Introduction to Digital Humanities

Get started with OpenRefine

DHers spend a TON of time cleaning and manipulating data. Luckily, there's a tool that makes all of this easier. It's called OpenRefine, and it's free!

This tutorial will walk you through some of the most common data-manipulation tasks you'll need to perform. When you're done, you should know how to:

- clean up spelling inconsistencies
- remove leading and trailing whitespace
- split cells into multiple columns

If you're using a computer that already has OpenRefine installed on it, you can skip Step One.

Before you get started, download this file somewhere onto your computer. It's a sample data file called NJShipwrecks.csv.

About This Class

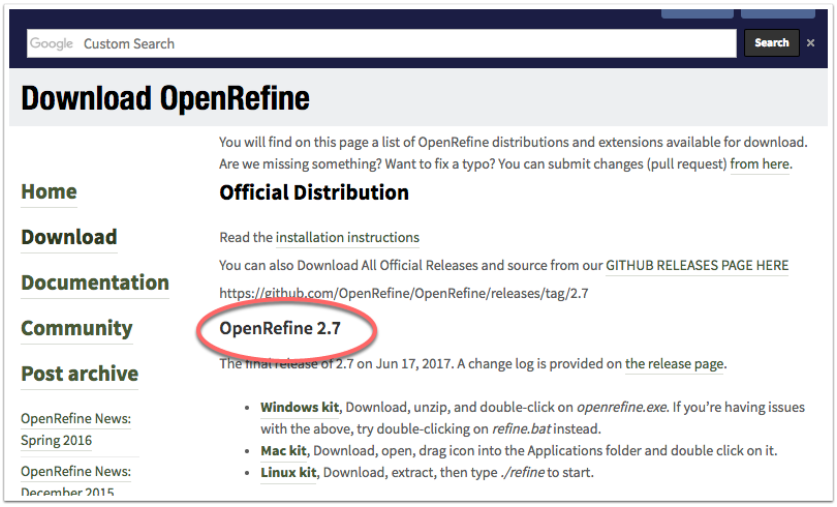
UCLA, Fall 2017
Professor Miriam Posner
TAs: Francesca Albrezzi and Dustin O'Hara
Lectures: M, W, 2-3:15, Young 2200
Labs: Fridays, Rolfe 2118 and YRL 11630F
Contact and office hours

Need help?

1. Install OpenRefine

Head to www.openrefine.org/download and download OpenRefine 2.7 as you would any software. It's available for both Windows and Mac.

NOTE: If you're on a Mac and, when you try to open OpenRefine, you get a message saying that you can't open software from an unidentified developer, do the following: Go to **System Preferences**, then **Security and Privacy**. On the **General** tab, click the lock to make changes, and then click on **Open Anyway.** You should now be able open the software.



2. Open OpenRefine

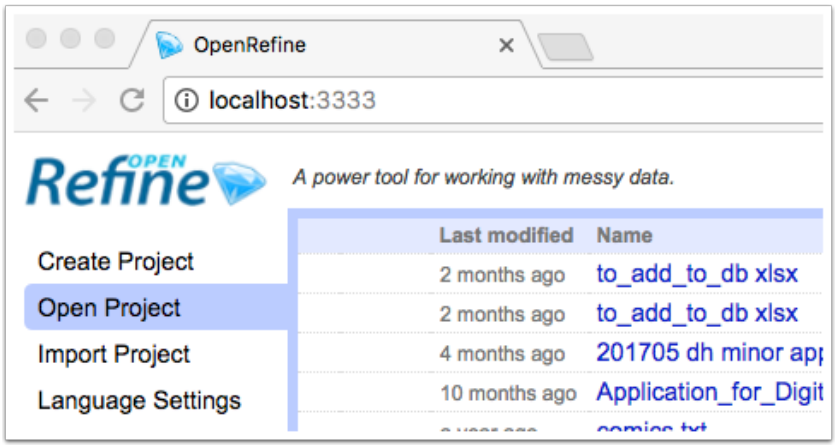
Double-click on the OpenRefine icon. It should open in your web browser. Occasionally, for whatever reason, OpenRefine doesn't launch when you double-click it. If this happens to you, enter **localhost:3333** in your browser's address bar and press return.

Quick Links

Reading schedule
Log in to blog
Lecture Podcasts
CCLE Site
Project Milestones

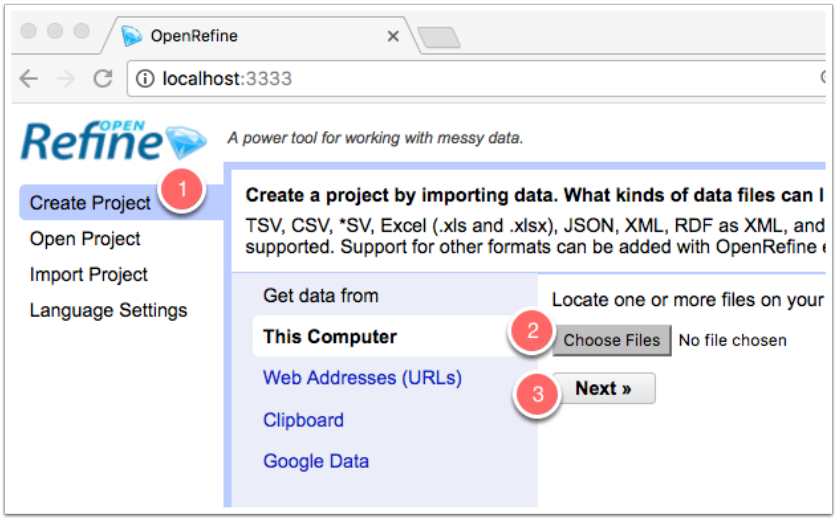
Recent Posts

- Blog Post 1: Walt Whitman Archive Makeup
- Blog Post 2: Heavy Metal Material MAKEUP
- Blog Post: Network Map
- Week 6: HTML
- Week 3 Blog Post: Listing of



3. Open your data file

Click on **Create Project** and then **Choose Files**. Navigate to the NJShipwrecks.csv file and then click **Next**.



4. What the heck is this?

This is just a preview of the way your data will look when you're working with it in OpenRefine. You

Active Businesses

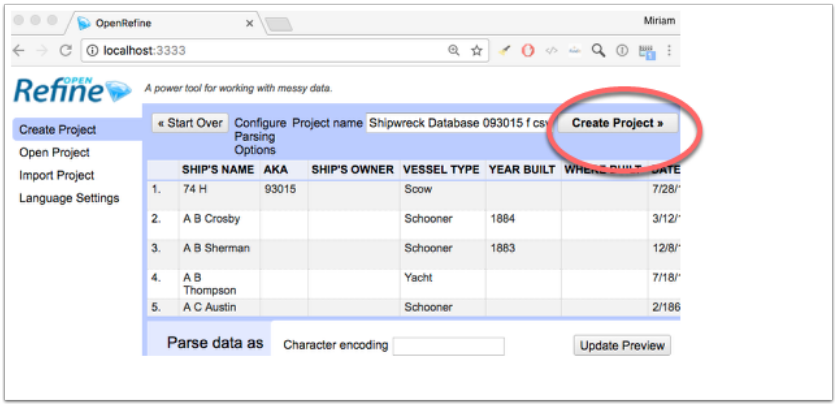
Meta

- Log in
- Entries RSS
- Comments RSS
- WordPress.org

License

Teachers: Please feel free to reuse any part of this syllabus you like!

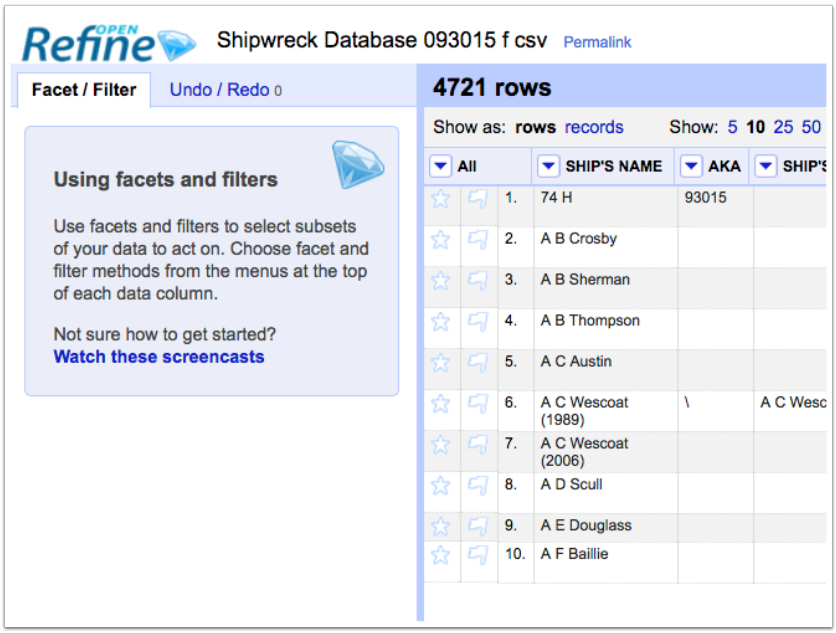
shouldn't have to make any changes; just click on **Create Project**.



5. What the heck is this (part 2)?

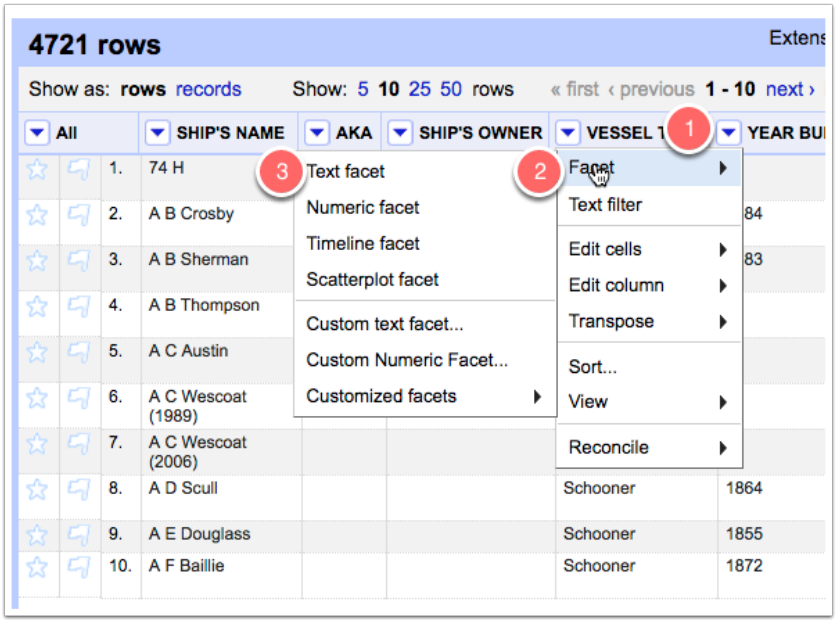
This is the main interface you'll use to work with your data. It sort of looks like Excel, but notice it shows you only 10 records at a time. That's because you're not supposed to be working with your data record by record; you'll find ways to group it into batches and then work with it. We'll try that next.

love it if you'd let me know! I'm trying to collect examples!



6. Create a facet

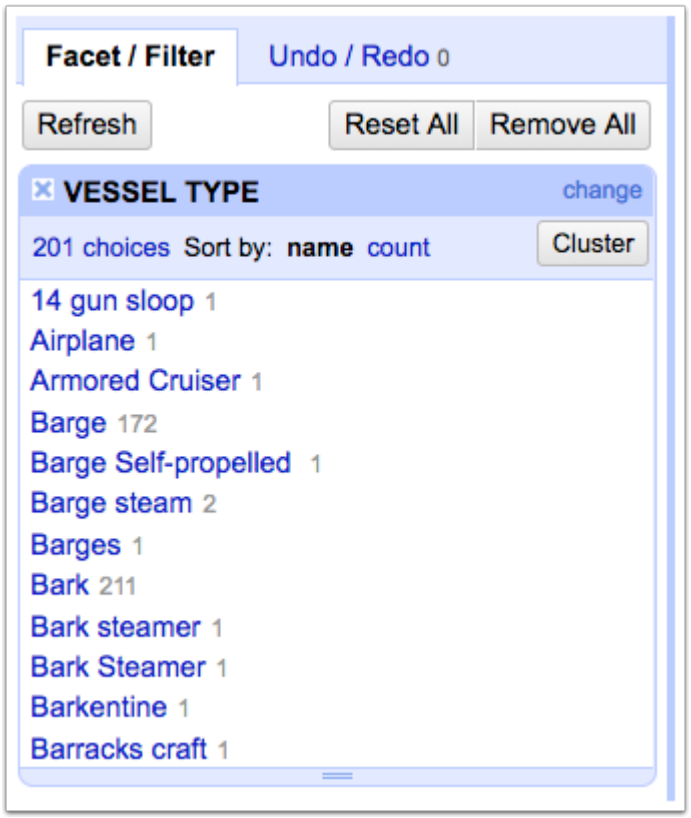
In OpenRefine, a **facet** is a way to isolate certain records that share features. It's easier to see what I mean when you try it yourself. Click on the down-arrow right next to the **VESSEL TYPE** column heading. Then select **Facet**, and then **Text Facet**.



7. Understanding facets

Look at the VESSEL TYPE list that appears on the lefthand side of the OpenRefine window. Can you tell what's going on there? OpenRefine's facet function has grouped together every term that appears in the VESSEL TYPE column, along with how many times it appears.

You can sort the list of terms alphabetically by name, or by count, according to how many times those terms appear on the list. If you click on one of the terms, only those rows that contain that term will be selected. This allows you to work on your data one chunk at a time.



8. Clean up some data

Look closely at that list of terms. You'll see that it includes two terms that are probably meant to be the same: **Bark steamer** and **Bark Steamer**. Even though a human can tell they're meant to refer to the same thing, a computer doesn't know that. So it's important to clean up this data to create accurate visualizations and analyses.

Hover over the **Bark Steamer** term in the facet list, so that you can see the **Edit** option. Press **Edit** and, in the box that appears, change **Bark Steamer** to **Bark steamer** and press **Apply**. Now the two terms will merge into one.

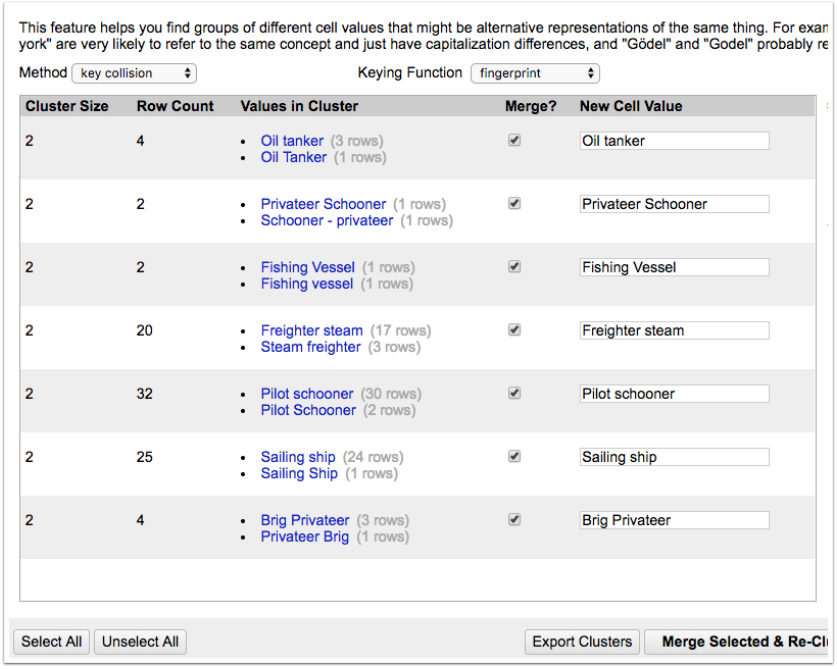
9. Another way to clean up some data

Look again at the **Facet** box. You'll see a button marked **Cluster.** Click it.

The resulting box shows you terms that OpenRefine thinks should be merged together. Check the boxes of the terms you think should be merged and then click **Merge Selected and Re-Cluster**.

Now experiment with some of the other items on the **Method** dropdown menu. What happens when you try different methods? Each uses a different algorithm to try to match terms.

When you're finished experimenting, click **Close**. You'll notice you have fewer terms in your facet list.

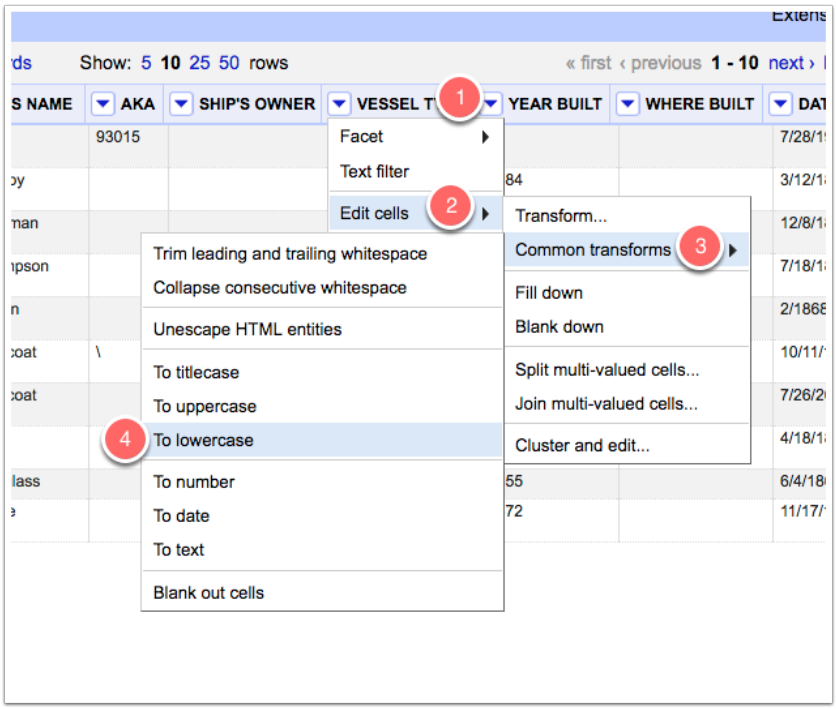


10. Change the case of an entire column

A lot of the problems with the data in the **VESSEL TYPE** were the result of variant cases (e.g., **Pilot**

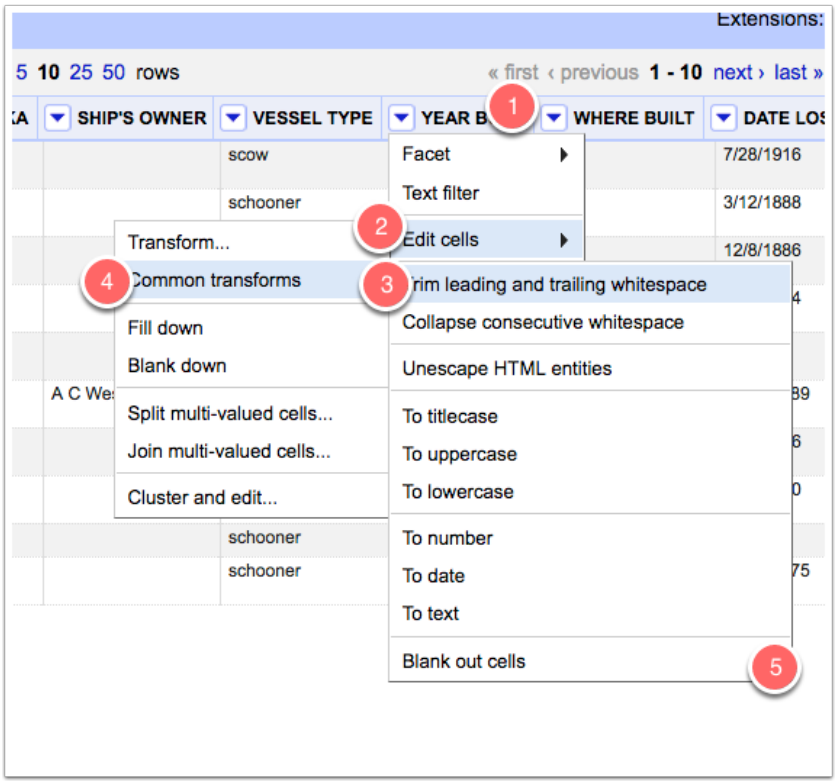
schooner versus **Pilot Schooner**). One way to eliminate these problems would be to make all of the terms lowercase. Let's do that now.

Click on the down arrow next to **VESSEL TYPE**. From the dropdown menu, click **Edit cells**, and then **Common transforms**. Finally, select **To lowercase**. Voila! All the vessel types are now lowercase.



11. Get rid of extra whitespace

One common problem with data is extra spaces before and after the values. Those are easy to get rid of with OpenRefine. On the **Year Built** column, click the down arrow, then click **Edit cells**, then **Common transforms**. Finally, click **Trim leading and trailing whitespace**. Much better!

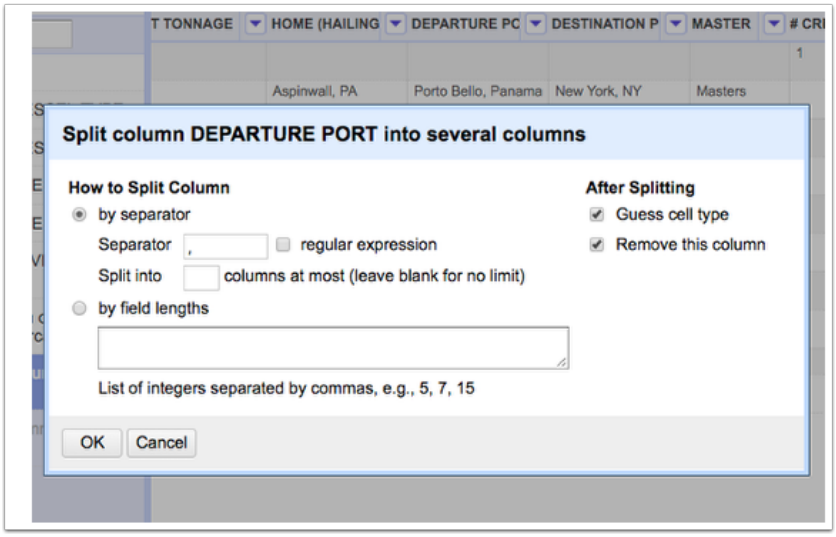


12. Split multi-valued columns

Several of our columns contain location, formatted as City, State. But let's say we want states to appear in their own column. That's easy to do with OpenRefine.

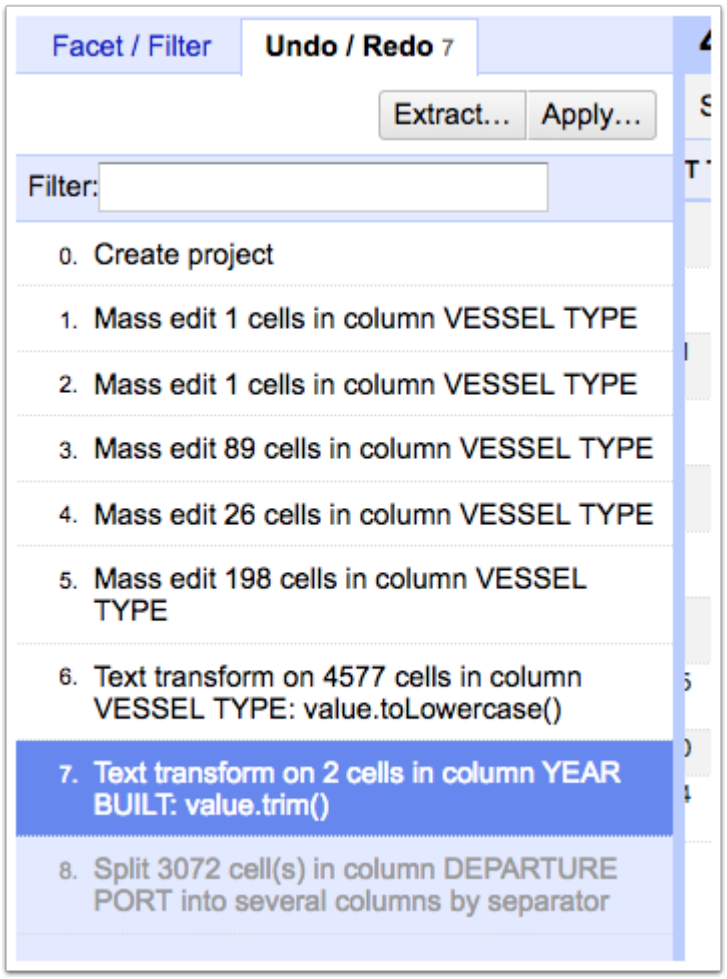
Scroll to the **Departure Point** column. Click the down arrow, then **Edit columns**, and finally **Split multi-valued cells**. The popup window asks which separator currently separates the values. Enter a comma and a space, since those are the two characters that lie between city and state. Then click **OK**.

You now have two columns! You can rename them by clicking on the down arrow, then **Edit column** and then **Rename**.



13. Undo an action

If you make a mistake in OpenRefine, no worries! It's easy to undo. Just click on the **Undo/Redo** link on the lefthand side of the screen. Then click on the next-to-last step in the list. Your last action will be reversed. If you change your mind about redoing it, you can just click the last step.



14. Add characters to selected data

Let's say we want to add the prefix **S.S.** to the name of any boat that has the vessel type **schooner**. We'll do that by first using our vessel type facet to select all the rows with the term **schooner** in the **VESSEL TYPE** column.

Once you have all of the schooners selected, head to the **SHIP'S NAME** column. Click on the down arrow, then select **Edit cells**, and then **Transform...**

The popup box that follows wants you to use a language called the Google Refine Expression Language (GREL) to transform your data. You don't have to actually know GREL; you just have to be able to look up the pattern for the expression you want to write.

When you want to add a prefix to some data in OpenRefine, the pattern looks like this:

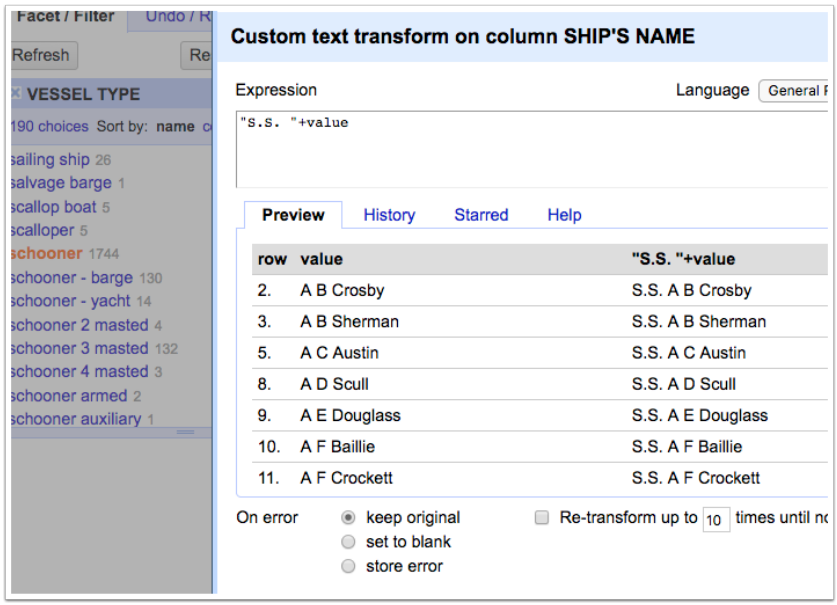
"prefix"+value

So in the blank text box, type

"S.S. "+value

You'll see a preview of how your data will look in the lower right-hand column. When you're satisfied, press **OK**.

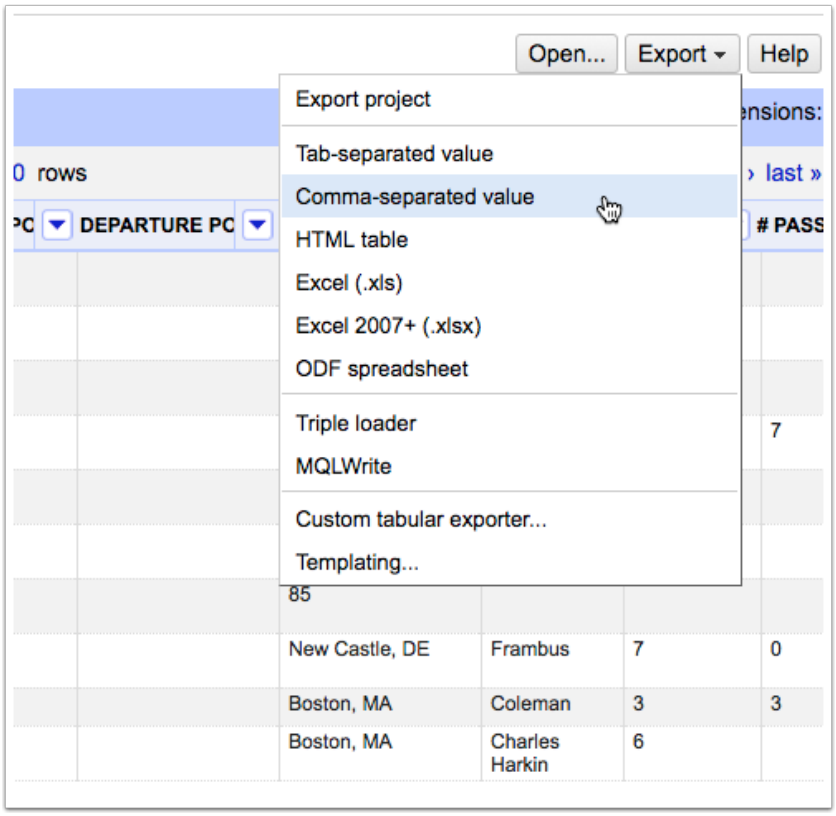
Now the title of every schooner is prefaced with "S.S."!



15. Export your data

Once you've cleaned up your data, you'll want to get it out of OpenRefine. To do that, click on the **Export** button in the upper right-hand corner. Then click on **Comma-separated value**. Your cleaned-up spreadsheet should begin downloading. You can download your data as many times as you want, at any stage of the project.

To close OpenRefine, just close the window or tab in your browser.



16. That's just the beginning!

These are some of the most common tasks you'll want to perform in OpenRefine, but OpenRefine can also handle tasks of much greater complexity. To get a sense of some of these tasks, see the resources on the **OpenRefine Resources** page: http://miriamposner.com/classes/dh101f17/tutorials-guides/data-manipulation/openrefine-resources/

OpenRefine Resources

On data-cleaning in general, see the School of Data's "Introduction to Data-Cleaning" for a helpful overview.

Our tool of choice is OpenRefine, which is installed on the lab computers and is also a free download.

OPENREFINE TUTORIALS

Maggie Pa's OpenRefine tutorial (thanks, Maggie!)

"Introduction to OpenRefine," developed by Owen Stephens on behalf of the British Library

"Cleaning Data with OpenRefine," by Seth van Hooland, Ruben Verborgh and Max De Wilde

Verborgh, Ruben, and Max De Wilde. *Using OpenRefine: The Essential OpenRefine Guide That Takes You From Data Analysis and Error Fixing to Linking Your Dataset to the Web.* Birmingham, UK: Packt Publishing, 2013. **(To get to the book, click on "EBSCO eBooks.")**