Data Cleaning in OpenRefine: Workshop Guide

JHU Data Services

February 17th, 2023

# Introduction

This guide is available to help you during the Data Cleaning in OpenRefine workshop offered by JHU Data Services. We encourage you to follow along during the live workshop and use this guide if you miss a step. For each step, terminology and buttons in the OpenRefine interface are bolded.

The dataset for this class comes from the National UFO Reporting Center (NUFORC). The data file will be shared with workshop participants via the Zoom chat. Find out more about this data source on their website: <http://nuforc.org/>

If you have any questions about OpenRefine or any of the transformations, ask the instructors in the Zoom chat. You can find more information about OpenRefine on its website: <https://openrefine.org/> and in the official documentation: <https://docs.openrefine.org/>

# 1. Create a new project in OpenRefine

1. Open the OpenRefine application. OpenRefine will open as a new tab in your web browser; Google Chrome or Microsoft Edge are recommended.
2. In the left column, select the **Create Project** tab.
3. In the middle column, select get data from **This Computer**.
4. Click **Choose Files**, select the data file on your computer, and click **Next**.
5. You should see a preview of the data at the top of your screen and option for parsing the data at the bottom. We will keep all default settings for this workshop.
6. In the top row, enter a new Project name (optional) and click **Create Project**.

# 2. Reduce number of options in Shape column

1. Open a **Text Facet** on the **Shape** column: Click the down arrow next to Shape, hover over Facet, and click Text Facet.
2. Hover over “Egg” in the facet window. Click **Edit** in the shaded section to the right of Egg.
3. In the pop-up window, delete “Egg” and type “Oval” in its place. Click **Apply**.
4. Fill-in blank cells: Scroll to the bottom of the text facet. Hover over “(blank)” and click **Edit**.
5. Type your desired missing value code in the text box (this workshop will use NA). Click **Apply**.
6. Remove the text facet by clicking the “**x**” in the top-left of the facet window, or clicking **Remove All** along the top of the Facet/Filter column.

# 3. Use clustering to standardize Duration column

1. Open a **Text Facet** on the **Duration** column.
2. Click **Cluster** in the top-right corner of the facet window.
3. Look through the identified clusters. Make sure they are related values and that the New Cell Value is the desired new value.
4. Change in the cluster for “5 sec” and “5 sec.”, change the New Cell Value to “5 seconds”
5. In the bottom-left corner of the Cluster window, click **Select All**.
6. Along the bottom row, click **Merge Selected & Re-Cluster**.
7. If there are no new clusters, change the **Keying Function** to “ngram-fingerprint”. New clusters should be identified.
8. Check the “**Merge?**” box for any cluster that you want to merge. Remember to edit the New Cell Value to adhere to our chosen standard of “hours”, “minutes” and “seconds” for units of time.
9. Click **Merge Selected & Re-Cluster**.
10. Once you are satisfied with your clusters, click **Close**.

More information on clustering methods: <https://docs.openrefine.org/manual/cellediting#clustering-methods>

# 4. Write GREL expression statements to standardize Duration column

1. Open a **Transform** window on the **Duration** column: Click the down arrow next to Duration, Hover over Edit Cells, and click Transform…
2. A **Custom text transform** box will pop up. Keep the default language: General Refine Expression Language (GREL).
3. In the Expression box, type: **value.replace(“mins”,“minutes”)**
4. Click **OK**. You should see a yellow box at the top of your screen declaring the number of cells that were transformed.
5. Open a new **Transform** window on the Duration column.
6. In the Expression box, type: **value.replace(“secs”,”seconds”).replace(“hrs”,”hours”)**
7. Click **OK**.
8. Fill-in blank cells: Scroll to the bottom of the Duration text facet. Hover over “(blank)” and click **Edit**.
9. Type your desired missing value code in the text box (this workshop will use NA). Click **Apply**.

Regular expression (regex) cheat sheet: <https://regexcheatsheet.com/>

More information on regular expressions in OpenRefine: <https://docs.openrefine.org/manual/expressions#regular-expressions>

# 5. Remove duplicate rows

1. Open a **Text Facet** on the **Summary** column.
2. Scroll to the bottom of the facet. Hover over “(blank)” and click **Edit**.
3. Type your desired missing value code in the text box (this workshop will use NA). Click **Apply**.
4. Within the **Summary** column, turn duplicate rows into blank cells: Click the down arrow next to Summary. Hover over “Edit Cells” and click **Blank down**.
5. Click the down arrow next to Summary. Hover over Facet, hover over **Customized facets**, and click the last option: **Facet by blank (null or empty string)**.
6. A new facet should have appeared in the Facet/Filter column with two options: false and true.
7. Click **true** in this facet to show only the rows where the Summary column is blank.
8. Click the down arrow in the **All** column (left-most column). Hover over **Edit Rows** and click **Remove matching rows**.
9. Close the Summary facets by clicking the “**x**” in the top-left of each facet window, or clicking **Remove All** along the top of the Facet/Filter column.

# 6. Split Date\_Time column into two separate columns: Date and Time

1. Click the down arrow next to **Date\_Time**. Hover over **Edit Column** and click **Split into several columns…**
2. Keep the button for **by separator** checked. In the **Separator** text box, remove the comma and enter one blank space.
3. Under **After Splitting**, uncheck the box for **Remove this column**.
4. Click **OK**. You should now have two additional columns in your dataset: Date\_Time 1 and Date\_Time 2.
5. Rename the new columns into Date and Time
   1. Click the down arrow next **to Date\_Time 1**. Hover over **Edit Column** and click **Rename this column**. Type “Date” in the text box. Click **OK**.
   2. Click the down arrow next to **Date\_Time 2**. Hover over Edit Column and click **Rename this column**. Type “Time” in the text box. Click **OK**.

More information on the datetime format: <https://www.w3.org/TR/NOTE-datetime>

# 7. Use reconciliation to retrieve full state names from abbreviations in State column

1. Click the down arrow next to **State**. Hover over **Reconcile**. Click **Start reconciling…**
2. Click **Wikidata** as the service.
3. Keep the first button checked: **state of the United States**. Leave all other fields alone.
4. Click **Start Reconciling**.
5. In the Facet/Filter column, look at the **State:judgement** facet. Click **none** to display only the rows that were not matched through reconciliation.
6. At the bottom of each cell in the State column, click **Search for match**.
7. Keep the button marked for **Match other cells with same content**. In the text box, choose the first option matching this cell to a **province of Canada**.
8. Repeat this process to search for matches for all unmatched states.
9. In the Facet/Filter column, click **Reset All** to bring our full dataset back into view.

# 8. Reconcile City names against State column to retrieve correct location

1. Clean-up variation in the **City** column:
   1. Click the down arrow next to City. Hover over **Edit cells**, hover over **Common transforms**, and click **To titlecase**.
   2. Open a **Text Facet** for the City column.
   3. Click **Cluster**.
   4. Merge desired clusters using the default clustering methods and close the Cluster pop-up box.
   5. Fill-in blank cells with “NA”.
2. Click the down arrow next to **City**. Hover over **Reconcile**. Click **Start reconciling…**
3. Click **Wikidata** as the service.
4. Keep the first button checked: **city of the United States**.
5. In the right-hand column, **check the box next to State**.
6. In the text box, type “**part of**” and choose it from the drop-down menu.
7. Click **Start Reconciling**.
8. Click the down arrow next to City. Hover over **Reconcile**, hover over **Actions**, and click **Match each cell to its best candidate**.

# 9. Add geographic coordinates for UFO sightings

1. Click the down arrow next to **City**. Hover over **Edit column**, click **Add columns from reconciled values…**
2. Click **coordinate location** from the list of Suggested Properties.
3. Click **OK**.

# 10. Export cleaned data and OpenRefine project

Option 1: Export dataset only

1. Click **Export** in the top-right corner of the screen.
2. Choose desired file format. Options include comma-separated value (CSV), Excel, Google Sheets and more.

Option 2: Export OpenRefine project

1. Click **Export** in the top-right corner of the screen.
2. Click **OpenRefine project archive to file**. This will save a .tar.gz file with the cleaned data and a record of all transformations, in a format that can be imported back into OpenRefine.