Government Documents Weeding Activity

This activity assumes no prior knowledge of OpenRefine. In this activity you will be importing a spreadsheet of data into OpenRefine and exploring it. The goal of this activity is to use a simple dataset to introduce you to the OpenRefine user interface and some of the basic types of tasks you can accomplish.

In this activity, you are going to:

A. Review the dataset and load it into OpenRefine
B. Perform some basic data cleanup to get familiar with the OpenRefine interface
C. Clustering, clean your data by combining similar values.
D. Converting a column from text to a numeric value.
E. Use OpenRefine to sort, filter and facet data
F. Restructure the dataset by removing columns and rows, and then work with Undo/Redo to roll those changes back
G. Export data from OpenRefine

Please download OpenRefine and the *Gov_Docs_Weeding_Project.xlsx* dataset prior to starting this activity. Instructions on how to download OpenRefine, the dataset, the slides and this handout can be found at **tinyurl.com/FDLC2019OPENREFINE**.

A. Review the dataset and load it into OpenRefine

1. Open the file *Gov_Docs_Weeding_Project.xlsx* in Excel and take a look at it. This is a dataset from a government documents weeding project from the University of Southern California Libraries. It contains metadata information, such as SuDoc numbers, OCLC number, title, publication date, etc., for items considered for discard in a weeding project. Notice the following: The data file has some blank columns and rows, and it has formatting applied. It has some rows at the top containing descriptive information not part of the data table. The "Permanent Call Number" column has leading spaces in it. We'll take a closer look at this in OpenRefine.
2. **Close the Excel file**. Next, **start up OpenRefine**.
3. Ensure that ***Create Project*** is selected. Click on ***Choose Files***. Browse to the file *Gov_Docs_Weeding_Project.xlsx*. Click ***Open***. Then, click the ***Next*** button.
4. You are now viewing the dataset in Preview view. Here you can see what data will look like when loaded, and make changes to what data OpenRefine will load.
   i. Notice that the descriptive text at the top of the Excel worksheet is showing in the preview, and is messing up OpenRefine's ability to identify the

column headings. We can instruct OpenRefine to ignore these rows that aren't part of the data table. **Select the check box beside *Ignore first*, and type *3* in the box** to ignore the first 3 line(s) at the beginning of the file.

ii. In the ***Project name*** box on the top of the screen, give the project a name of your choice.

iii. Click ***Create Project***.

5. Your data has now been loaded into OpenRefine. Note that it has stored a copy of this data with the OpenRefine installation files on your computer. When you make edits using OpenRefine, you are not editing the original data file you uploaded, all edits are made to the copy OpenRefine has created.

B. Perform some basic data cleanup to get familiar with the OpenRefine interface

6. In the top toolbar, select ***50*** in order to show more rows on the screen at once.
7. Let's remove the blank column. Look for the pull down menu (button with downward- pointing arrow on it) for the column named "Column". **From the pull down menu, select *Edit Column -> Remove this Column***. Do the same thing for the column named "Column2."
8. Next, look at the "Permanent Call Number" column. Remember earlier we noticed that there were leading spaces? **Hover your cursor over a cell in this column and click e*dit***. You'll see that the leading spaces are still there. **Click *Cancel*** on the edit window. These "invisible" leading spaces could cause problems down the road, so let's remove them altogether. **From the "Permanent Call Number" column pull down menu, select *Edit cells -> Common transforms -> Trim leading and trailing white space***. Check a cell to verify that the leading spaces are gone.

C. Clustering, clean your data by combining similar values.

9. Let's try the clustering function on this dataset. Go to the "Publication Date" column. **From the *"Publication Date"* column *pull down menu*, select *Facet->Text facet***. Notice we did not select "Numeric Facet." It is because OpenRefine has not recognized the numbers in the years in this column, because of the other characters in this column. We need to clean this column first! You'll see that some of them are a bit unusual, and in those cases, you may want to edit them; however, in other cases, you'll see that there are two facets that look very similar, but just have different characters added like brackets, questions marks and periods. When we have facets that look similar, we can use OpenRefine's clustering features to help improve the consistency of the values in that column. Let's take a look. **Click on the *Cluster* button** at the top of the facet window.
10. At the top of the screen, you'll see that there are different methods and keying functions you can chose from to find clusters. They roughly go from more

strict/unforgiving to looser. **Let's keep the default for now**.
Note: In this case, you should see that the column values catch small differences, but clustering can also catch differences in capitalization, typos, and plural vs singular.

11. You can see that it has found entries that it thinks are all referring to the same thing and suggests merging them under one recommended facet. You can put a check mark next to the ones you agree with, and edit the heading that you want to merge them into. **Go through and merge the entries found into new terms that only contain the year in numbers (deleting brackets and other characters) by adding a check mark under *Merge?* and editing the *New Cell Value*.**
12. When done, **click on *Merge Selected & Re-Cluster***. You might've noticed that as you did a merge, it flashed at the top of the screen how many rows were affected/mass edited.
13. If you no longer have any options with the current method, you could try the **nearest neighbor** method to get more options. As you can see, it is an iterative process to normalize your data. When you're done, with no more options to consider for merging, you can **click on *Close***.
14. In the facet window on the left side, notice that not all the years were corrected for this column after we clustered. You can still delete additional characters like brackets in the remaining years. **Hover your cursor over a year in the facet window and click e*dit. Update the year by deleting the brackets and clicking Apply.*** It has now updated multiple cells at once. Do the same for the remaining years.
15. Finally, notice that there is a value labeled **blank.** This means there is no publication year value for these cells. If none are available, you can hover your cursor over the **blank** value and click **edit**. You can add a better description and type "not available" and click **apply.** When you are done, click on the **X** on the upper left hand corner of the facet window to close the window.

D. Converting a column from text to a numeric value.
What do you do when OpenRefine incorrectly labels a value "text" when it is really a number? In this case, after we clean the "Publication Date" we need to sort by numbers, however OpenRefine still thinks this column only has textual values. We need to convert them to "numeric values."

16. **Click on the "Publication Date" pull down menu, select *Edit Cells, Common Transformations, To number.*** Now all the values in this column should be green. This is how we know that OpenRefine identifies theses values as numbers.

E. Use OpenRefine to sort, filter and facet data

17. Rows of data are initially loaded in the order they appear in the original data file. You can change how they appear by sort as text or by numbers. **From the**

Publication Date column pull down menu, select Sort, and sort by numbers, from largest first. Click on OK. Now the most recent publication dates appear first.

18. Filtering allows us to search for certain information within our dataset. Let's say we want to display only the rows from the VKC Library. **From the Library Name column pull down menu, choose *Text filter***. The text filter appears in the left-hand sidebar, under the "Facet / Filter" tab. Type ***VKC Library*** in the search box. OpenRefine automatically removes any rows that don't match from the display, leaving a total of 25 rows remaining (out of 313 total).
19. We can have text filters on more than one column at a time. **From the Publication Date column pull down menu, choose *Text filter***. Type ***1990*** in the search box for that filter. The two filters are combined, showing us the 4 government documents published in 1990 in the VKC Library.
20. You can remove a filter by clicking on the ***x*** in the top left-hand corner of the filter box. ***Remove both filters now***. You should have all 313 rows displayed again.
21. Next let's explore an even more sophisticated way of selecting which data to work with. A facet summarizes all the values that appear in the column, and lets you select which data to view, as well as provides ways to edit the data. **From the Library Name column pull down menu, choose *Facet -> Text facet***. The facet appears in the left-hand sidebar, in the same area where the filters were previously. Have a look at the facet. It shows you how many total values there are in this column, and how many rows contain each value. It allows you to sort the values by name or by count. Click on **count** to see which library has the most gov docs for potential weeding. It looks like Grand Depository has the most rows (266).
22. **Click on *VKC Library* in the value list**. This has the same effect as using the text filter to search for the VKC Library, leaving 25 matching rows. However, from there we can do more than the filter allowed. We can select a second value at the same time. **Hover your cursor over *Grand Depository* in the value list and choose *include***. You will notice that the included values are now in red. And you now have 291 matching rows. You can then exclude one or both of the selections at any time. **Hover your cursor over *the VKC Library* in the value list and choose *exclude***. Now only items from the Grand Depository rows are shown.
23. Like with filters, you can combine multiple facets at the same time. **Add another text facet on the Resource Type – Bibliographic Details by going to the column pull down menu, and choose *Facet -> Text facet*.** Now you can see what resource types are in the Grand Depository. **Click on "Books-Physical."** It looks like there are 257 matching rows for physical books at the Grand Depository. Once you are done you can, ***reset*** **both facets** by clicking on the **X** on the corner of each facet.

F. Restructure the dataset by removing columns and rows, and then work with Undo/Redo to roll those changes back

24. Let's say we're unhappy with entire columns – data we can't do much with, and don't want. Well we can remove whole columns. There are a couple ways to do this. One way is from the pull down menu. Let's say we don't want the column ***MMS id***. **From the MMS id column *pull down menu*, select *Edit column->Remove this column***.
25. Another way to bulk remove columns is to go to the special ***All*** column pull down menu on the far left. **From the *All* column *pull down menu*, select *Edit columns->Re-order / remove columns...* From here you can drag columns from the left to the right to remove them – do this for *Location Name, Material Type- Physical Item Details* and *Network Number***. We can also reorder columns. **Move Title to the top before Permanent Call Number and Publication date after Permanent Call Number** to move those columns more to the left. Once we're done, **click on OK** to make the changes.
26. Instead of working with columns, we can also work with rows. Another feature of OpenRefine is the ability to flag or star certain rows and facet by this flag. An easy way to flag rows is to just click on the flag symbol next to a row of interest – **try flagging the first few rows of our dataset**.
27. We can also facet our dataset to show certain rows and then automatically flag those rows. For example, to see how many rows have a blank value for a particular column, you can facet by blanks. **From the *Resource Type-Bibliographic Details pull down menu*, select *Facet->Customized facets- >Facet by blanks.* Click on true** to show only the row where that column is blank. This shows that we have a row with no information. To delete the row click the downward arrow on the **All column → Edit Row → Remove all matching rows**. You have just deleted the row that provided no information. You can click on the **X** on the **Facet** window on the left side to see all rows.
28. Let's flag the rows that are journals in the **Material Type-Bibliographic Details** column. **From the *Material Type-Bibliographic Details* column *pull down menu*, select *Facet->Text facet* Select the *Journal* facet.** Now we have a subset of rows that are just journals. Let's say that these 11 rows were no good to use. We could flag them (or star them or remove them). **From the *All* column *pull down menu*, select *Edit rows->Flag rows***
29. Finally, reset all facets by **clicking on the *Reset All button*** on the left above the facet windows. Now you should see all the rows in your dataset again, some are flagged and some are not. Click on the **X** in the facet window to close the window.
30. Later if you decide that you want to remove those flagged rows that you were unsure of, you can. **From the *All* column *pull down menu*, select *Facet->Facet by flag*** and then **select true** from the facet window to show only your 11 flagged rows. Finally, you can delete all of them. **From the *All* column *pull down menu*, select *Edit rows->Remove all matching rows***. All the flagged rows should now be removed from the dataset.
31. Reset all facets again by **clicking on the *Reset All button*** on the left above the facet windows to see your remaining rows. Click on the **X** in the facet window to close the window.
32. We've done a lot to our dataset. But what happens if you do a few things, and then wish you could take some of it back! Well you can with OpenRefine's undo/redo

features. **Click on the Undo/Redo tab** above where the facets show up.

33. You'll see a number of steps that outlines everything you did to this dataset. It is a great way to keep track of what you've done. You can also roll back your changes to a previous version by clicking on the last step you were happy with. Then everything after that has been rolled back. You can go back and forth in time to take a look at the dataset at a particular point. For example, **click on the item that says *Reorder columns***. You'll see that the steps after that have greyed out, which means they haven't happened yet. So for this example, those flagged rows have now not be deleted, and you should see them in your dataset. If you go back to a previous step (like we've just done), and then start making new changes/transformations - all the subsequent steps will be deleted permanently.
34. **Go ahead and try it by starring some rows this time.** You should see that the steps we did to flag and delete rows have been replaced by our new starring rows action.

    Note: If we had a similarly structured dataset – perhaps for a different snapshot in time – and we wanted to perform the same steps that we had done on this dataset, we could, by clicking on the ***Extract button****. We would then select the steps we wanted to repeat. You'll see code in the window to the right describing the steps. You would then copy that code and save it in a text file to keep a copy of your steps. Later if you load up your new dataset, you could go back to the Undo/Redo tab and select Apply and paste in this code into the window to run those steps on the new dataset.*

G. Export data from OpenRefine

35. **In the top right-hand corner of the screen, pull down the *Export* menu and choose Excel** or whatever format you would like to download.