# Intro to Open Refine

Rebecca Olson, Business and Social Science Informationist, Research & Data Services, University of Cincinnati Libraries. Rebecca.Olson@uc.edu

Open Refine is a free, open source tool used to clean data from spreadsheets. It is a powerful tool to clean messy data. This is an introductory class.

|  |
| --- |
| **Installation**:  <http://openrefine.org/download.html>  Follow directions to download the updated version appropriate to your operating system.  OpenRefine is a Java application, and you need to have a ‘Java Runtime Environment’ (JRE) installed on your computer to run OpenRefine. If you don’t already have one installed then you can download and install from <http://java.com> by going to the site and clicking “Free Java Download”.  Please use **Firefox, Chrome or Safari**. IE or Edge will not work. |

**What Data Can I use**:

TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents

Create Your First OR Project:

 - Look for this image and click on it (exe file) to launch Open Refine in a browser. NOTE –if OR won’t launch, enter <http://127.0.0.1:3333/> into your browser window.

1. Click  - Create Project and then Get data from This Computer.

2. Click on  Choose Files and locate the file Menu.csv Click Next.

3. A Preview will appear. You don’t need to change any of the data below. Enter Project Name and click “Create Project”

|  |
| --- |
| **Note: Your original data (raw data) file will NOT change! The Project is a copy!** |

## Using Open Refine

1. With the project open, click on the arrow at the EVENT column
2. Click “SORT.” What do you see?
3. Try sorting on other columns
4. Click remove Sort.
5. Click on the PLACE column. Choose “Edit Column – Rename this column.” Rename it to **location\_of\_meal** (or other term, just follow naming convention)

## Faceting

1. From this column, click on the arrow – Facet – Text Facet.
2. Facets list groupings of the same text (or number or value). Take a look at “KONIGEN LOUISE” AT SEA. What do you notice?
3. Place your mouse at the end of an entry – click on EDIT and change it to match the other entry.
4. Note the “include” option – click on it and see what happens. Click on “exclude” to go back.
5. Try a Text Facet on another column and try editing it.

## Clustering

1. CLUSTERING. Click on “Cluster” from the Facet Box OR from the column menu, Edit Cells – Cluster and Edit.
2. Instead of editing entries one by one from the facet, you can change/edit entries in groups at a time. Look at Delmonico’s NY. Note the Cluster beneath – Delmonico’s, New York, NY. These are obviously the same – but how do you want the entries to read? Edit both to have the same entry, click the boxes, then click **Merge Selected & Re-cluster** button at the bottom.
3. Repeat this a few times. Try changing the clustering options.
4. Close the cluster.

## Transformations

1. In the Facet box for the **Location\_of\_meal** column, note the first entry, 9 W 53rd St. Click edit – note that there is a space at the beginning of the entry. You could delete that space manually, or you can do it automatically.
2. From the column menu, select Edit Cells – Common Transformations – Trim leading white space
3. Let’s get rid of the brackets. [ ]
4. Edit Cells – Transform.
5. Type in this: **value.replace('[','').replace(']','')**
6. See the resources below for additional GREL help.

## Automation

1. Let’s apply this to the EVENT column.
2. Click on the undo/redo tab.
3. Click on **EXTRACT.** A window with coding pops up. Unselect the first two options, just keeping checked “text transform on cells… using grel”
4. In the coding text on the right, everywhere you see “location\_of\_meal” replace with “event”. We are changing the column name to what the column we want to apply this to. Don’t forget to change the “Description” field, at the bottom, too.
5. Highlight the code and copy it. (ctrl c). Close the window and then click “**apply**”. Paste your copied text in the new window and click “**Perform operations**.” The brackets ([ ]) should have disappeared.

## Data clean up

1. Good spreadsheets have one variable per cell. However, this is not always possible. The notes field for example. What if you want to separate the notes into their own columns?
2. Notes column – click on the arrow menu and select **Edit Column** – **Split into several columns**
   1. Note – we have a semicolon (;) that designates what we want split into separate columns
   2. In the most recent versions of OpenRefine, you can split by field lengths, namely integers. This could be helpful for splitting dates or other number combos.
3. Select “split the column by separator” and enter a semicolon **; .** Leave the number of columns blank. Click **OK**
4. You now have several new columns. You can now rename them if applicable.

## Fixing mistakes

1. **UNDO/REDO**. You can join the split columns together through the edit column option, or you can use the “UNDO/REDO” option.

## Dates

1. Click on the Date column and make a copy. Rename it Date2
2. Edit Cells, Common Transforms, Trim Leading White Space
3. Edit Cells, Common Transforms, Date. (Play around with the Facet timeline)
4. Edit Cells, Transform – cut and paste
5. toString(toDate(value),"MM/dd/yyyy") OR value.slice(5, 7) + '/' + value.slice(8, 10) + '/' + value.slice(0, 4)

## Saving/extracting your data

1. Once finished with your cleanup, you can save the data – by clicking the EXPORT button in the top right corner. You have a variety of formats to save it as. We advise saving it as a CSV form.
2. If you would like to save your project, click on Export your project and save it (either locally or on a google drive). It will be saved as a tar.gz file (a compressed file). You will have to open Openrefine and then load the project to get it to open. This is helpful to share with others to review what you have done or for reproducibility.
3. Remember- your original data has not changed. Save this newly cleaned data with a new name!

## More help:

This lesson was inspired by the Openrefine lessons from the Carpentries.

<https://librarycarpentry.org/lc-open-refine/>

<https://datacarpentry.org/OpenRefine-ecology-lesson/>

General Refine Expression Language (GREL)

<https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language>

GREL Recipes

<https://github.com/OpenRefine/OpenRefine/wiki/Recipes>

University of Illinois Library’s Libguide to OpenRefine

<https://guides.library.illinois.edu/openrefine/prepnormal>

## For more workshops, please see:

## <https://libraries.uc.edu/research-teaching-support/workshops.html>

## For questions, comments, or suggestions, please contact us at:

## [ASKDATA@UC.EDU](mailto:ASKDATA@UC.EDU)