Poynter.

How To's Digital Strategies

How journalists can use Google Refine to clean 'dirty' data sets

by **Matt Wynn**
Published Dec. 21, 2011 6:28 am
Updated Dec. 21, 2011 9:06 am

The first attempt at a lead for this post, it turns out, was pretty much the same lead I wrote five years ago when reviewing a book about dirty data.

My lapse illustrates two things: First, that I have the memory of a goldfish and some bad habits to address. Second, that dirty data is a constant thorn in the sides of data journalists.

Luckily, we now have a tool to address it.

Google Refine bills itself as a "power tool for working with messy data," and it does not disappoint. While not a turnkey solve-all for data integrity, it makes this tedious task far less intimidating. In this tutorial, we'll cover how to install and a take advantage of one trick that will make your work easier.

Understanding the problem

Before diving into what the tool can do, let's take a minute to understand the problem it solves.

Calling data "dirty" means that it's unreliable for analysis. Consider the origin of most government data sets, specifically, and it's easy to understand how that happens. Most government data starts when a politician mandates that an agency track a certain issue. Full stop. End of story. Rarely is there a mandate to analyze the data; even more rare is the allocation of funding for its creation. The result is data created by minimum-wage temp employees with little guidance, which leads to all kinds of inconsistencies.

Names are a classic example of the variations that can happen. A name can be saved in a database in a number of ways. The single name "Jon Doe" could also be stored as "Jonathan Doe," "J Doe," "J. DOE," J.P. Doe," and "Jon Paul

Doe," and that doesn't even count misspellings. Computers read each permutation as its own entity. If it was a database of, say, campaign donations to Mr. Doe, our calculations would miss the mark by a wide margin.

So clean we must.

Refine to the rescue

Installing Refine will vary slightly depending on your operating system. You can find instructions here. Once you've installed it, the program itself runs out of your Web browser.

We're going to use a table we received from the Omaha police department to see one of Refine's tricks. Data is easily imported from CSV files or directly from Excel. Once it's loaded into the program, I'm taken to a screen like this:

We're going to focus on the field called "description," which in this case 911 responders use to type a free-text interpretation to describe the call.

Right away it becomes clear that we have data integrity issues. Even in this small sample, you can see that third-degree sexual assault shows up four times, under four different names — just one of many, many clear duplications.

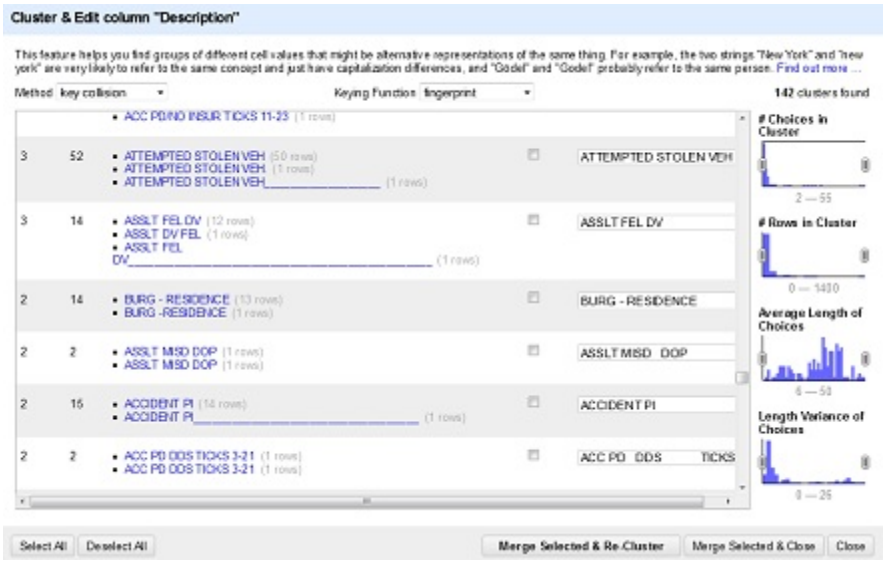
Clicking the arrows next to each field name shows that Refine does many of the same basic tricks as a spreadsheet program. You can sort and add columns, which is important when cleaning data because you always want to be able to explain your logic later. To get started, I'll choose the handle on the description field and pick Facet, Text Facet.

That's going to create a new facet box over on the left-hand side of the screen.

The window lists all the various offenses that have been entered over time. It also gives me a summary of what I have (my records contain 1,538 different

values in the description field). I know there should be only about 200 types of crimes in the data, so this is no good.

The window includes a button labeled “Cluster,” which is where the whizbang lives. Whereas the facet box lists exact matches — treating each of our third-degree sex assaults as separate labels — clustering shows matches based on a variety of approaches that you can fine tune. Clicking the button opens the following:



Rather than rely on precise matches, the Cluster and Edit window uses more sophisticated matching algorithms to group like values. Spacing issues, nonsense characters and misspellings have all been accounted for, at least to some degree.

These results are a starting point; not the final outcome. At the top of the Cluster & Edit window you’ll see a few choices for controlling the clustering process. The defaults are the key collision method and fingerprint function, which give me 142 options for new clusters. The other options use different rules: matching based on the letters in a cell, the sound of the words, the number of steps needed to make string match and more. There’s even a technique designed specifically to solve the name problem described earlier.

To use Refine, you’ll want to be familiar with the ins and outs of each approach. Google provides a walk-through of the various options. Each project will call for a different technique or combination, and you can mix and match to your heart’s content.

Each method can yield dramatically different results. For example, choosing the nearest neighbor method with the PPM function gave me 1,077 clusters. It lumped together descriptions that are less cut-and-dry than using the previous

method. If I accepted the defaults, this approach would combine DUIs and other traffic offenses into one catch-all traffic field — no good for my purposes.

For that reason, it's important to go through each clustering option one by one and make sure it looks good before accepting Refine's suggestion. To do this, you'd read through the "Values in Cluster" field to make sure the options listed merit clustering. If so, you click the "Merge" checkbox and, if needed, apply a new value to those records.

Since this is just a demo, I'm going to go ahead and accept all the defaults. Right off the bat, my list of values has fallen off a cliff, from 1,538 values to 893. I can keep going back through the process, using other techniques to get to the results that most accurately represent what my data really want to tell me.

And if I make a mistake? No sweat. Refine has an "Undo/Redo" option that persists even after you close out. Any step you take can be rolled back with the click of a button.

Next steps

The clustering function alone is a huge boon to data wonks. But it's only one of many things Refine can do. There are several walkthroughs out there to help get you started:

- Refine provides screencasts and sample data sets.
- Paul Bradshaw and the Online Journalism blog outline different transformations that can be used specifically for cleaning data.
- Refine developer David Huynh created an excellent tutorial specifically designed for journalists. (Sample data available here.)

Dirty data is here to stay. But with Refine, at least it's no longer the daunting task it once was.

HACKS HACKERS

This piece is part of a Poynter Hacks/Hackers series featuring How To's that focus on what journalists can learn from emerging trends in technology and new tech tools.

Tags: Hacks/Hackers

RELATED POSTS

How journalists can use Backbone to create data-driven projects

10 tools that can help data journalists do better work, be more efficient

Can Journalists Entrust their Data to Google?

RELATED TRAINING

Leading an Online Newsroom: What You Need to Know

Writing Better Headlines

Online Media Law: The Basics for Bloggers and Other Publishers

How journalists can use Excel to organize data for stories

Google's +1 button adds sharing ability

What Makes a Great Blog: Principles for Success

10 Things You Can Do For Free on Your Web Site