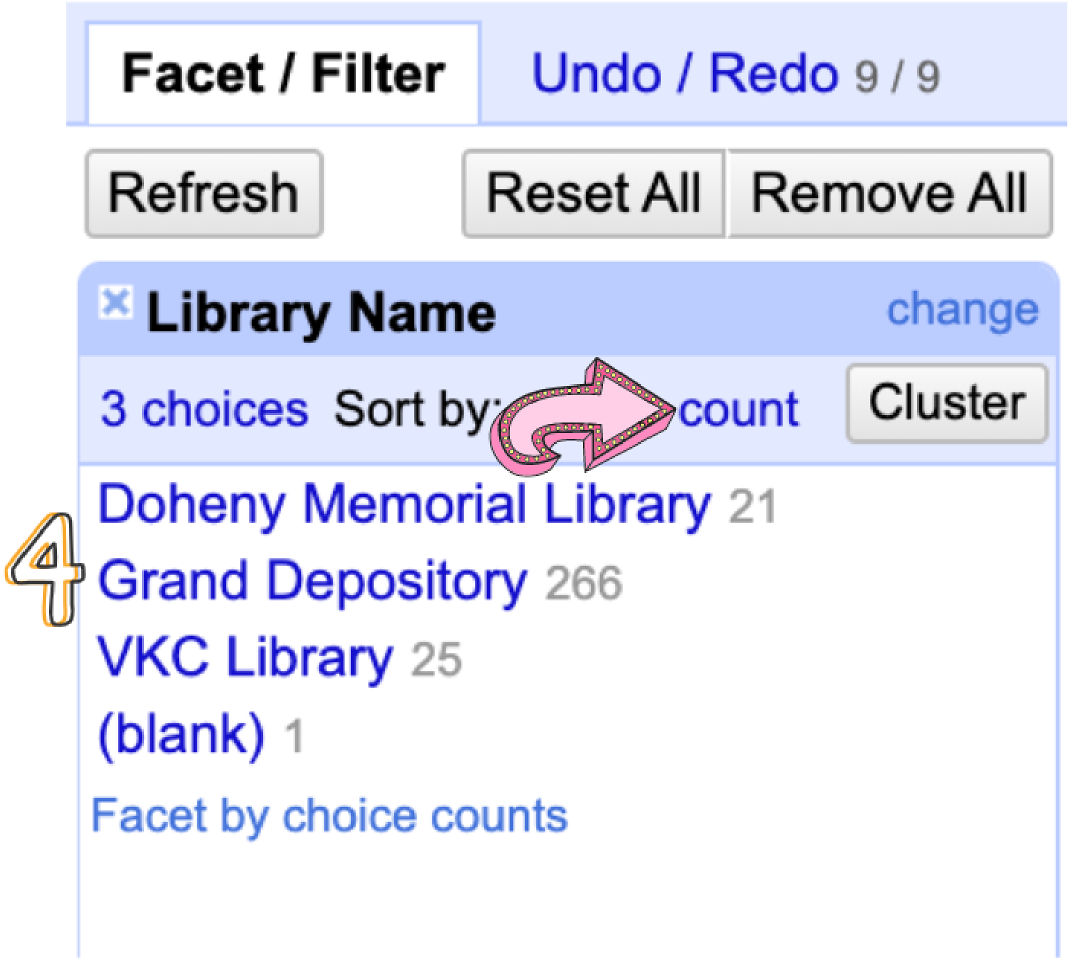


VKC Library

□ case sensitive □ regular expression

FACET



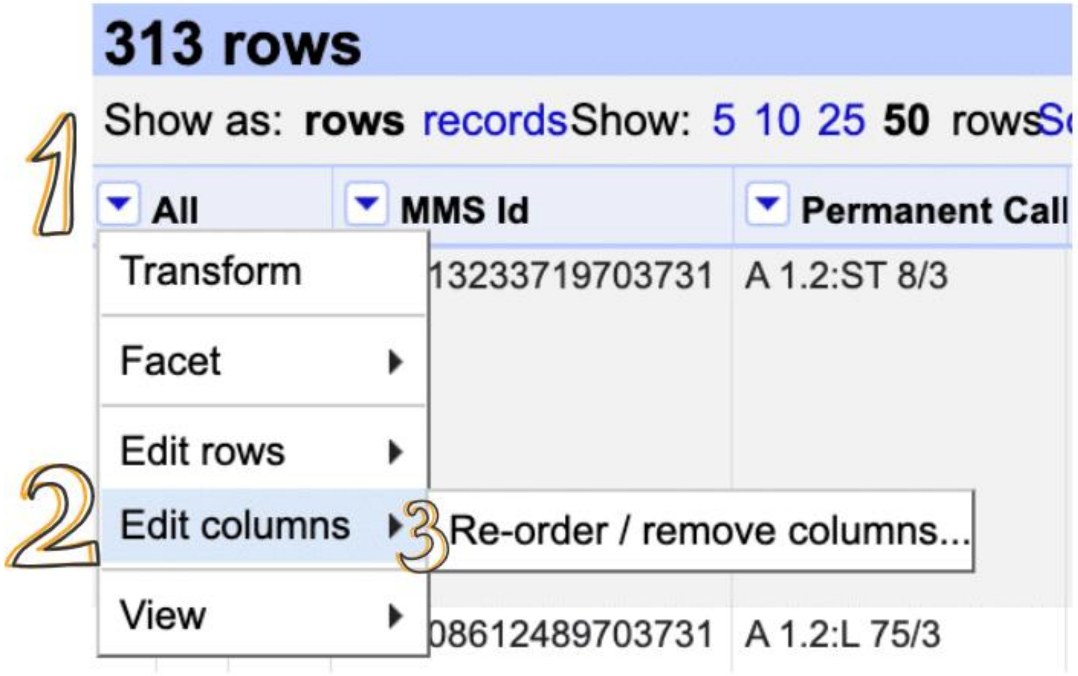


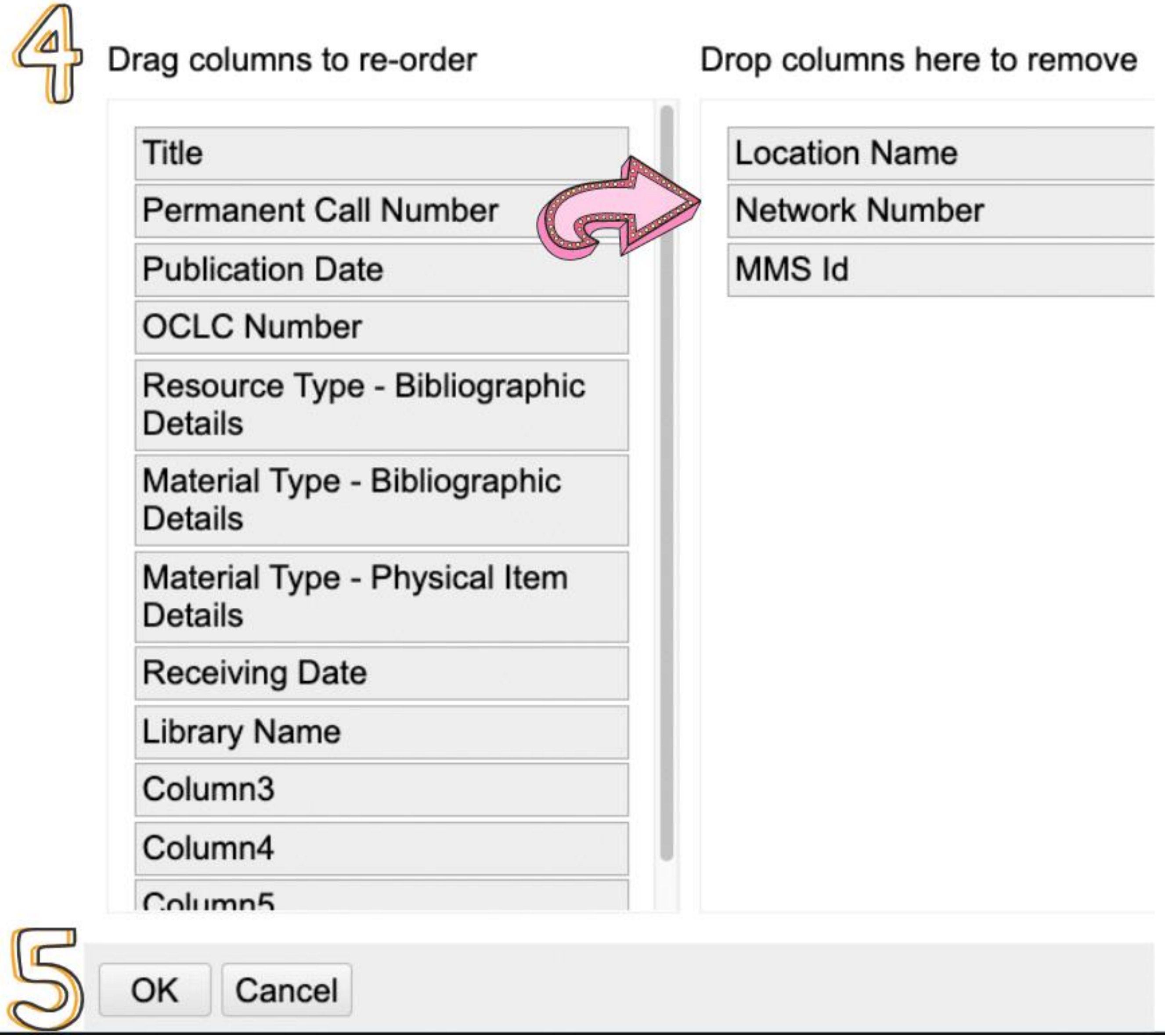
Library Name

3 choices Sort by: count

Doheny Memorial Library 21
Grand Depository 266
VKC Library 25
(blank) 1

RE-ORDER / REMOVE COLUMNS





Drag columns to re-order

- Title
- Permanent Call Number
- Publication Date
- OCLC Number
- Resource Type - Bibliographic Details
- Material Type - Bibliographic Details
- Material Type - Physical Item Details
- Receiving Date
- Library Name
- Column3
- Column4
- Column5

Drop columns here to remove

- Location Name
- Network Number
- MMS Id

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