OpenRefine Tutorial
Olin/Uris Library, Cornell University
Eliza Bettinger, Digital Humanities Librarian (ecb4)
March 6, 2017

digitalhumanities.library.cornell.edu/openrefine-resources

Set Up & Install

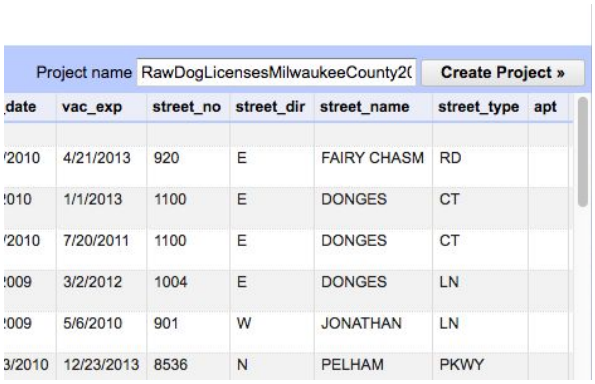
1. Download Workshop Data. Save it on your Desktop.
2. Download OpenRefine for your operating system. Unzip and Install.
3. Double-click on the OpenRefine icon. The software will open in your browser (either Firefox or Chrome). If you ever need to get back to your OpenRefine window, enter this URL in your browser: http://127.0.0.1:3333/

Create a Project

1. Click "Create Project"

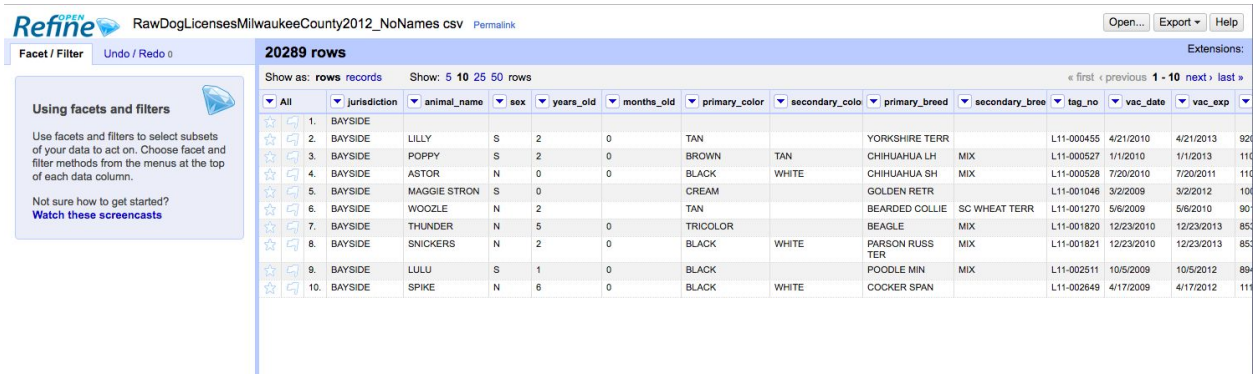
Click "Choose Files" and upload the Workshop Data from the place where you saved it on your computer.

2. Click "Next".
3. Examine the parsing of the data. This screen gives you a preview of the data so that you can make sure that OpenRefine has parsed it correctly. Are column headers in the right place? Are the columns separated correctly? (The workshop data should produce correctly parsed data at the first try.)
4. If everything looks right, click "Create Project >>"



date	vac_exp	street_no	street_dir	street_name	street_type	apt
'2010	4/21/2013	920	E	FAIRY CHASM	RD	
:010	1/1/2013	1100	E	DONGES	CT	
'2010	7/20/2011	1100	E	DONGES	CT	
:009	3/2/2012	1004	E	DONGES	LN	
:009	5/6/2010	901	W	JONATHAN	LN	
3/2010	12/23/2013	8536	N	PELHAM	PKWY	

5. After a few seconds, your project will be created, and you'll see something like this:

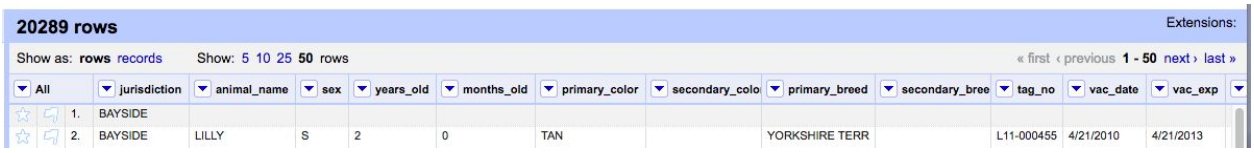


Explore the Data

This dataset consists of a list of more than 20,000 dogs licensed in Milwaukee County, Wisconsin, as of 2012. Take a few moments to get a sense of the dataset, and what cleaning a researcher might want to perform on it.

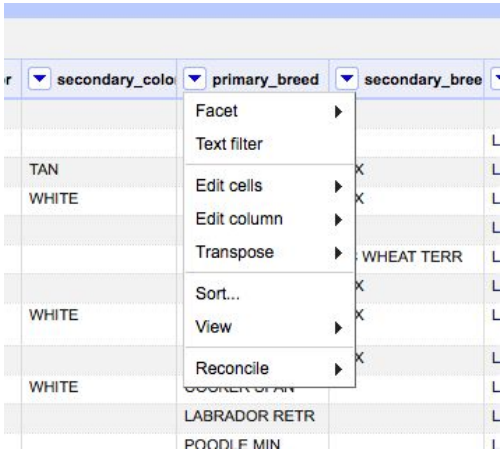
1. Examine the size of shape of the dataset.

You can see the total number of items in the dataset at the "20289 rows" report at the top left.
To see more of the rows at once, click the "50" above the data grid.
Scroll to the left to see all the column headings and type of attributes included in the dataset.
To see items beyond the 50th, click "next >" or "last >>" on the upper right.



2. Apply Filters.

Filters let you get a quick look at different sets of your data.
Click the arrow next to the "primary_breed" column heading. Click "Text filter".



A new window opens up on the left sidebar. Type in the name of a breed you're interested in, such as "Labrador".
Notice that the data visible in the main window changes as you type.
When you're finished typing, notice that you can easily refer to the number of rows that match your filter, on the upper left. In the case of "Labrador," there are 3774 matching rows.

Now add a second filter, such as color. You'll quickly see, for example, that there are 1554 matching rows for primarily black Labrador Retrievers.

Continue to explore the data -- add more filters or change the ones you have.

When you're ready to move on, remove all filters by clicking the "X" on the filter box in the left sidebar.

4. Ask questions.
Let's say you are interested in the geography of dog ownership in Milwaukee County. You want to investigate questions such as: Which breeds are most common in which cities and which zip codes? In which areas are dogs more likely to be spayed or neutered, rather than intact? Where are there concentrations of animals whose vaccinations are soon to be exprie?
What kind of cleaning tasks might you want to perform?

Facets

Facets allow you to get more up close and personal with your data than filters do.

1. Click the arrow next to "primary_breed", then "Facet," then "Text Facet."
A new window will appear in the left sidebar. Notice that the facet tells you that there are 437 different breeds listed in the dataset, and they are displayed in alphabetical order by default.

2. Click and pull on the bottom of the facet window to expand its length and see more of the list.
Note the small gray numerals after each breed name. These indicate the number of entries for breed.

3. At the top of the facet window, click "count" to sort the list by number of members rather than alphabetically.
Very quickly, you'll notice that Labrador Retrievers appear to be the most popular breed by far, followed by German Shepherds and Golden Retrievers.

4. Create filters from your facet.
Click on the name of one of the breeds, such as "PIT BULL"
Notice that immediately, the list of dogs in the main window is limited to pit bulls.
Hover over the name of another breed, such as "BORDER COLLIE". Click "include." Now the data list consists of only pit bulls and border collies.
At the top of the facet window, click "invert". Now the data list consists of every breed EXCEPT pit bulls and border collies.

5. A new facet
Clear all your filters in the Primary_breed facet. (Click "reset" in the upper-right corner of the facet window.)
Now add a new text facet, based on the column "animal_name."

You now have a new facet window, but there are 5547 different animal names represented, so they can't all be listed.

Click “Facet by choice counts.” A third facet window opens, featuring a histogram. Play with this histogram and see if you configure out what it represents. Be sure to watch the data change in the data window AND ALSO the content change in the first “animal_name” facet window that you created.
What’s the most popular dog name in Milwaukee County?

6. Combining Facets
Clear all filters. (Click “reset” in the upper-right corner of each of the facet windows.)
In the first primary_breed facet window, click on the name of a breed, such as “SIBERIAN HUSKY”.
What happens in the “animal_name” facet window? What’s the most common name for Siberian Huskies in Milwaukee County?

7. Batch editing with Facets
Create a facet from the column “sex.”
S and N stand for “spayed” and “neutered”, while F and M stand for “female” and “male.”
Let’s say you want all the S and N cells to be replaced by “O” for operation and the all the F and M cells to be replaced by “I” for intact.

In the facet window, click on “F” so it is highlighted.
Then hover over the highlight. Click on “edit.”
In the box that appears, type “I” (or whatever you’d like). Click “Apply”.
Continue for the other items in the facet window.
Notice how your counts change.

Some Questions:
What’s the most common name for pit bulls?
What’s the most common name for dogs in the jurisdictions of Cudahy and South Milwaukee combined?
Is there a difference in spay/neuter rates between the city of Milwaukee and it suburbs, as a whole? (The suburbs are every jurisdiction except Milwaukee itself.)

Undo/Redo
Experiment with confidence! No change is permanent in OpenRefine.

Try it out:

1. In the left-hand panel, click Undo/Redo

You'll see a list of all the changes you've made to the dataset.

2. Click on one of the steps. Your data will be restored back to its state right before you made the change that you clicked on.

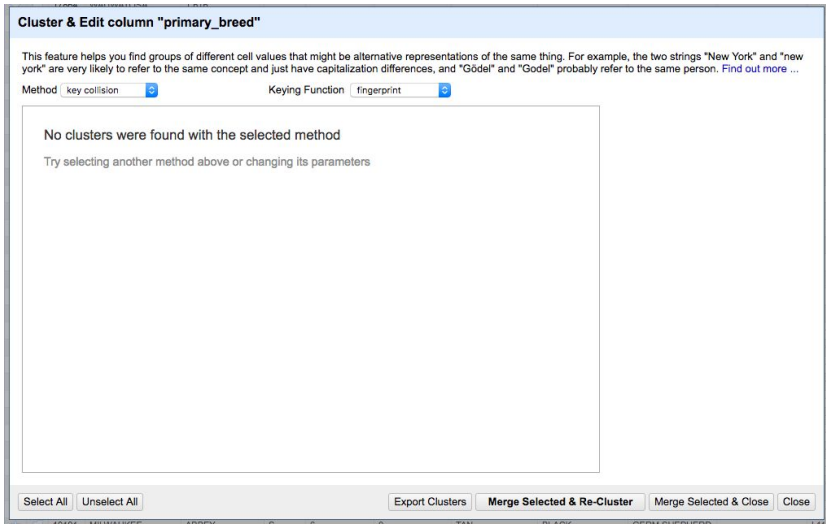
3. To redo the steps that you just undid, scroll back down and click on your desired step.

Clusters

Inconsistencies in spelling, formatting, and capitalization are the bane of data workers. They're also incredibly common. Clustering lets you more easily identify and correct these errors and inconsistencies.

1. Close your "animal_name" facet windows and reset your "primary_breed" window.

2. In the upper right of the "primary_breed" facet window, click the Cluster button.

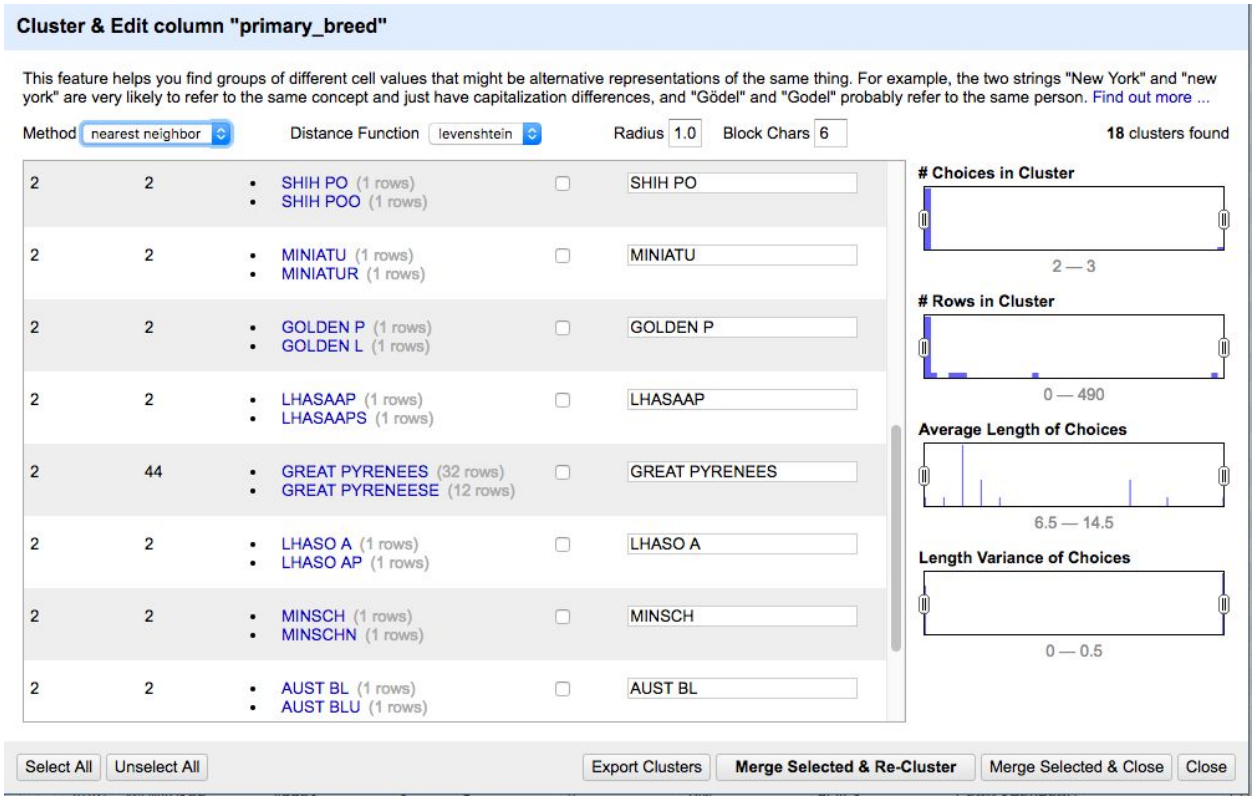


3. Upon opening the window, no clusters are found. Try changing the keying function and then the method.

Soon, you'll see clusters of text that have different spellings, but which may refer to the same breed.

For example, there are 32 rows that refer to “GREAT PYRENEES” and 12 rows that refer to “GREAT PYRENEESE”. Probably, these dog owners all have the same breed: A Great Pyrenees, but some owners spelled it differently on the registration form.

(For more information on clustering methods, see the link on the website.)



To correct all the entries featuring Great Pyrenees, check the “Merge” box, and fill in the box on the right with the correct spelling.
On the bottom of the window, click “Merge Selected & Re-Cluster.” Now all the Great Pyrenees are spelled consistently and will be counted correctly.

Go ahead and merge some more of the clusters.

You’ll notice that some suggested clusters are not correct -- you probably don’t want to merge “Rat Terrier” and “Australian Terrier” because they’re two different breeds. Other clusters will be up to interpretation. Should Long Hair (LH) Dachshunds be counted separately from Short Hair (SH) Dachshunds? It probably depends on your analysis.

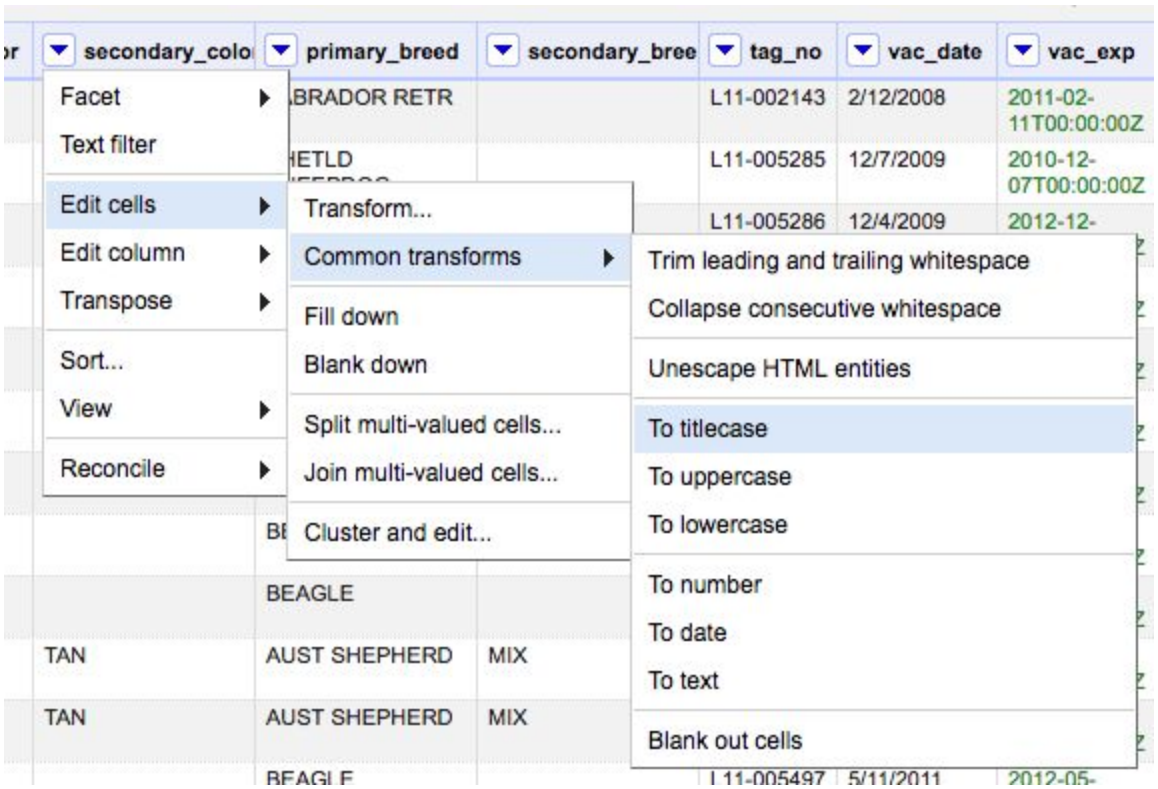
Close your Cluster & Edit window. Hit the Refresh button. Notice that the number of choices (originally 437) in your “primary_breed” facet window has been reduced by however many clusters you chose to merge.

Transformations

When your data needs batch editing, Transformations are a very useful tool.

1. Try a pre-set (common) transformation.

On the arrow next to “secondary_color”, click Edit cells>Common transformations>To titlecase



Notice the list of common transformations. Banal formating errors like leading and trailing whitespaces can cause big problems in data analysis. But this menu makes quick work of them.

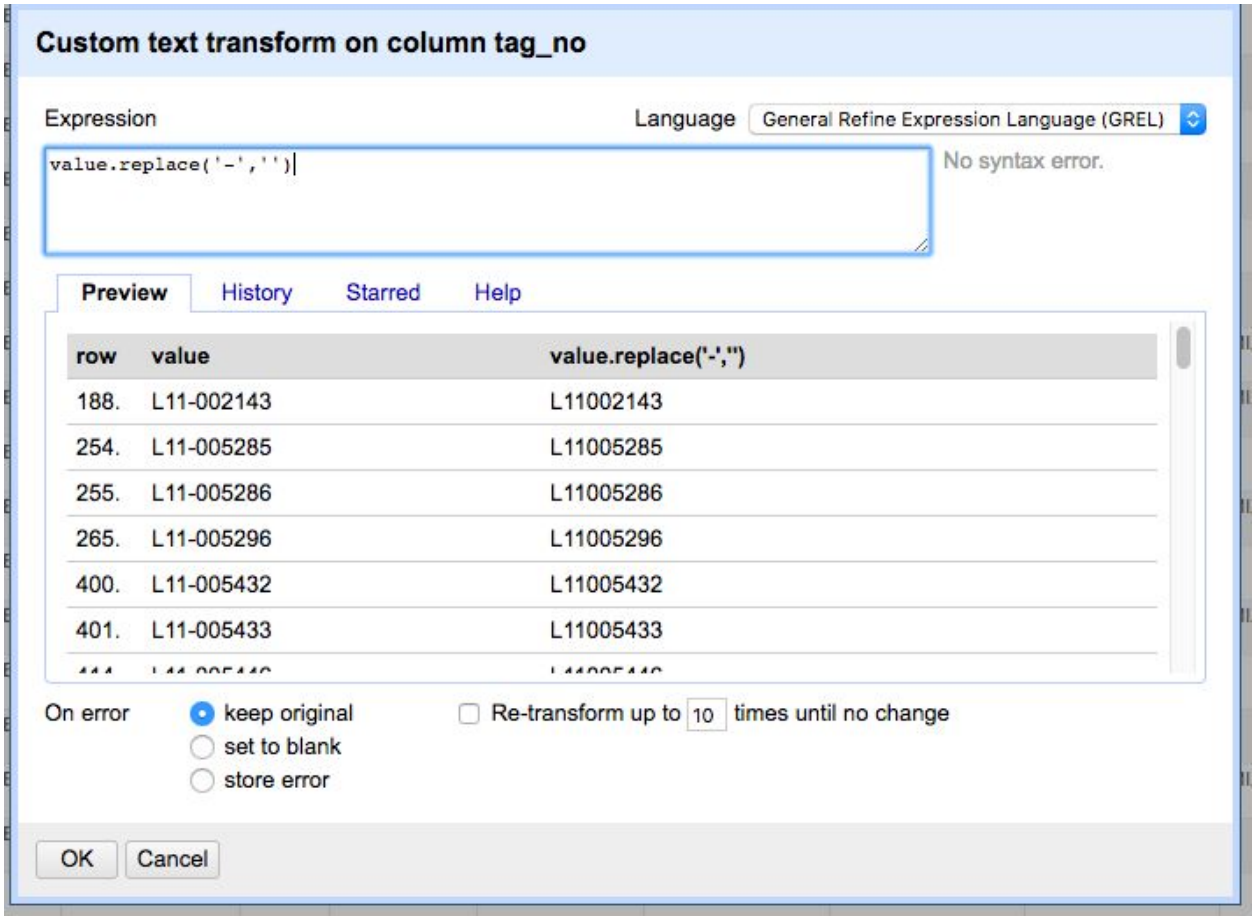
In this case, we don’t need to delete whitespace, so we’ll just change the capitalization of this column to see how it works.

2. Write your own transformation with GREL.

Sometimes you want to make the same change to all the data in a column, but it’s some change that’s specific to your data. That’s where the scripting language GREL comes in. For a good reference on GREL, check the link on the website.

Click the arrow next to the column header “tag_no”. Click Edit cells>Transform.

Enter a GREL expression in that will delete the hyphens from the tag numbers.



Export Your Data

Once you have a dataset you want to work with in your analysis software, it's easy to export.

1. Click the "Export" button in the upper right of the screen.
Choose whether you'd like your exported data to be in CSV, TSV, XLS, XLSX, or ODF format.