Splitting addresses with OpenRefine
By MaryJo Webster
@MaryJoWebster
Mjwebster71@gmail.com
Created: March 2015

Data file: stpaul_crime_Q4_2014.xlsx

We're going to split apart the addresses in this crime data file using OpenRefine. These addresses are particularly nasty because some are intersections and the police department lumps multi-word names together (i.e. “Old Hudson” is listed as “OldHudson”) and they list avenue as “AV” and lane as “LA” and a bunch of other weird stuff. So, to make it easier to clean up some of these problems, we need to split the various pieces into separate fields.

Step 1: Import the data into OpenRefine. This workbook has multiple sheets, so on the first page of the import, make sure you are selecting the “incidents” worksheet.

Step 2: Make a copy of the address field, call it “NewAddress” (we’re going to preserve the original, as is). To add a new column: go to the address field and choose “edit column” >>> “add column based on this column”.

Step 3: Some basic cleanup. Go to the NewAddress field and go to Edit Cells >>>Common Transformations

---first run “trim leading and trailing whitespaces”
--then run “collapse consecutive whitespace”

Step 4: Now we’re going to split out the intersection-based addresses into their own fields. In this case, all the intersection records have an ampersand (&) between the names of the first street and the second.

But first we need to filter our dataset so that we’re only working on the records with the ampersand.

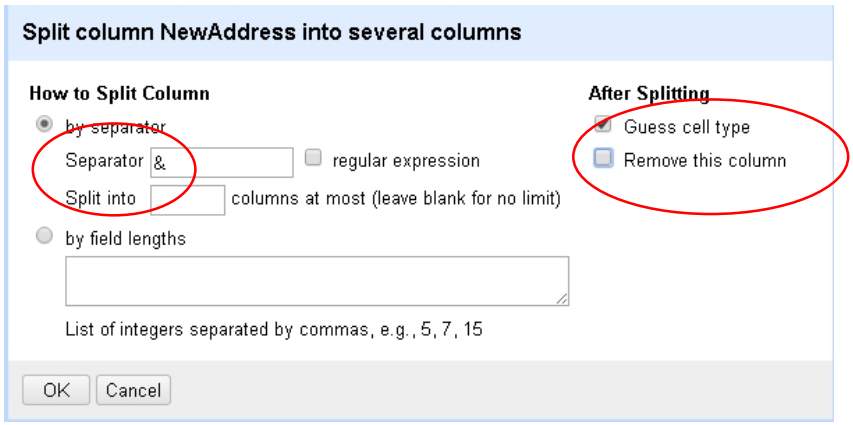
On the “newaddress” field go to “Text filter”

Then in the little box that comes up, type an ampersand (&) and you’ll see that your data filters down to 685 rows.

Go to New Address and "Edit Column" and then choose "Split into several columns."

Put an ampersand (&) in the "separator" box and uncheck where it says "remove this column"

Hit OK



Step 5: You will now have two new columns called "newAddress1" and "newaddress2". Go to the first one and choose "Edit column" >>>>"Rename this column" and call it "Intersect1". Do the same for the second, but call it "Intersect 2"

You'll see that it should have only populated these fields for the addresses that had an ampersand.

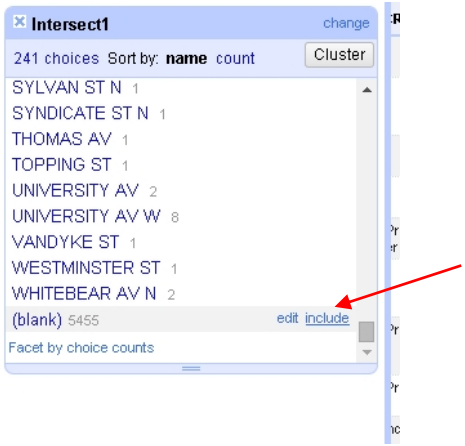
Step 6: Now let's deal with the other addresses. First we need to isolate them from the intersections, so that we only do the remaining steps on the non-intersection addresses. The easiest way to do that is to Facet and then choose which records to "include" (alternatively, you can "exclude" certain records but that doesn't quite work for this example)

Think about the structure of your data and if there is a field that easily indicates which are intersection addresses and which are regular addresses. Figure it out yet?

Here's the answer: Our new intersect1 and intersect2 fields are filled out for intersection records and are blank for all regular addresses.

So go to intersect1 and choose Facet >>>>Text Facet

In the facet box that comes up, scroll to the bottom and find the blanks. Hover in the area to the right of that and get the "include" hyperlink to come up. Click on "include" and your record set will be filtered to only the ones that are blank in the intersect1 field.



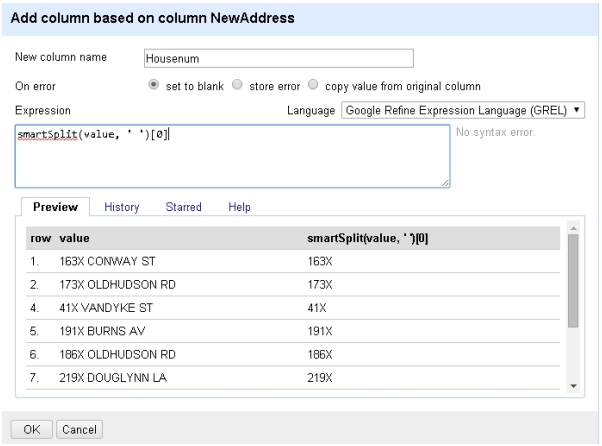
Now we can split apart these records.

There are a couple ways to split apart a field, such as an address. One is using the smartSplit function, but that requires you to do separate formulas for each "chunk" that you want to pull out to separate fields. The other is using a regular expression to tell OpenRefine to split the field at each space.

It's possible to do that with this particular dataset because of one of the problems I pointed out above. The police turned any two-word names into 1 word (i.e. "St Peter" is "StPeter"). So when we split it on the spaces, we should get the house number, street name, street type and direction.

(option1) Here's how to use smartSplit:

Go back to "newaddress" and choose Edit Column >>>Add column based on this column.



Put this formula in the expression box:

smartSplit(value, ' ')[0]

The "value" refers to whatever column you are on and will always be "value".

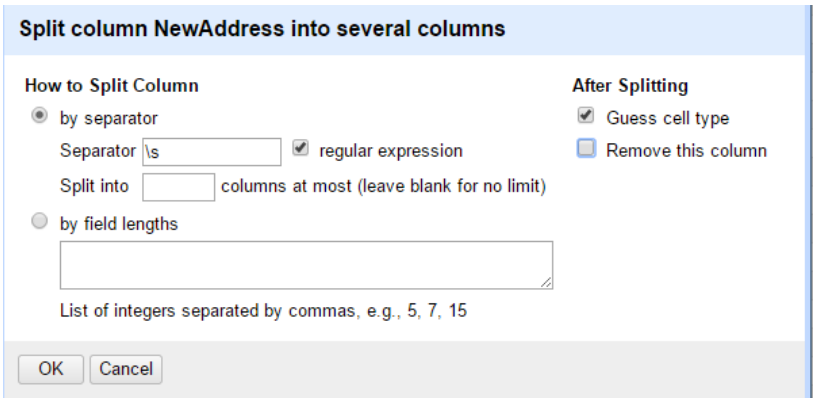
The ' ' tells the function that we're looking for a space (i.e. if you want to split on commas, you would put a comma between the single quotes)

The [0] says we want to grab the first item. We would use [1] for the second item, [2] for the third item, etc.

After using the above syntax for a new field called "Housenum", the repeat the process for a new field called "StreetName", another for "StreetType" and another for "Direction". Remember to adjust the number within the brackets accordingly. (I found that Refine has trouble with the Direction one – i.e. "N", "SE", etc) because it's not used on every record)

(option 2) Here's how you'd split it using a regular expression. (Regular expressions are very powerful tools that can be used in OpenRefine, text editors and other software)

Go to "newaddress" and choose Edit Column >>>Split into several columns.



In the "Separator" box put

\s

Click the box that says "regular expression"

And uncheck the box that says "remove this column"

You'll see that it makes 4 new columns.

Step 7:

To return back to the full dataset, close the facet box (using the "x" in the upper left corner) and you're data will return to the full set.

Now let's do a little standardizing and also check our work.

Go to the Housenum field and choose Facet >>>>Text Facet

Scroll down through the variations and you'll find "WhiteBear" – clearly a problem. Click on WhiteBear and it will filter your dataset to just those two problematic records.

Here we can see that this is supposed to be an intersection, but for some reason both the intersection fields and the housenum, streetname, etc fields got populated. Go to those two records in the dataset and you can manually enter the problem fields (delete "&" from direction, "WhiteBear" from Housenum, etc)

Go back to the facet box and push the "reset" link in the upper right corner of the box.

Now let's look at the ones that have "X" or "XX"

These appear to be ones where the housenumber was completely withheld. You might want to standardize these so that you either have a single X or a double XX for these records where no house number is listed (not even a partial).

Close the Housenum facet and do a text facet on the StreetName field. First thing you'll see is that numbered streets (1st street, 10th street, etc) are listed only with the number. For geocoding or displaying online you might want to add the "st" or "th" or "nd", as

necessary. You can do that by going in the facet box and hovering over the number you want to edit and click “Edit” and in the little box that comes up, change it as you wish (it will change all records with that value)

Finally, you may want to edit the street names that were mashed together. For example, “BargeChannel” should be “Barge Channel.” It will require a familiarity with the street names in St. Paul or access to a mapping program to make sure you are making the correct changes. To change these, you would use the same editing process as used above for the numeric streets.

You can export your results, using the export button in the upper right corner.

Extra:

Need to geocode your addresses? Check out these directions on how to use an API to geocode them within OpenRefine. Note, there are some restrictions if you want to use Google Maps.

https://github.com/OpenRefine/OpenRefine/wiki/Geocoding

More training materials by MaryJo Webster:

Mjwebster.github.io/DataJ