Open Refine Tutorial 1 of 3:
Cleaning Messy Data

In this tutorial on cleaning messy data you'll resolve variations of titles into a single identity using Open Refine. You'll also learn how to undo functions and export your refined spreadsheet to your desktop.

Getting Started

First you need to download a dataset to work with. This tutorial uses open data on public housing from the City of Los Angeles' Housing and Community Development Department, found here [https://data.lacity.org/A-Livable-and-Sustainable-City/HCIDLA-Affordable-Housing-Projects-Catalog-And-Lis/u4mj-cwbz].

On the webpage, click the blue 'export' box and export the file as a 'CSV for Excel' to your computer.

Now download Open Refine's installation package [http://openrefine.org/download.html]. Once installed, click on its application icon, and it will pop open in your default web browser. Create a new project and click 'browse' to upload your file. Click 'Next'. On the following screen, check that the box, "Parse cell text into numbers, dates, etc", is ticked, then click 'create project' on the following screen.

You'll see a familiar looking spreadsheet format. You can also see that you're able to edit each individual cell, and that each of the column headings has an arrow – clicking on the arrow brings up a dropdown menu of operations.

Resolving Variations in a Name

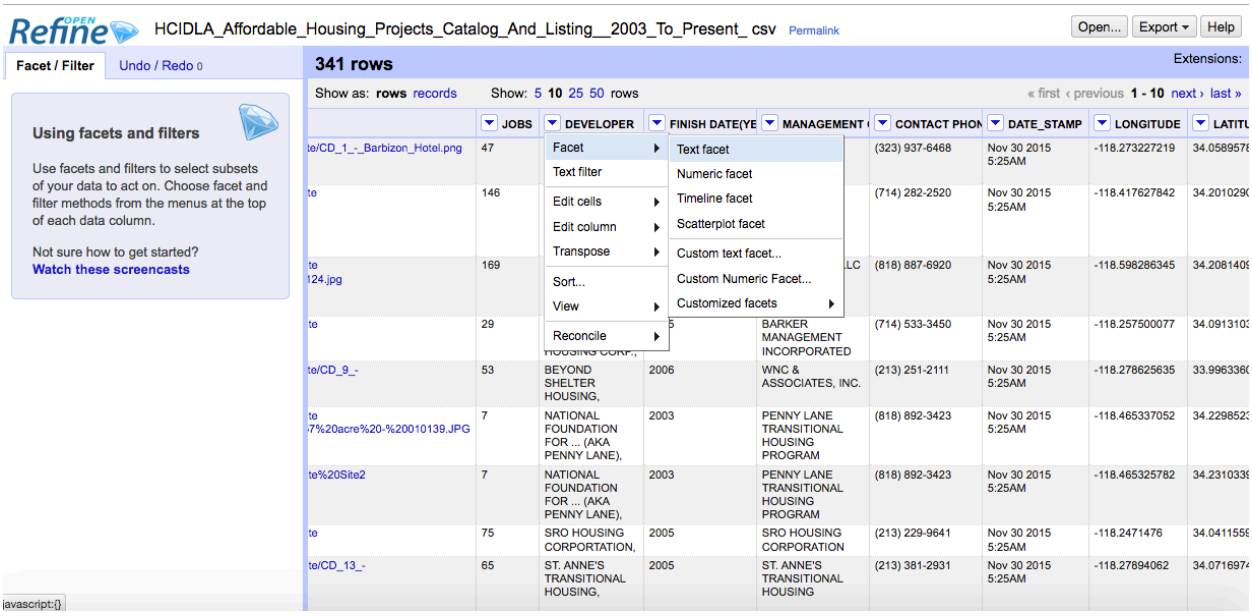
Open Refine helps find and clean variations or misspellings of titles, resolving them into a single term. This function is useful because spreadsheets often contain human error or suffer from poorly designed records keeping, with multiple name variations for the same entity. For instance, the author William Burroughs might be entered as "William Burroughs", "William S. Burroughs", "William Seward Burroughs", or "William Seward Burroughs II."

Furthermore, computers can read two titles as distinct if one is capitalized and the other isn't, or if one entry has a blank space at the end and another doesn't, a difference that isn't even apparent to the human eye. Consequently, when you sort or search for a term, you may miss variants in the retrieved results.

To locate and correct these inconsistencies, you'll learn two features: faceting and clustering.

Faceting

You'll be faceting the 'Developer' column to see if there are any inconsistencies in the titles.



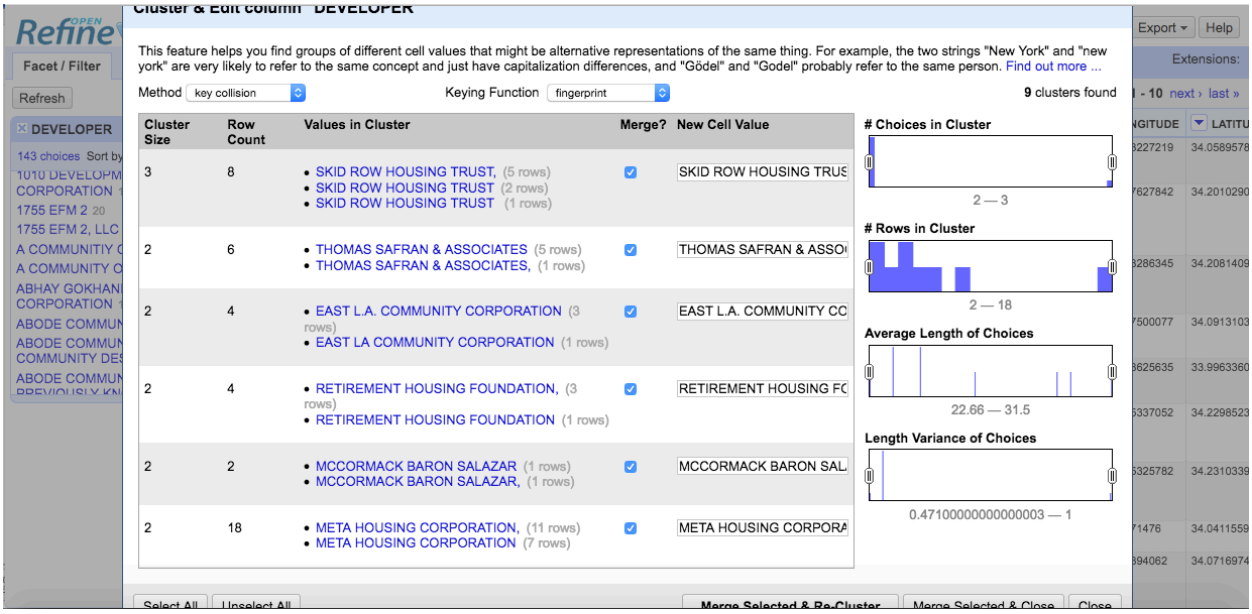
In the dropdown menu of the ‘Developer’ column, you’ll see the ‘Facet’ feature. In the submenu, click on the ‘facet text’ option. Now, in the panel to the left of the spreadsheet, you’ll now see a list of all the unique values within that column – 136 in total. Each unique value is assigned a number, meaning the number of times it appears in the column. By default the faceted list is sorted in alphabetical order. You can now easily compare the values to see if there are any variations of the same name.

Clearly this list has problems with consistent data entry. 1755 EFM is most likely the same entity as 1755 EFM, LLC, and the name Community of Friends has been entered both with a comma at the end and without one.

To resolve these differences, we will next use the *clustering* feature.

Clustering

Now click the ‘Cluster’ button. A pop-up window shows all groups of different cell values that might be alternative representations of the same thing:



The cluster function relies on a few different algorithms to determine how aggressively it will apply criteria to determine sameness. You can switch between these algorithms by selecting within the 'Keying Functions' drop down menu. Fingerprinting will be the strictest heuristic – this formula detects values that differ only in punctuation and capitalization. For instance, 'Atwood, Margaret' and 'MARGARET ATWOOD' would both be read as equivalent with the fingerprinting formula, but 'Margaret E. Atwood' would not.

The metaphone function is looser, so it will return more results; it groups entities together by how they sound. Try selecting among each keying function and go through the clusters, using your best judgement to determine if the algorithm has found similarities or delivered false positives. Click the box under 'Merge?' next to any clusters that need resolving, and adjust the title if necessary.

Some cases are less clear. When you use the Metaphone keying function, for example, you get the result of MCCORMACK BARON SALAZAR, MCCORMACK BARON & ASSOCIATES, and MCCORMACK BARON SALAZAR, INC. Should we assume that these are all the same company?

For clues, you can point your mouse over the row and click on 'browse this cluster'. Clicking this function allows you to isolate these rows and scrutinize them further. You can see that all three of these companies have the same Management company in the 'Management' column, so it's likely that they can be resolved as one entity.

Manual Cleaning

Now examine your facet again. If you have resolved all clusters retrieved by the cluster function, that will have narrowed down your list of unique entities. But the cluster function doesn't catch all similar matches. There still are some entities that should be merged that will need to be edited by hand. For example, "Women Organizing Resources Knowledge & Services" is most likely

the same as “W.O.R.K.S.,”. To resolve these, click on the ‘edit’ function next to one of these names, and rename it with the same title as its match.

Additional functions: Creating a custom facet

You can use facets also to isolate rows of specific items. For instance, perhaps we want to find out all of the developers of public housing in Hollywood. To do this, go to the “Community” column and click ‘Facet’ → ‘Custom facet’. In the text box write:

```
value.contains("HOLLYWOOD")
```

The command is case sensitive, so you’ll need to keep the text in all-caps.

The results generates a boolean value of either ‘true’ or ‘false’ depending on whether the current value in the cell contains the text ‘HOLLYWOOD’ anywhere.

NOTE: A Boolean is a data type with only two possible values: true or false.

Next, you’ll note that the results also return ‘E. HOLLYWOOD’ and ‘N. HOLLYWOOD.’ To refine the results further, go to the drop down menu for the “Community” column and click ‘Facet’ → ‘Text Facet’. Now, in your facet box, you can select ‘HOLLYWOOD’ by clicking on the text ‘include’. The results are now isolated to 14 rows.

Note: This command is actually an example of a programming language called Google Refine Expression Language (GREL). Several advanced features like this are available. If you want to explore more commands visit the Google Refine Expression Language (GREL) reference [https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users#reference] or this tutorial [http://www.meanboyfriend.com/overdue_ideas/wp-content/uploads/2014/11/Introduction-to-OpenRefine-handout-CC-BY.pdf].

Note also: You can get to the same results by using the Text Filter. In the “Community” column click the drop down menu, then ‘Text filter’. Type in ‘hollywood’ – in this instance the text does not need to be case sensitive.

Other Types of Facets

You’ll have seen in the drop down menu that Open Refine allows other types of facets, including Numeric facets, Timeline facets for dates, Custom facets, and Scatterplot facets.

- Numeric and Timeline facets display graphs instead of lists of values.

- Scatterplot facets, which allow you to visualize the relationships between numeric values in columns. For more information, you can visit this tutorial:

http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial#Exploring_the_data_with_scatter_plots

- Custom facets allow a range of commands, including isolating cells in a column that contain a particular word (as shown above), finding duplicates, finding blanks, and faceting by length of characters in a cell.

Note: See the 2nd tutorial in this series for more information on Numeric facets.

Undo/Redo

With the undo and Redo function you can go back to any past action you've made on your dataset and restore the data to its state at that point. In this way you can experiment on your data but still access all prior versions.

'Undo' and 'Redo' are accessed via the lefthand panel. Through the 'extract' and 'apply' functions you can also save the operations you've carried out on one data set and apply them to another data set by copying and pasting them.

Exporting Your Cleaned Data

Click on the 'Export' box on the upper right hand corner. You can export as HTML, Excel, csv (comma separated values), tsv (tab separated values, or a custom export.

Additional Resources

Video tutorials on Open Refine:
http://openrefine.org/

Text tutorials:

"OpenRefine Tutorial" by David François Huynh
http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

"Cleaning Spending Data with OpenRefine," by School of Data
http://schoolofdata.org/handbook/recipes/cleaning-spending-data-open-refine/

"Using Google Refine to Clean Messy Data," by Dan Nguyen for ProPublica
https://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning

"Introduction to OpenRefine," developed by Owen Stephens on behalf of the British Library.
http://www.meanboyfriend.com/overdue_ideas/wp-content/uploads/2014/11/Introduction-to-OpenRefine-handout-CC-BY.pdf

"Cleaning Data with OpenRefine," by Seth van Hooland, Ruben Verborgh and Max De Wildo
http://programminghistorian.org/lessons/cleaning-data-with-openrefine

Credits:
Written by Morgan Currie, September 2016.