OpenRefine Data Standardisation

Last Update: 2019 August 22, Anna Leary

Process

1. Facet based on at least one common column value/theme
   1. This should be a datapoint that is consistent throughout the dataset and, theoretically, a main way the data will be organised
      1. Examples: catalogue/identification number, locality number, locality description, collector name, taxonomic name, physical location in the collection

***Tip:*** *If working with a physical collection, the way in which it is organised/housed is a good place to start with faceting to standardise the data*

1. Use several strategies to begin cleaning and standardising the data
   1. Cell Editing
      1. Text Facets
         1. Single datapoint: On the left side where the facets appear, select the option to ‘edit’ the text of the datapoint
         2. Multiple similar values: If you have similar values in a column, clustering the values will sort similar values and suggest a new value to replace those cells with.

***Tip:*** *This works the best with values without a large amount of text in the cells.*

* + 1. Transforming
       1. If you have more than one datapoint in the facet that needs to be changed to a common value, select ‘include’ to select only those that need to be changed

**Tip:** If you know there is a particular word, phrase, or symbol in the cells that needs to be changed, you can search based on that using the ‘Text Filter’ option to find those data points. NOTE: This will only search one column at a time.

* + - 1. Select the column and go to ‘cell edit’ and then ‘transform.’ From here you can change the value of the selected cells in this column
    1. Clustering
  1. Column Editing
     1. Merge/Concat Columns
        1. Merging or concatenating columns is used to combine at least two columns to create a single

**Recipe:**

cells[‘NAME OF COLUMN’].value + ‘ ’ + cells[‘NAME OF 2nd COLUMN’].value + ‘ ’ + … + cells[‘NAME OF FINAL COLUMN’].value

***Tip:*** *When merging columns make sure to include spaces between data points to avoid having one long string of information*

* + 1. Splitting Columns
       1. If the cells in a column have data that should be separated into multiple columns there are two methods.
          1. Using the ‘Split column into several columns’ option in the dropdown menu. This will spilt the column into as many columns as needed based on the separator chosen to split on (space, pipe, comma, etc.)

***Tip:*** *If you want to keep the original column, make sure to uncheck the box that says, ‘Remove this column’. Keeping the original column is also helpful in case information is accidentally deleted or changed.*

* + - * 1. Using the ‘Add column based on this column’ option in the dropdown menu. This will allow you to create new columns and splits based on more specific separators (i.e., word, position of word in data point)

***Tip:*** *This method is best used later in the data standardisation process after you are certain on how data points should be split to avoid creating unnecessary columns and/or deleting data.*

**Overall Tips:**

* Focusing on one theme/content area of your data (i.e., locality, specimen remarks, etc.) at a time makes standardising the data easier.
* Keep track of the way data points are organised/phrased (i.e., capitalisation of certain words, abbreviations used)
* Keep track of any discrepancies/issues noticed when cleaning data and how they were resolved - there will be many and you will forget

Data Enhancement

Types

Adding columns from other projects

Adding columns based on other columns