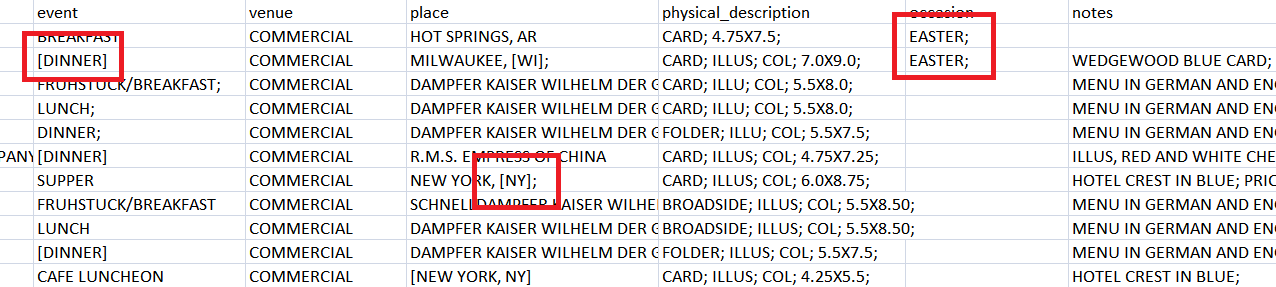
**DATA 201 – Assignment 3**

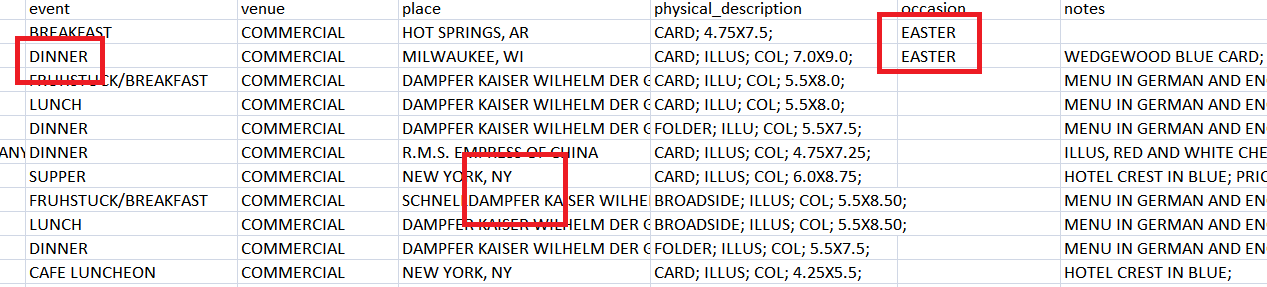
Alex Stevenson – 30073617

**Cleaning 1: Unnecessary Punctuation Marks**

Before:

****

After:



**Data Quality Issue:** Data Integration Error

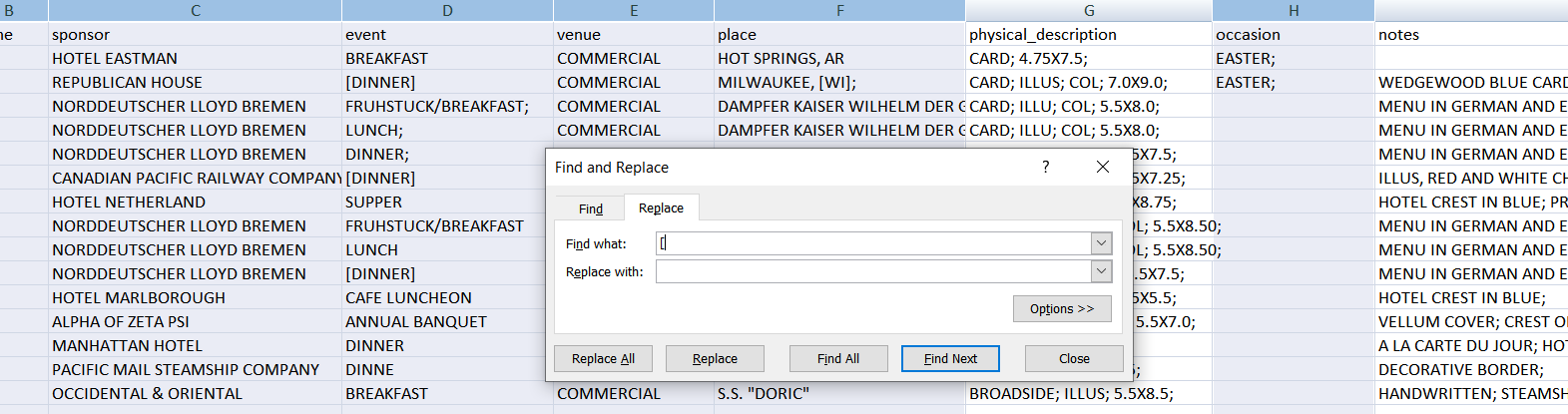
It looks like some sources of data use semicolons to mark the end of lines, and square brackets to highlight certain elements. There are many semicolons at the end of certain points of data that seem to have come from the same source, while various data points are surrounded by square brackets that could denote a differing source of data. An example of this would be the difference between “[DINNER], DINNER;, and LUNCH”

**Why these entities need to be cleaned:**

These entities should be cleaned as their presence makes it more difficult to read and compare data. As with the above example, different tuples contain the same information (“DINNER” and “[DINNER]”) should be formatted the same way. Cleaning these entities will make them clearer without erroneous punctuation marks, make them easier to group together as identical elements will be formatted the same way, and will be easier to analyze and compare different rows that have the same elements.

**Steps to Clean:** (Excel)

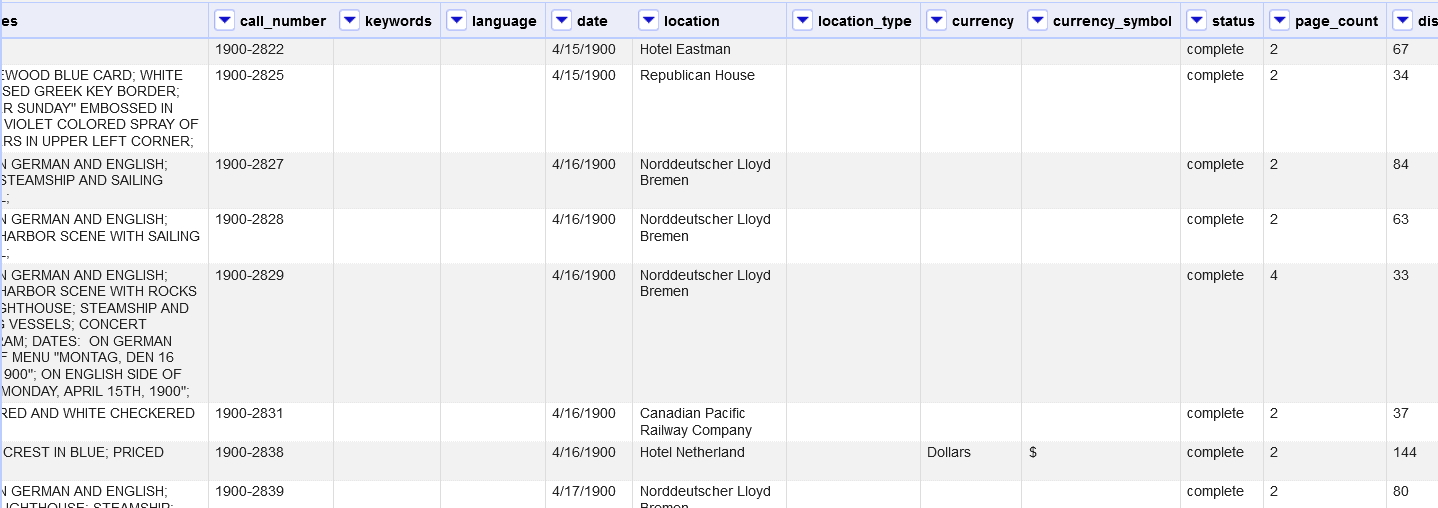
1. Select each column not including physical\_description and notes.
2. Ctrl-H to open the Replace menu, find “[“ and replace it with nothing



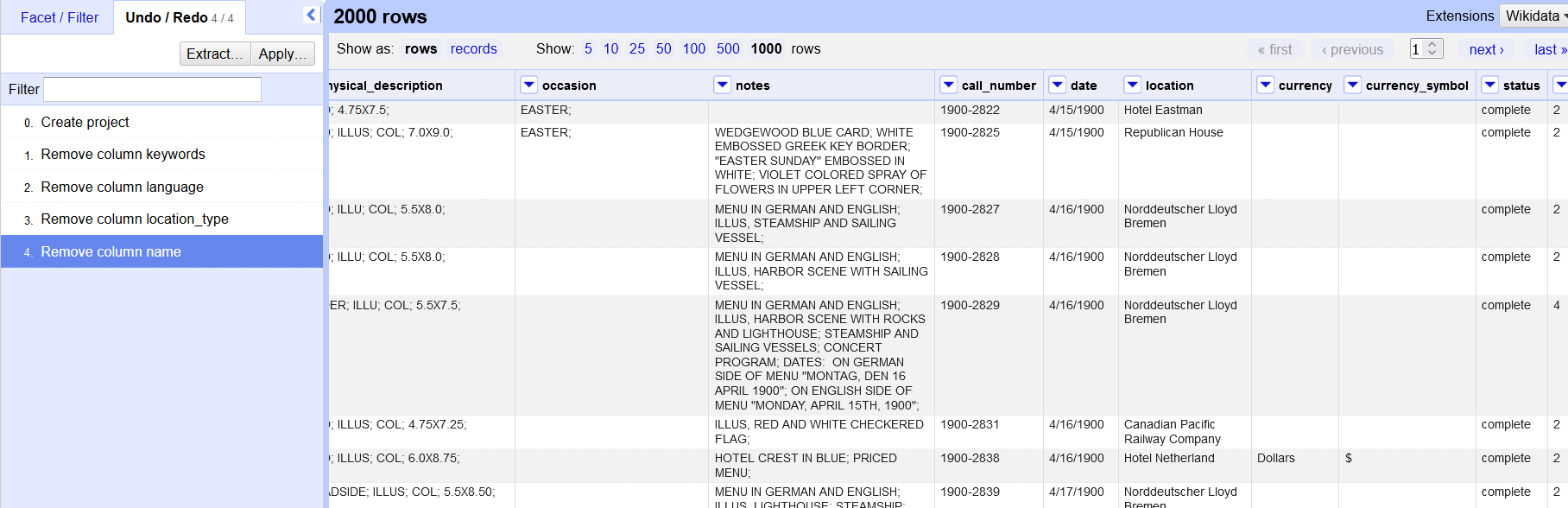
1. Do the same with “]” and “;”

**Cleaning 2: Empty Columns**

Before:



After:



**Data Quality Issue:** Missing Data

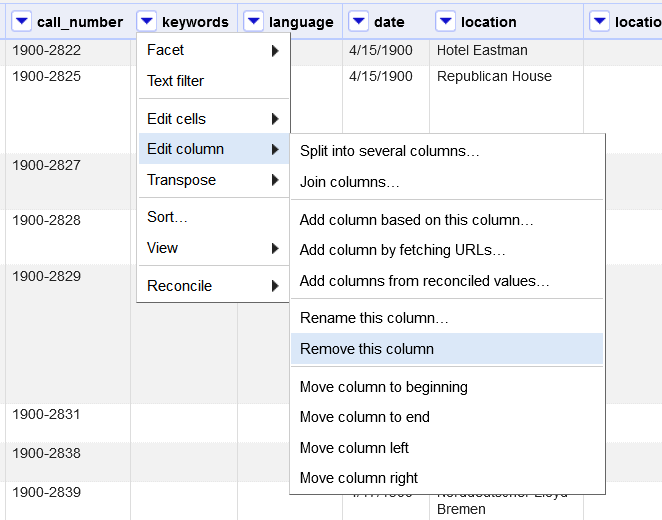
It looks like the data set was supposed to contain the attributes name, keywords, language, and location\_type. At some point this data was lost or removed after being collected as zero entries have any of these fields populated.

**Why these entities need to be cleaned:**

These columns are unnecessary and clutter up the data with random empty attributes, making it more difficult to read. Particularly as there are so many other fields that could be important to a user. Additionally, if you are taking a look at a small amount of rows at once, the existence of these columns implies that other rows will have that data available and that the current rows are not complete.

**Steps to Clean:** (OpenRefine)

1. Click the name of an unnecessary column (name, keywords, language, and location\_type)
2. Click Edit Column
3. Click Remove This Column

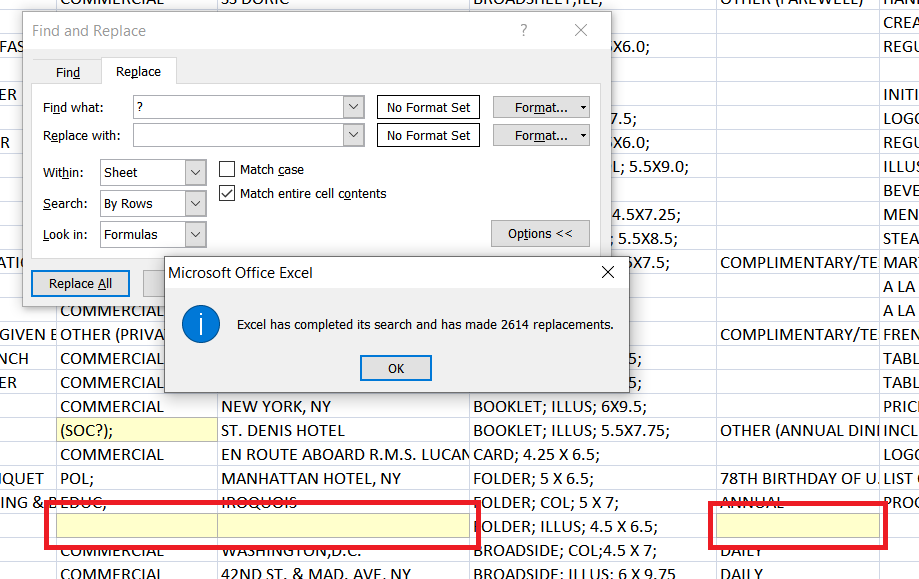


**Cleaning 3: Multiple Values for Unknown Fields**

Before:



After:



**Data Quality Issue:** Entity Resolution

Some unknown entries in the database are labelled with “?”, while others are left blank. You can also see that this issue is not consistent per row, as some rows have several question marks along with several empty attributes. This is an entity resolution error as there are multiple different entries that mean the same thing. I don’t think this is a data integration error as the error is not consistent within individual rows.

**Why these entities need to be cleaned:**

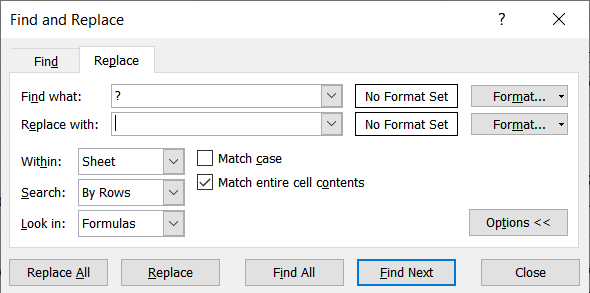
These entities consisting of "?" should simply be deleted as their presence adds nothing to the dataset but inconsistency. There are many blank fields that denote information that is not available, two different formats to explain the same thing is unnecessary and confusing.

Removal of these entities should help with analysis by reducing confusion, making the format of unknown fields consistent, and make it obvious when a field is unknown without a dubious question mark.

Note that in the screenshots above, the entry (SOC?); is highlighted in yellow. This is to make sure that not all question marks are removed, only the ones that fill the entire cell on their own.

**Steps to Clean:** (Excel)

1. Ctrl-H to open the Replace menu
2. Find “?” and replace it with nothing
3. Click Options to see the advanced options
4. Make sure that Match entire cell contents is checked, this will prevent the necessary question marks from being removed



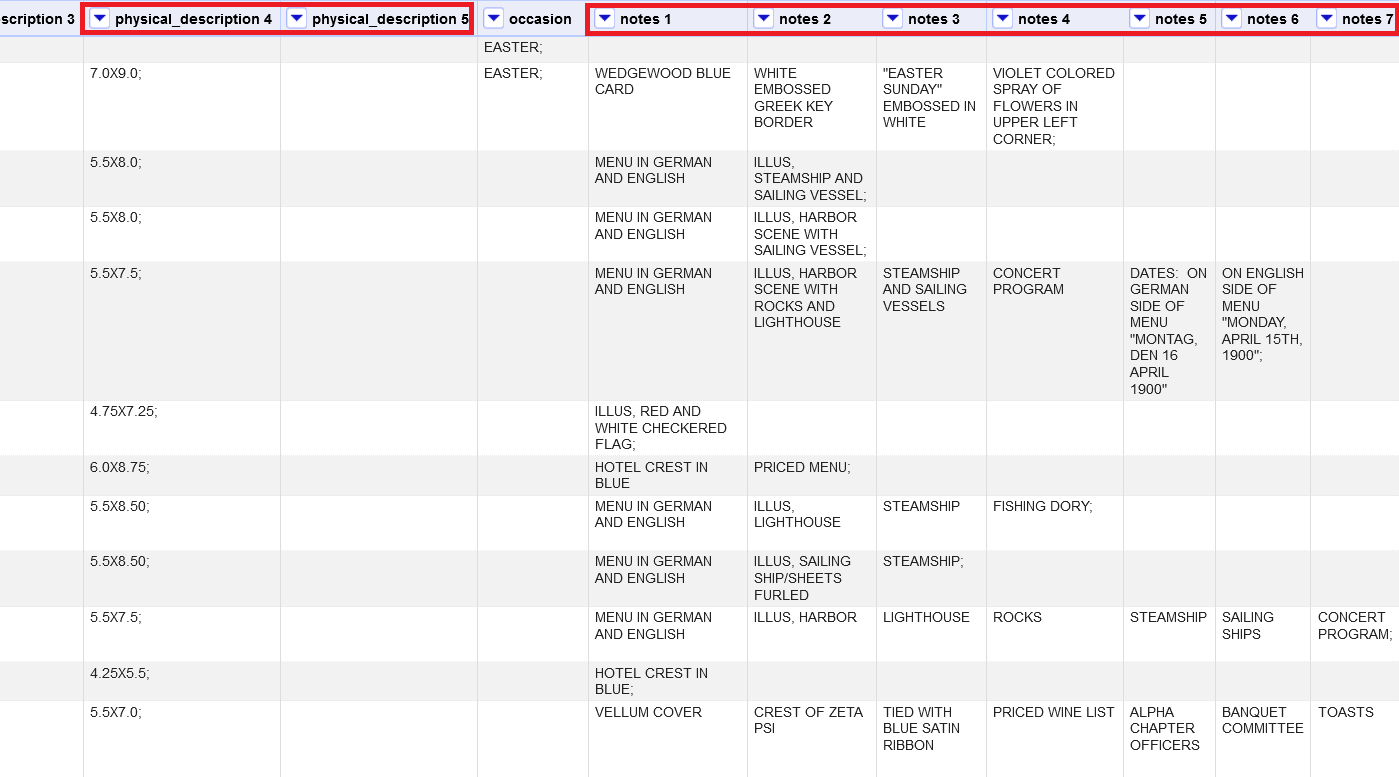
1. Click Replace All

**Cleaning 4: Breaking Big Attributes into Multiple Columns**

Before:



After:



**Data Quality Issue:** Entity Resolution

This is a entity resolution issue because the data has been collected, but is stored in a confusing and difficult to parse way. Both physical\_description and notes have several entries for a majority of different rows, however all of those entries are listed off in a impenetrable block of capitalized text.

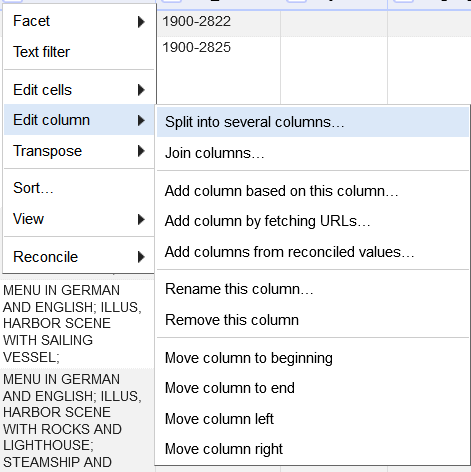
**Why these entities need to be cleaned:**

It should be useful to break these two fields into multiple columns. Both of them consistently have 2+ entries separated by semicolons, which makes it difficult to read and understand the data provided. By separating each entry into its own column, it will become clearer and easier to compare the different tuples.

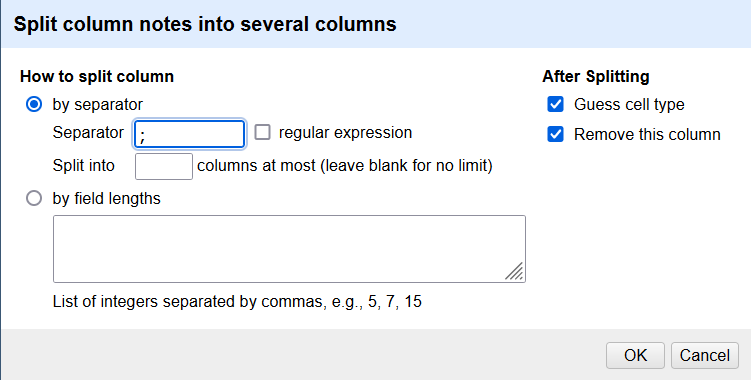
As it is now, users have to read the entire line of notes/physical descriptions in order to find specific ones that they are looking for as it is just a line of capitalized text. By breaking it into their own columns you are able to skim those fields and quickly find what you are looking for, while also being able to compare different rows more easily.

**Steps to Clean:** (OpenRefine)

1. Click the name of the column to edit, Edit Column, and Split into several columns



1. The separator should be “;” as that is what denotes the separation of list elements in the data



**Cleaning 5: Differing Date Formats**

|  |  |
| --- | --- |
| Before: | After: |

**Data Quality Issue:** Data Integration

It looks like there are two different date formats used in this data set. This implies either that the data has come from at least two different sources that use different date formats, or that the schema has changed over time to use a different date format.

**Why these entities need to be cleaned:**

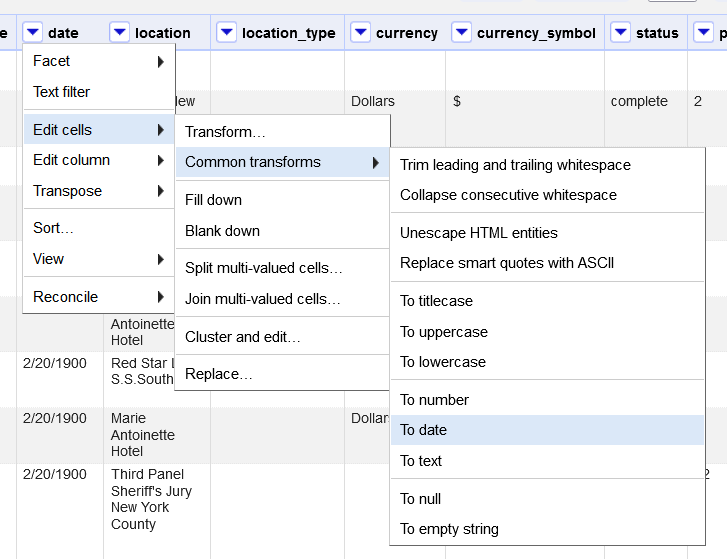
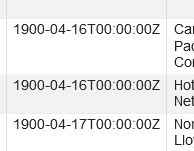
The current date formats can be confusing to readers when they look over multiple tuples with different formats. In the screenshots above we see two different date formats in use. This requires more work to parse the data as a human reader or a program, as they will need to account for both formats.

There are also potential uncertainties between the ordering of months and days in the data. An example of this is 9/5/2000, that could denote either May 9th or September 5th.

These different date formats should all be made consistent to prevent confusion and uncertainty as to what the data means.

Additionally, leading zeros should be added to the month and day sections of the dates, to make each date uniform and easier to compare with one another as the length of the dates will all be consistent

**Steps to Clean:** (OpenRefine)

1. Click the date column header, select Edit Cells, Common Transforms, To date, to transform the data in the column to a date object.  
   
2. Do the same as step 1, but transform the column into text.   
   This step is necessary as the Dates object seems to store additional text than what is displayed
3. Each date should now look like this:  
   
4. Click the date column header, select Edit Cells and Replace, replacing the string “T00:00:00Z” with nothing.  
   