Issues in Science and Technology Librarianship

Cleaning Collections Data Using OpenRefine

Elizabeth Sterner, MLIS, MS
Health and Human Services Librarian
Assistant Professor
Governors State University
esterner@govst.edu

Collection maintenance, including weeding, is a key component of my position as an academic science librarian. In an ideal world we receive perfect data that are clean and ready to use. But unfortunately, that is not always the case. In large deselection projects you might receive holdings and circulation records in separate files which, once combined, may contain many undesired duplicated line items. I will demonstrate how you can effectively and quickly use the facet row feature in OpenRefine to deduplicate data. The benefit of this method is that you select which of the duplicated items will be kept and which will be deleted. Once OpenRefine is downloaded and opened, you work in a web user interface to upload your data, clean and transform the data, and then download from the browser to a CSV file in Excel. With practice, I have found that this only takes a few minutes for thousands of line items, and ensures I am able to select the data I want.

Why OpenRefine?

I choose to use OpenRefine, an open source tool available on GitHub, because it is free, easy to learn and use, and allows for the selection of a union of rows after applying facets and filters. OpenRefine offers a platform for data management; and users, including non-programmers, do not need to know how to code to clean and transform data. With OpenRefine, I can ensure I am deleting specific line items while confirming I am keeping preferred line items.

Method

Basically, I divide my data into numerous sets, and then I select the union of desired sets. I have used this method during large weeding projects when I received two Excel files.

In this example, I received the data necessary for the deselection process in two different files. The first file of all holdings included the following fields: DISPLAY_CALL_NO, TITLE, PUB_PLACE, PUBLISHER, PUBLISHER_DATE, ITEM_BARCODE, and LOCATION_ID (Figure 1).

DISPLAY_CALL_NO	TITLE	PUB_PLACE	PUBLISHER	PUBLISHER_DATE	ITEM_BARCODE	LOCATION_ID
RT31 .A25	Book of Nursing	Amherst, N.Y.	Publisher A	2018	123456	1
HV91 .H424	Social Work Book	Chicago, IL	Publisher B	2016	987654	2
QD33 .S73	Chemistry	New York, NY	Publisher C	2015	345678	2
QH442 .G4462009	Genetic Engineering	Detroit, MI	Publisher B	2009		1

Figure 1: Example of records holdings.

The second file of items circulated included the following fields: DISPLAY_CALL_NO, TITLE, PUBLISHER, PUBLISHER_DATE, ITEM_BARCODE, HISTORICAL_CHARGES, and LastOfCHARGE_DATE (Figure 2).

DISPLAY_CALL_NO	TITLE	PUBLISHER	PUBLISHER_DATE	ITEM_BARCODE	HISTORICAL_CHARGES	LastOfCHARGE_DATE
RT31 .A25	Book of Nursing	Publisher A	2018	123456	1	20-Aug-19
QD33 .S73	Chemistry	Publisher C	2015	345678	3	10-May-18

Figure 2: Example of circulation statistics.

Combining Files

Before I can go any further, I want to combine the two data files into one file. I want all the information available in one file to increase my efficiency while in the stacks reviewing the collection, and to provide one file of data to the faculty members of my liaison areas. I combine the data with some simple manipulations in Excel. First, I make sure that all columns match (Figure 3). Then I copy and paste the smaller circulation file to the significantly larger file of holdings (Figure 4). The smaller circulation files typically holds approximately 200-3,000 lines of data. The larger files of holdings hold between 5,000-20,000 lines of data.

DISPLAY_CALL_NO	TITLE	PUB_PLACE	PUBLISHER	PUBLISHER_DATE	ITEM_BARCODE	LOCATION_ID
RT31 .A25	Book of Nursing	Amherst, N.Y.	Publisher A	2018	123456	1
HV91 .H42	Social Work Book	Chicago, IL	Publisher B	2016	987654	2
QD33 .S73	Chemistry	New York, N	Publisher C	2015	345678	2
QH442 .G4	Genetic Engineering	Detroit, MI	Publisher B	2009		1

DISPLAY_CALL_NO	TITLE	PUB_PLACE	PUBLISHER	PUBLISHER_DATE	ITEM_BARCODE	LOCATION_ID	HISTORICAL_CHARGES	LastOfCHARGE_DATE
RT31 .A25	Book of Nursing		Publisher A	2018	123456		1	20-Aug-19
QD33 .S73	Chemistry		Publisher C	2015	345678		3	10-May-18

Figure 3: Mismatch of combined data.

DISPLAY_CALL_NO	TITLE	PUB_PLACE	PUBLISHER	PUBLISHER_DATE	ITEM_BARCODE	LOCATION_ID	HISTORICAL_CHARGES	LastOfCHARGE_DATE
RT31 .A25	Book of Nursing	Amherst,	Publisher A	2018	123456	1		
HV91 .H42	Social Work Book	Chicago, IL	Publisher B	2016	987654	2		
QD33 .S73	Chemistry	New York, N	Publisher C	2015	345678	2		
QH442 .G4	Genetic Engineering	Detroit, MI	Publisher B	2009		1		
RT31 .A25	Book of Nursing		Publisher A	2018	123456		1	20-Aug-19
QD33 .S73	Chemistry		Publisher C	2015	345678		3	10-May-18

Figure 4: Example of combined data with duplicates.

This first step creates duplicates of items, books in this case, that had been previously checked out. This is not something that I want to deduplicate by hand. I opt to use OpenRefine so that I can control which duplicate line item I am deleting. OpenRefine provides the functionality of the facets for my desired data cleanup. After downloading and opening OpenRefine, I create a project (Figure 5).

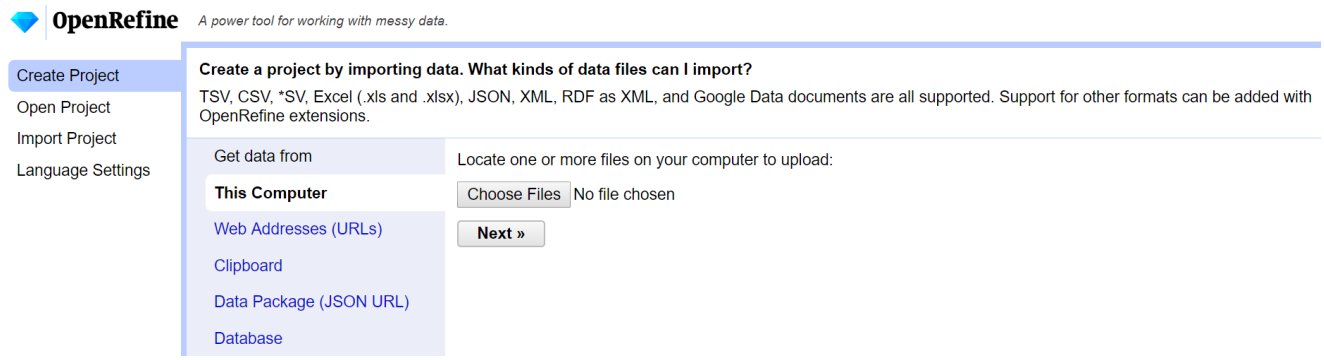

Figure 5: Screenshot of creating a project in OpenRefine.

Selecting Non-Duplicated Items

My first step is to select items which are not duplicated. That is, they are in the collection, but they have not been checked out. To do this, I need to transform the data into "Text" first. I select "Edit cells," "Common transformations" and "To text" under ITEM_BARCODE (Figures 6-7).

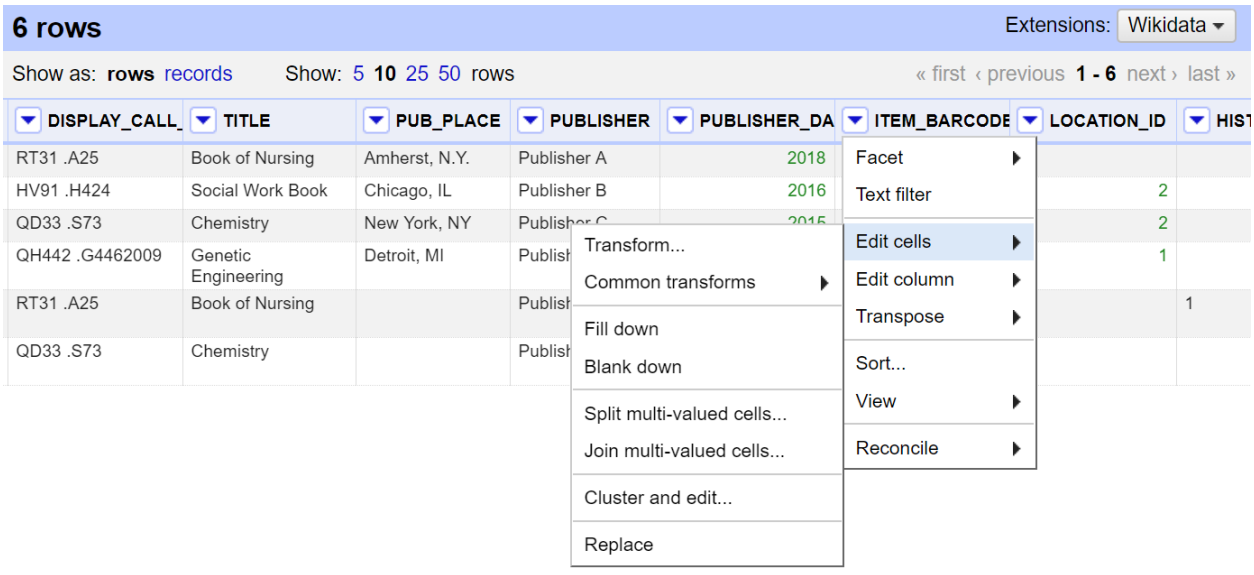

Figure 6: Screenshot of preparing ITEM_BARCODE data.

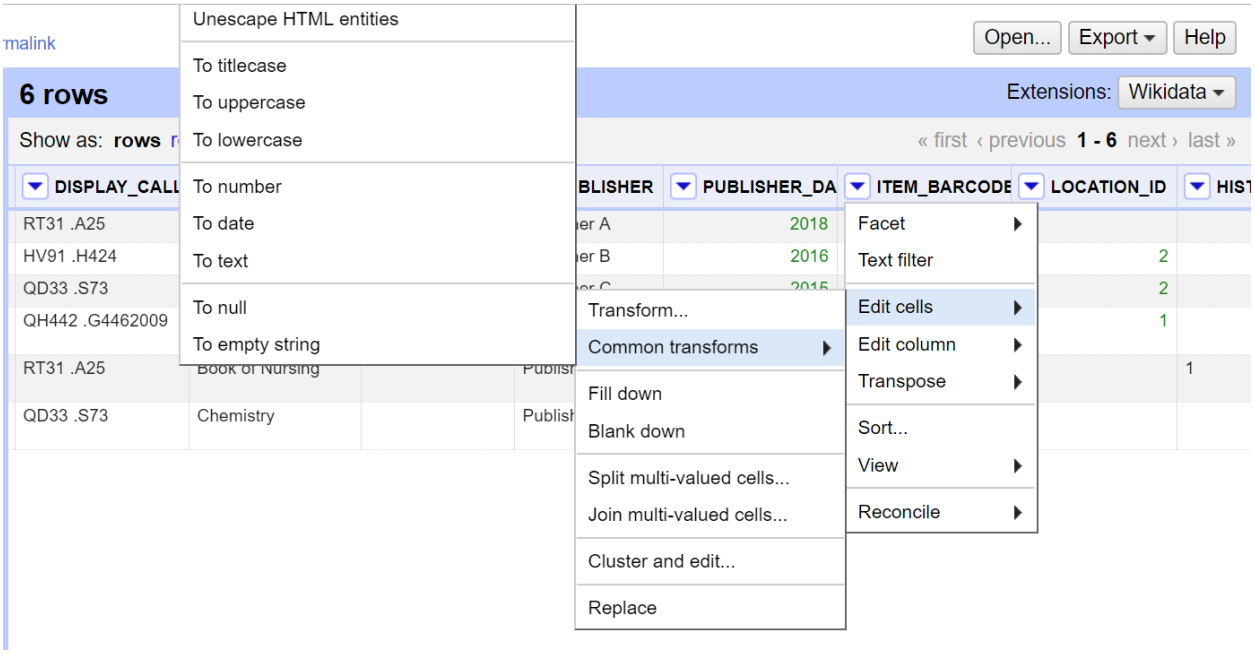

Figure 7: Screenshot of transforming ITEM_BARCODE data.

Next, I want to find the values that are not duplicates. I select "Facet," "Customized facets" and "Duplicates facet" on ITEM_BARCODE. I select "false" in the ITEM_BARCODE box on the left-hand side. This will find all items that do not have a duplicate in the same column. After selecting "false," only the non-duplicated items will show (Figures 8-10).

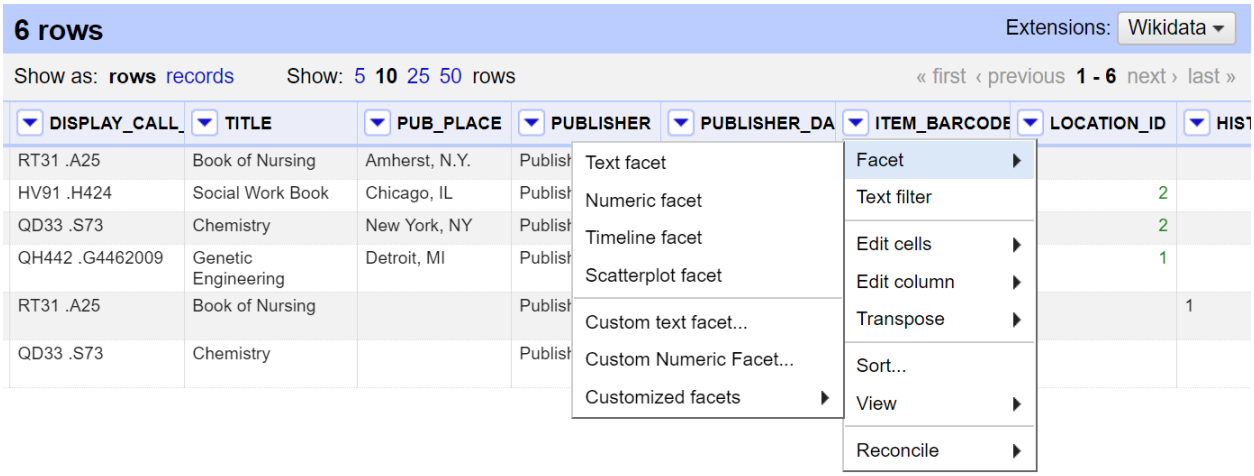

Figure 8: Screenshot of selecting the type of facet in ITEM_BARCODE.

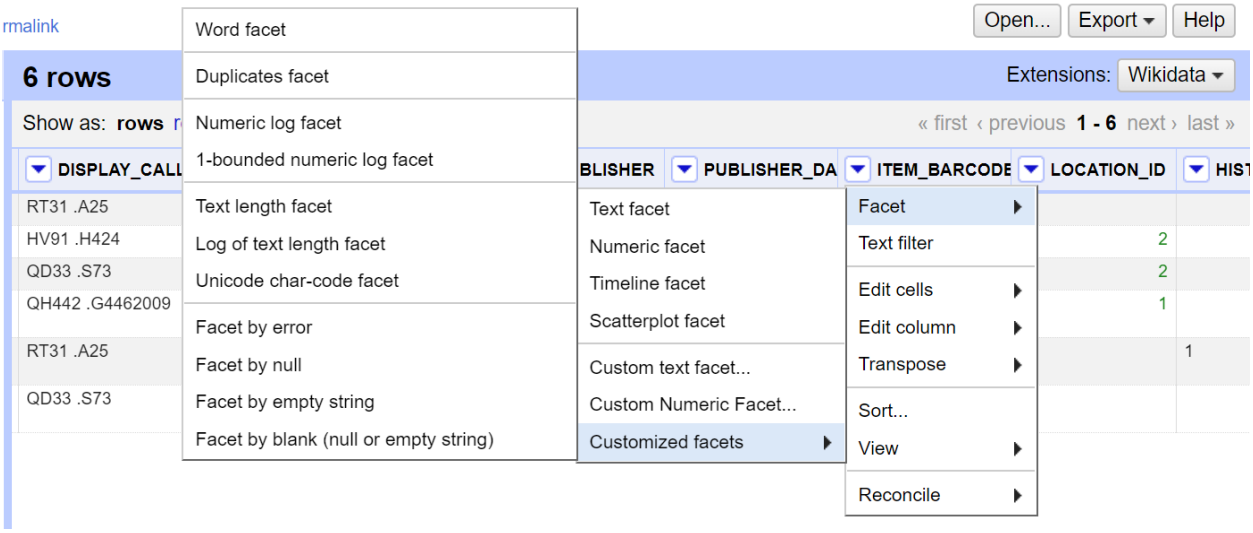

Figure 9: Screenshot of faceting ITEM_BARCODE by duplicates.

Figure 10: Screenshot of "false" selection of ITEM_BARCODE duplicates.

After selecting "false" to find all items that do not have a duplicate, I navigate to the All column with stars. I then select "Edit rows" and "Star rows," and then close the ITEM_BARCODE box on the left-hand side (Figures 11-12). Closing the ITEM_BARCODE box will display all the data while maintaining the stars on the non-duplicated data (Figure 13).

Figure 11: Screenshot of starring row in All column.

Figure 12: Screenshot of selected starred rows.

Figure 13: Screenshot of selected data starred within all data.

Selecting all Duplicates & 'Starring' Certain Items

Next I want to select all duplicated items and star only those items which contain LastOfCHARGE_DATE data. To do this, I select "Facet," "Customized facets," and "Facet by blank (null or empty string)" on the LastOfCHARGE_DATE column (Figures 14-15).

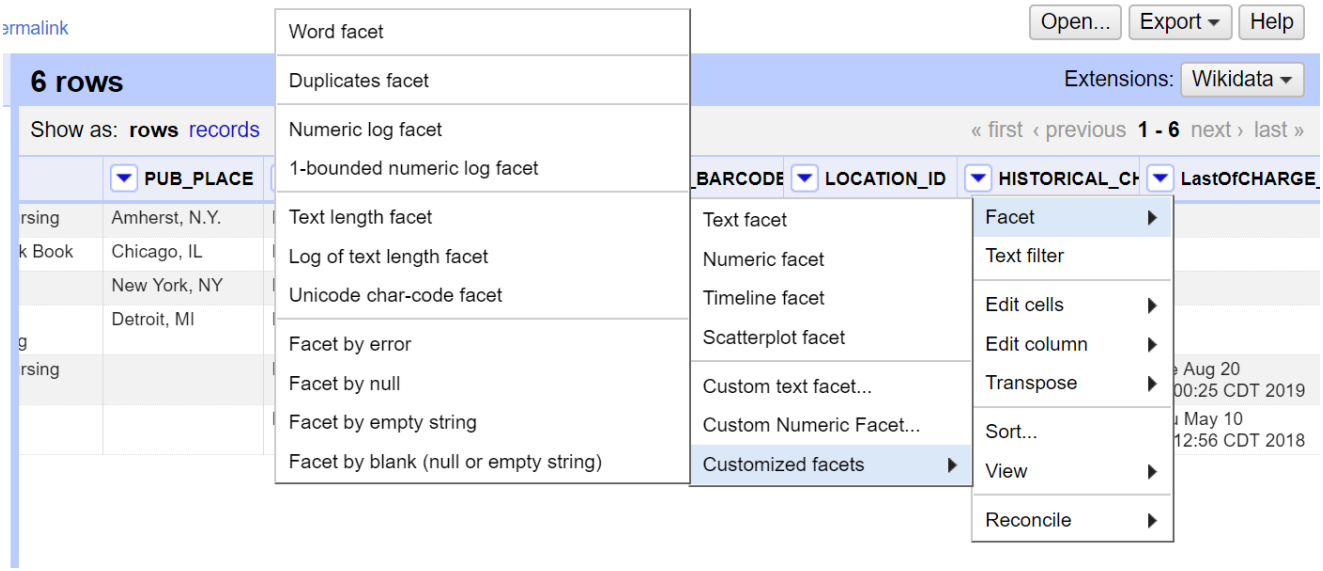

Figure 14: Screenshot of selecting the type of facet in LastOfCHARGE_DATE.

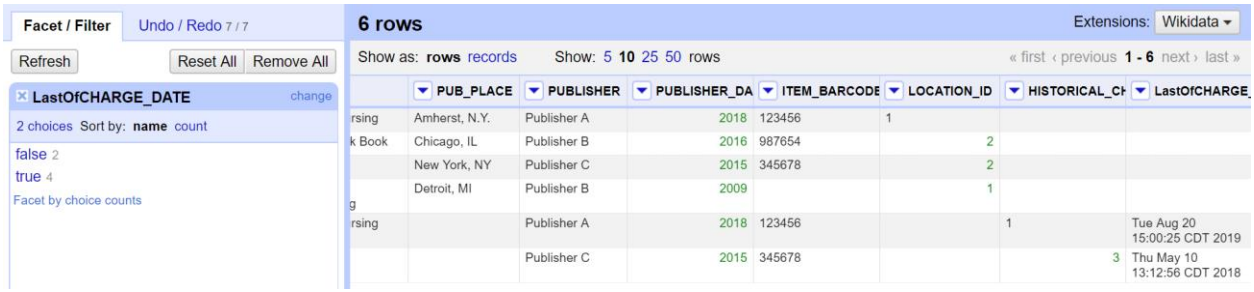

Figure 15: Screenshot of "false" selection of LastOfCHARGE_DATE.

I want the items that are not blank, and so I select "false." After selecting "false" to find all items that have information in the LastOfCHARGE_DATE column, I navigate to the All column with stars. I select "Edit rows" and "Star rows," and then close the LastOfCHARGE_DATE box (Figures 16-17). Closing this box will display all the data while maintaining the stars on the non-duplicated data and the duplicated item lines with LastOfCHARGE_DATE data (Figure 18).

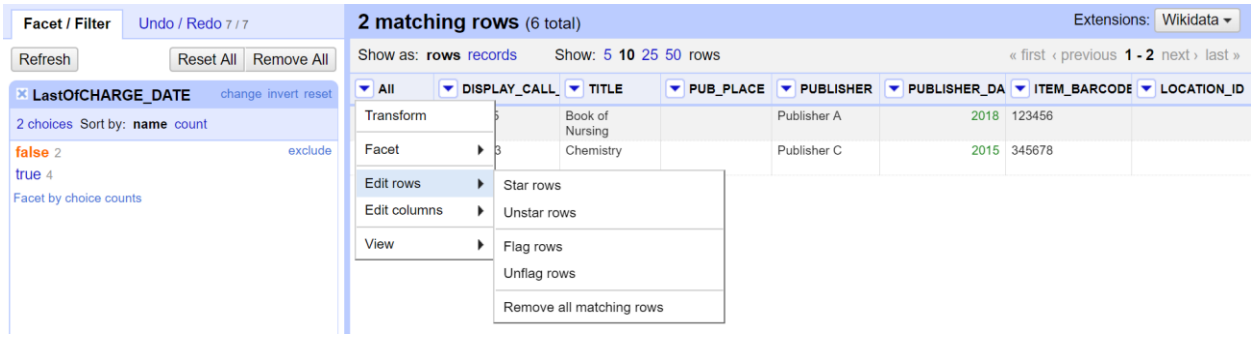

Figure 16: Screenshot of starring row in All column.

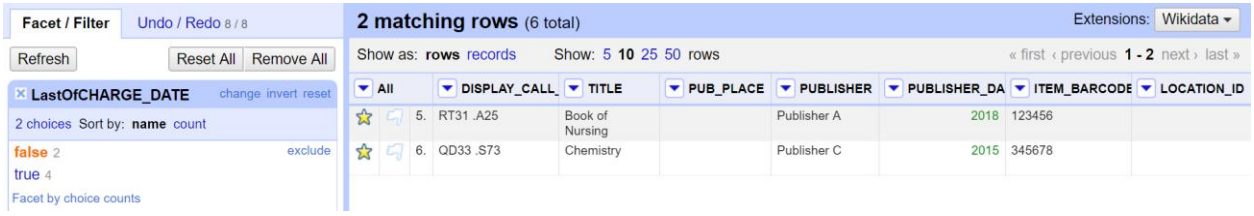

Figure 17: Screenshot of selected starred rows.

Figure 18: Screenshot of selected data starred within all data.

Ensuring All Data are Included

Then, I want to ensure that I have included all data. It is possible that I did not have a barcode number for all items. In my example, the single blank barcode was detected when I looked for duplicated items. However, if you have multiple blank barcode items, this would have been overlooked. To check for blank barcodes, select “Facet,” “Customized facets” and “Facet by blank (null or empty string)” on the ITEM_BARCODE column (Figures 19-20).

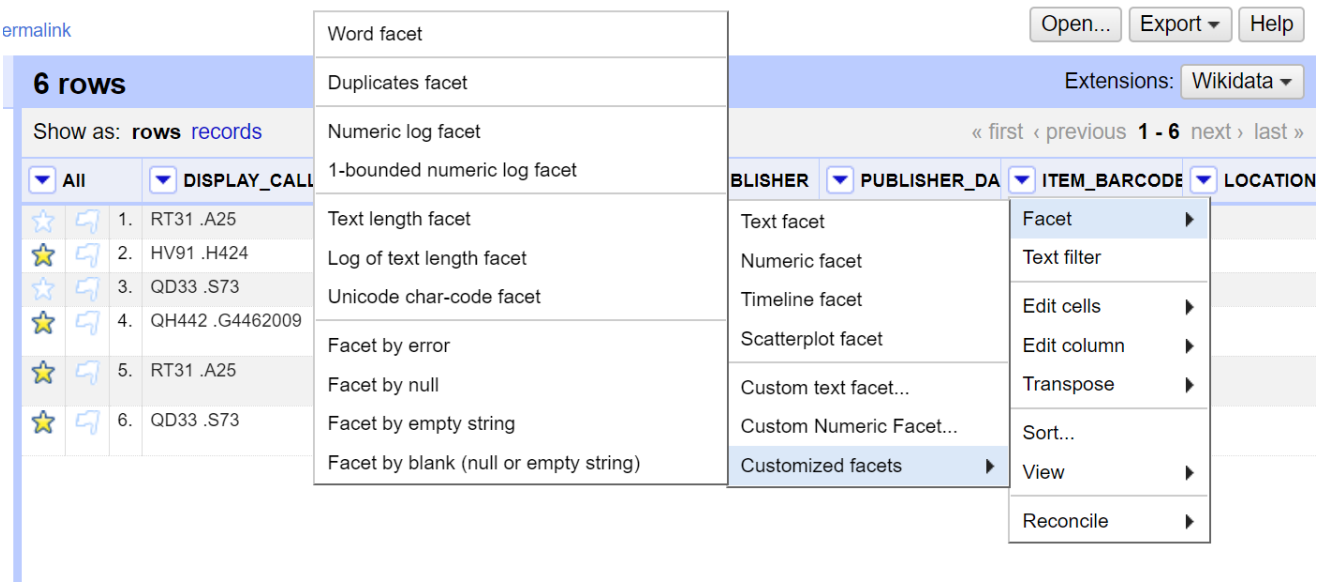

Figure 19: Screenshot of selecting the type of facet in ITEM_BARCODE.

Figure 20: Screenshot of "true" selection of ITEM_BARCODE.

In this case, I want to select "true" because I am looking for blank barcode lines. Again, select "Edit rows" and "Star rows," and then close the ITEM_BARCODE box (Figures 21-22). Closing this box will display all the data while maintaining the stars on the non-duplicated data, the duplicated item lines with ITEM_BARCODE data, and any items without barcodes (Figure 23).

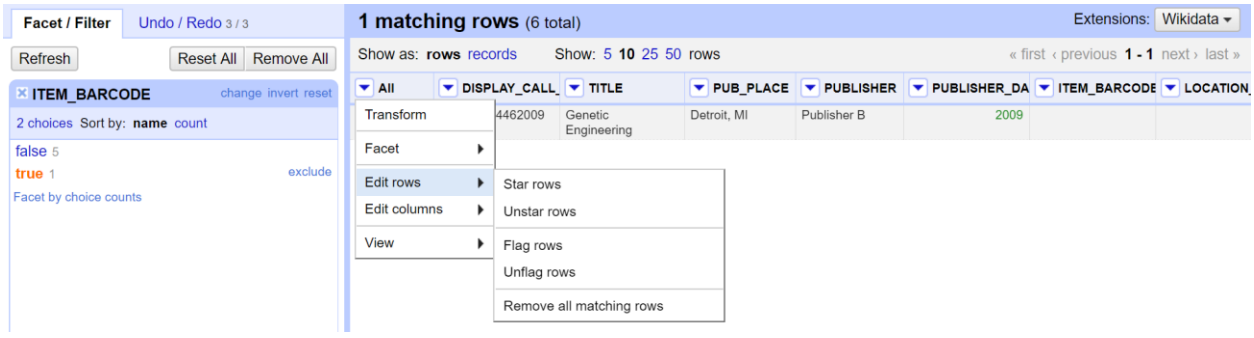

Figure 21: Screenshot of starring row in All column.

Figure 22: Screenshot of selected starred rows.

Figure 23: Screenshot of selected data starred within all data.

Selecting Only "Starred" Items

Next, I want to select only the data that have been marked with a star. This includes all items that have no barcodes, do not have a duplicate, and duplicated items with the historical charge information. To do this, select "Facet" and "Facet by star" on the All column with stars (Figure 24). Then select "true" (Figures 25-26). This will display all the desired items.

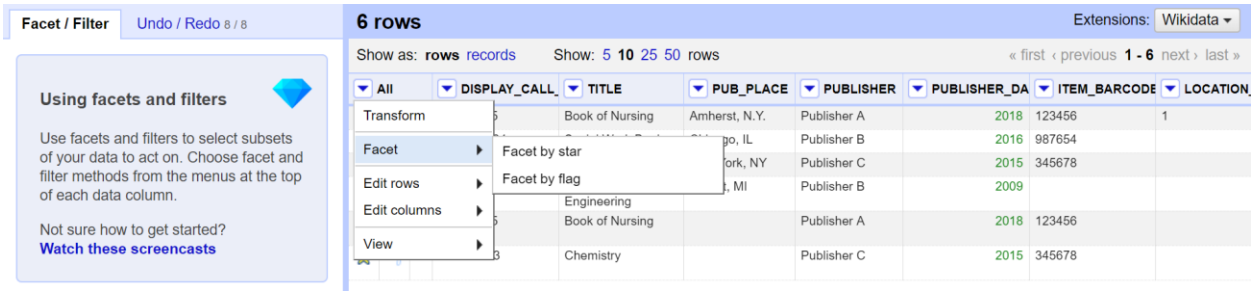

Figure 24: Screenshot of selecting the type of facet in All.

Figure 25: Screenshot of selected starred rows.

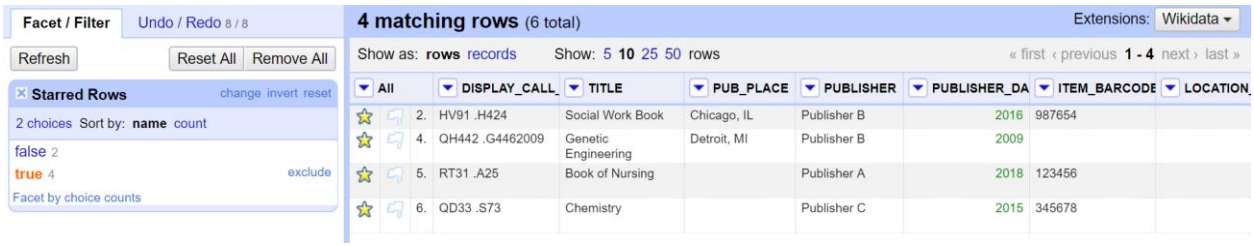

Figure 26: Screenshot of selected data starred.

Almost Done! Exporting Data

Finally, you can export your data into many different forms. Do not close the "Starred Rows" box. Select "Export" in the upper right hand corner and select your desired format (Figure 27). I'll use the Excel format (Figure 28). Once in Excel, I prefer to format columns, e.g. wrapping text, in order to make the table more visually pleasing.

Figure 27: Screenshot of Export options.

DISPLAY_(	TITLE	PUB_PLAC	PUBLISHE	PUBLISHE	ITEM_BAR	LOCATION	HISTORIC/	LastOfCHARGE_DATE
HV91 .H42	Social Wor	Chicago, IL	Publisher B	2016	987654	2		
QH442 .G4	Genetic En	Detroit, MI	Publisher B	2009		1		
RT31 .A25	Book of Nursing		Publisher A	2018	123456		1	Tue Aug 20 15:00:25 CDT 2019
QD33 .S73	Chemistry		Publisher C	2015	345678		3	Thu May 10 13:12:56 CDT 2018

Figure 28: Screenshot of exported data from OpenRefine.

Conclusion

With these few steps, I have used the facet row feature to ensure that I have kept the correct duplicate line with the historical charge information while also ensuring that I have kept non-duplicated items or items without barcodes in the file. OpenRefine has free online courses available (OpenRefine), and other articles provide detailed descriptions of how to get and use OpenRefine (Groves 2016; Hill 2016). I used this method as part of the weeding project, but it could be applied to any situation with duplicated data. After completing this process, I have effectively and quickly deleted selected duplicated information to improve my workflow during the deselection process.

References

Groves, A. 2016. Beyond Excel: how to start cleaning data with OpenRefine. *Multimedia Information & Technology* 42(2): 18-22.

Hill, K.M. 2016. In search of useful collection metadata: using OpenRefine to create accurate, complete, and clean title-level collection information. *Serials Review* 42(3): 222-228. DOI: 10.1080/00987913.2016.1214529.

OpenRefine [Internet]. [cited 2019 July 15]. Available from http://openrefine.org/documentation.html