

# 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](mailto:eimmysol@usc.edu)

[tinyurl.com/FDLC2019OPENREFINE](https://tinyurl.com/FDLC2019OPENREFINE)

# Agenda

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

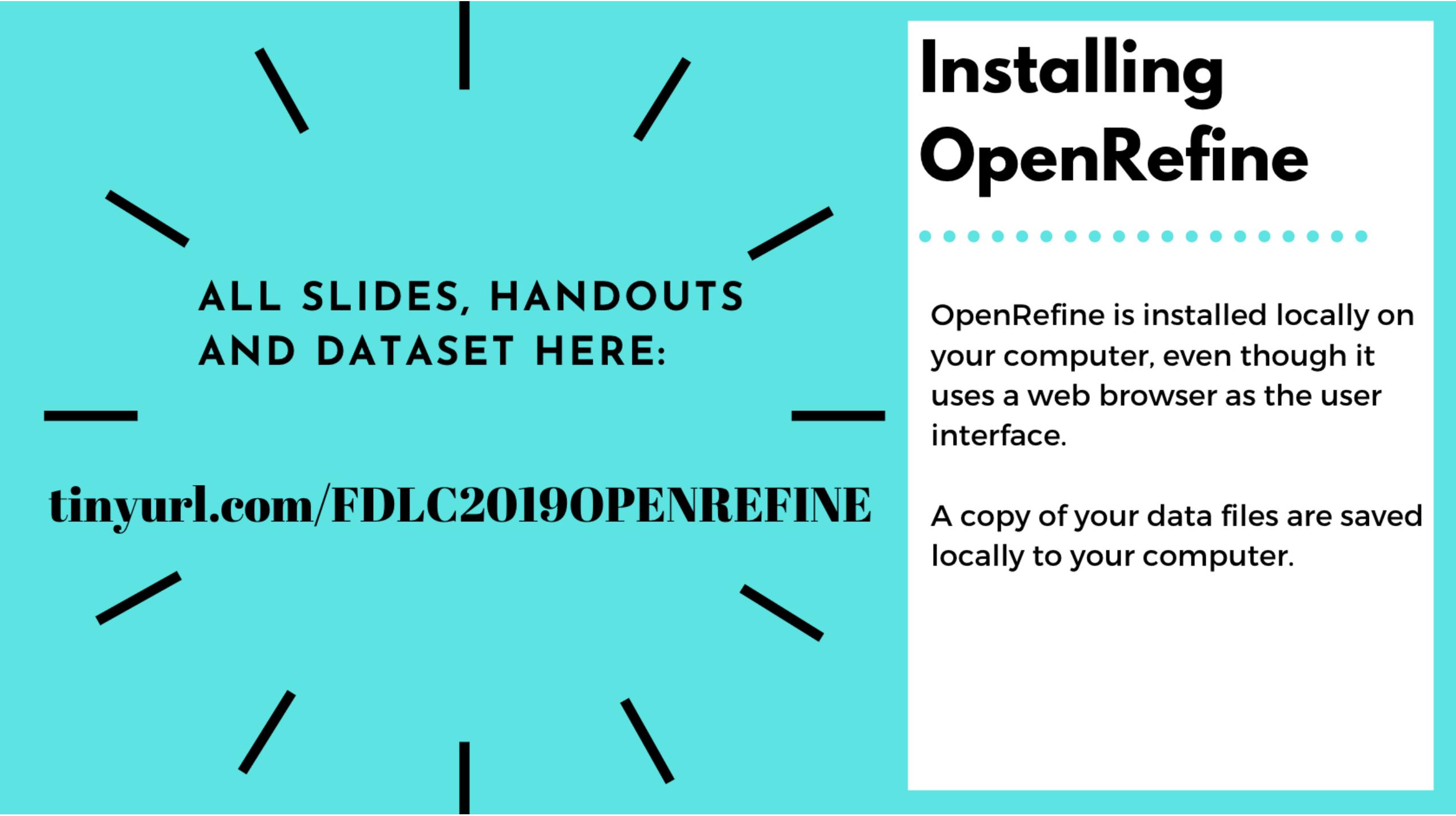
# Learning Objectives

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](https://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](https://tinyurl.com/FDLC2019OPENREFINE)**

# IMPORTING A DATASET INTO OPENREFINE

1



**OpenRefine**

*A power tool for working with messy data*

**New version! [Download OpenRefine v3.2 now.](#)**

Create Project

Open Project

Import Project

Language  
Settings



Version 3.0-beta  
[TRUNK]

**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 documents be added with OpenRefine extensions.

Get data from

**This Computer**

Web Addresses (URLs)

Clipboard

Data Package (JSON URL)

Database

Google Data

2

Locate one or more files on your computer to upload:

Choose Files

Gov\_Docs\_We...roject.xlsx

Next »

3

# IMPORTING A DATASET INTO OPENREFINE

4

Project name  Tags

5

**Create Project »**

# REMOVING A COLUMN

OpenRefine Gov\_Docs\_Weeding\_Project.xlsx [Permalink](#)

Facet / Filter

Undo / Redo 0 / 0

313 rows

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

All	MMS Id	Column	Permanent C	Column2	OCLC Num
☆	1.	991003163989703731	/3:		ocm30989877
☆	2.	991003150309703731			
☆	3.	991003144159703731			
☆	4.	991003147769703731	A 1.		
☆	5.	991003165389703731	A 1.		ocm32473383

## 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](#)

1

2

3

# CLUSTERING

The image shows a software interface with a menu open. The menu is titled 'Facet' and contains several options. Three callouts are present: a '1' pointing to the 'Publication D' dropdown, a '2' pointing to the 'Facet' menu title, and a '3' pointing to the 'Text facet' option.

<input type="checkbox"/> Publication D	<input type="checkbox"/> Resource Type	<input type="checkbox"/> Material Type
<b>Facet</b>		
▶ <b>Text facet</b>		
Text filter		
▶ Edit cells		
▶ Edit column		
▶ Transpose		
Sort...		
▶ View		
▶ Reconcile		
▶ Numeric facet		
▶ Timeline facet		
▶ Scatterplot facet		
▶ Custom text facet...		
▶ Custom Numeric Facet...		
▶ Customized facets ▶		

# CLUSTERING

The screenshot shows a data visualization interface with a blue header bar. The header contains the text "x Publication Date" on the left, "change" on the right, and "61 choices Sort by: name count" in the center. A yellow circle highlights the number "4" in the "count" column. Below the header is a list of publication dates and their corresponding counts, sorted by name. A "Cluster" button is located on the right side of the interface.

Publication Date	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 ...](#)

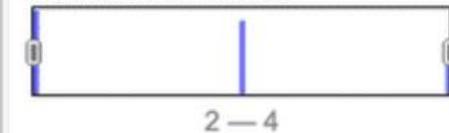
Method

Keying Function

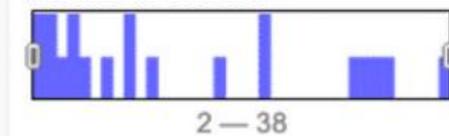
20 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	31	<ul style="list-style-type: none"><li>1987. (19 rows)</li><li>[1987] (10 rows)</li><li>1987] (1 rows)</li><li>[1987- (1 rows)</li></ul>	<input checked="" type="checkbox"/>	<input type="text" value="1987"/>
4	32	<ul style="list-style-type: none"><li>1981. (28 rows)</li><li>[1981?] (2 rows)</li><li>[1981] (1 rows)</li><li>[1981]. (1 rows)</li></ul>	<input type="checkbox"/>	<input type="text" value="1981."/>
4	38	<ul style="list-style-type: none"><li>[1980] (24 rows)</li><li>1980. (12 rows)</li><li>1980- (1 rows)</li><li>1980] (1 rows)</li></ul>	<input type="checkbox"/>	<input type="text" value="[1980]"/>
4	18	<ul style="list-style-type: none"><li>[1991] (11 rows)</li><li>1991] (4 rows)</li><li>1991. (2 rows)</li><li>[1991?] (1 rows)</li></ul>	<input type="checkbox"/>	<input type="text" value="[1991]"/>

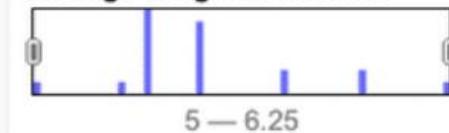
### # Choices in Cluster



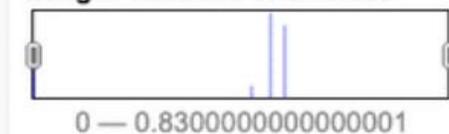
### # Rows in Cluster



### Average Length of Choices



### Length Variance of Choices



Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close

Close

# Sort

1

Publication D	OCLC Number	Resource Type	Material
cm50634760		Book - Physical	Book
cm45917369		Book - Physical	Book
cm45132113		Book - Physical	Book

2

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort
- View
- Reconcile

3

- Sort...
- Reverse
- Remove sort

### Sort by Publication Date

Sort cell values as

- text  case-sensitive
- numbers
- dates
- booleans

4

Position blanks and errors

- Valid values
- Errors
- Blanks

Drag and drop to re-order

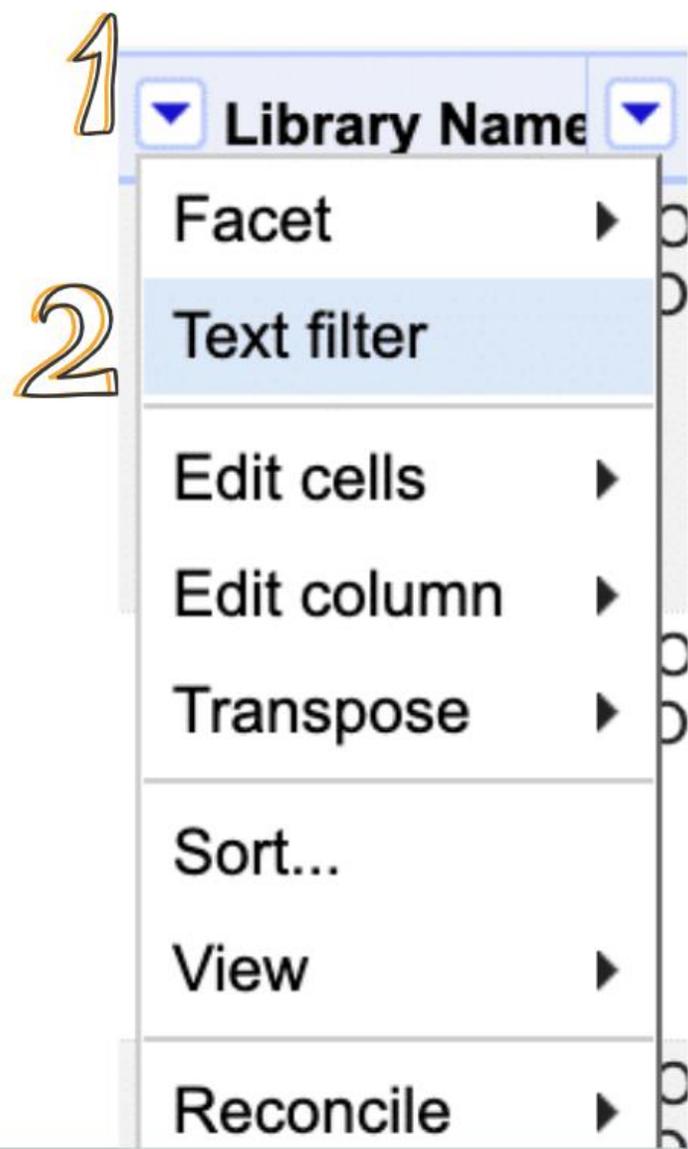
smallest first  largest first

5

6

OK Cancel

# FILTER



3



# FACET

1 Library Name Location Name Column3

2 Facet

3 Text facet

- Text filter
- Numeric facet
- Edit cells
- Timeline facet
- Edit column
- Scatterplot facet
- Transpose
- Custom text facet...
- Sort...
- Custom Numeric Facet...
- View
- Customized facets
- Reconcile

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

**1** 313 rows

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

All	MMS Id	Permanent Call
Transform	13233719703731	A 1.2:ST 8/3
Facet		
Edit rows		
<b>Edit columns</b>		
View	08612489703731	A 1.2:L 75/3

**2** **3** Re-order / remove columns...

**4** 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

**5** OK Cancel

# Closing OpenRefine

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

# Helpful Resources

- 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](mailto:eimmysol@usc.edu)**

**[tinyurl.com/FDLC2019OPENREFINE](https://tinyurl.com/FDLC2019OPENREFINE)**

