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?

<table>
<thead>
<tr>
<th>Customer Name</th>
<th>B: Doe, John</th>
</tr>
</thead>
<tbody>
<tr>
<td>John K. Doe Jr.</td>
<td>Doe, John</td>
</tr>
<tr>
<td>Mr. Doe, John</td>
<td>Doe, John</td>
</tr>
<tr>
<td>Jane A. Smith</td>
<td>Smith, Jane</td>
</tr>
<tr>
<td>MS. Jane Smith</td>
<td>Smith, Jane</td>
</tr>
<tr>
<td>Smith, Jane</td>
<td>Smith, Jane</td>
</tr>
<tr>
<td>Dr Anthony R Von Fange III</td>
<td>Von Fange, Anthony</td>
</tr>
<tr>
<td>Peter Tyson</td>
<td>Tyson, Peter</td>
</tr>
<tr>
<td>Dan E. Williams</td>
<td>Williams, Dan</td>
</tr>
<tr>
<td>James Davis Sr.</td>
<td>Davis, James</td>
</tr>
<tr>
<td>James J. Davis</td>
<td>Davis, James</td>
</tr>
<tr>
<td>Mr. Donald Edward Miller</td>
<td>Miller, Donald</td>
</tr>
<tr>
<td>Miller, Donald</td>
<td>Miller, Donald</td>
</tr>
<tr>
<td>Rajesh Krishnan</td>
<td>Krishnan, Rajesh</td>
</tr>
<tr>
<td>Daniel Chen</td>
<td>Chen, Daniel</td>
</tr>
</tbody>
</table>
What is OpenRefine?

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

OpenRefine vs Excel vs R
Demonstrations & Hands-on Practice

tinyurl.com/FDLC2019OPENREFINE
IMPORTING A DATASET INTO OPENREFINE

1. Create Project
   - Open Project
   - Import Project
   - Language Settings

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.

Get data from

This Computer
- Web Addresses (URLs)
- Clipboard
- Data Package (JSON URL)
- Database
- Google Data

Located one or more files on your computer to upload:

Gov_Docs_We…roject.xlsx

Next »
IMPORTING A DATASET INTO OPREFINE

4. Project name: Gov_Docs_Weeding_Project
5. Create Project »
# REMOVING A COLUMN

**OpenRefine** Gov Docs Weeding Project.xlsx  Permalink

<table>
<thead>
<tr>
<th>Facet / Filter</th>
<th>Undo / Redo</th>
<th>0 / 0</th>
</tr>
</thead>
</table>

**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](#)

### 313 rows

<table>
<thead>
<tr>
<th>MMS Id</th>
<th>1. 991003163989703731</th>
<th>2. 991003150309703731</th>
<th>3. 991003144159703731</th>
<th>4. 991003147769703731</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Column</strong></td>
<td>/3:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Edit column</strong></td>
<td>Split into several columns...</td>
<td>Transpose</td>
<td>Add column based on this column...</td>
<td>Add column by fetching URLs...</td>
</tr>
<tr>
<td><strong>Reconcile</strong></td>
<td>Rename this column</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Remove this column</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Move column to beginning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Move column to end</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Move column left</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Move column right</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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...

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Row Count</th>
<th>Values in Cluster</th>
<th>Merge?</th>
<th>New Cell Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>31</td>
<td>• 1987. (19 rows)</td>
<td></td>
<td>1987</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• [1987] (10 rows)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• [1987] (1 rows)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• [1987-] (1 rows)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|              |           | • [1981] (2 rows) |       |
|              |           | • [1981] (1 rows) |       |
|              |           | • [1981-] (1 rows)|       |

| 4            | 38        | • [1980] (24 rows)|       | [1980]         |
|              |           | • 1980. (12 rows)|       |
|              |           | • 1980- (1 rows) |       |
|              |           | • [1980] (1 rows)|       |

| 4            | 18        | • [1991] (11 rows)|       | [1991]         |
|              |           | • 1991 (4 rows)  |       |
|              |           | • 1991 (2 rows)  |       |
|              |           | • [1991-] (1 rows)|       |

# Choices in Cluster

<table>
<thead>
<tr>
<th># Choices in Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 — 4</td>
</tr>
</tbody>
</table>

# Rows in Cluster

<table>
<thead>
<tr>
<th># Rows in Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 — 38</td>
</tr>
</tbody>
</table>

Average Length of Choices

<table>
<thead>
<tr>
<th>Average Length of Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 — 6.25</td>
</tr>
</tbody>
</table>

Length Variance of Choices

<table>
<thead>
<tr>
<th>Length Variance of Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 — 0.8300000000000001</td>
</tr>
</tbody>
</table>
Sort

1. Publication Date
2. OCLC Number
3. Resource Type
4. Material
5. Sort by Publication Date
6. Sort cell values as
   - text
   - case-sensitive numbers
   - dates
   - booleans
7. Position blanks and errors
   - Valid values
   - Errors
   - Blanks
8. Drag and drop to reorder
   - smallest first
   - largest first
9. OK
10. Cancel
FILTER

1. Library Name
2. Text filter
3. Facet / Filter
   - Undo / Redo
   - Refresh
   - Reset All
   - Remove All

Library Name:
- VKC Library
- case sensitive
- regular expression
FACET

1. Select Text facet from the menu.
2. Choose 3 choices from the library name.
4. Facet by choice counts.
RE-ORDER / REMOVE COLUMNS

1. Show as: rows recordsShow: 5 10 25 50 rows Show:
2. Edit columns
3. Re-order / remove columns...
4. Drag columns to re-order
5. Drop columns here to remove

- Location Name
- Network Number
- MMS Id

Columns:
- 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
Closing OpenRefine

• Click on OpenRefine icon and type Command- Q.

• Wait until there’s a message that says the shutdown is complete.
Helpful Resources

Questions?

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