

## OpenRefine - Instructions for Exercise

OpenRefine: <http://openrefine.org> and tutorials

Installation instructions: <http://openrefine.org/download.html>

Data Carpentry lesson on “OpenRefine for Ecology”:

<http://www.datacarpentry.org/OpenRefine-ecology-lesson/>

**Formatting issues:** Tables A, B, C (wide tables, each table contains a months worth of data). These tables all have the same format, except:

- Table A has the same column meanings as B and C, however the column names of A have different spelling.
- Table A doesn't contain the flag column.

### Solution:

1. Edit column names in text editor, so that names are consistent for all three tables.
2. Add column “flag” to Table A by means of adding name in line 1.
3. Bind Tables A, B and C row-wise into one Table: “Table\_1.csv”.
  - a. In OpenRefine “Create project” (Figure 1).
  - b. Import data: select three tables A,B,C from “This Computer”.
  - c. Edit “parsing options” (pressing on upper right hand corner). Enter project name (here “Table\_1”). Press “Create Project” (Figure 2).
  - d. Tables are automatically concatenated row-wise. Additional column is created with “tablename” from which data in row originated.

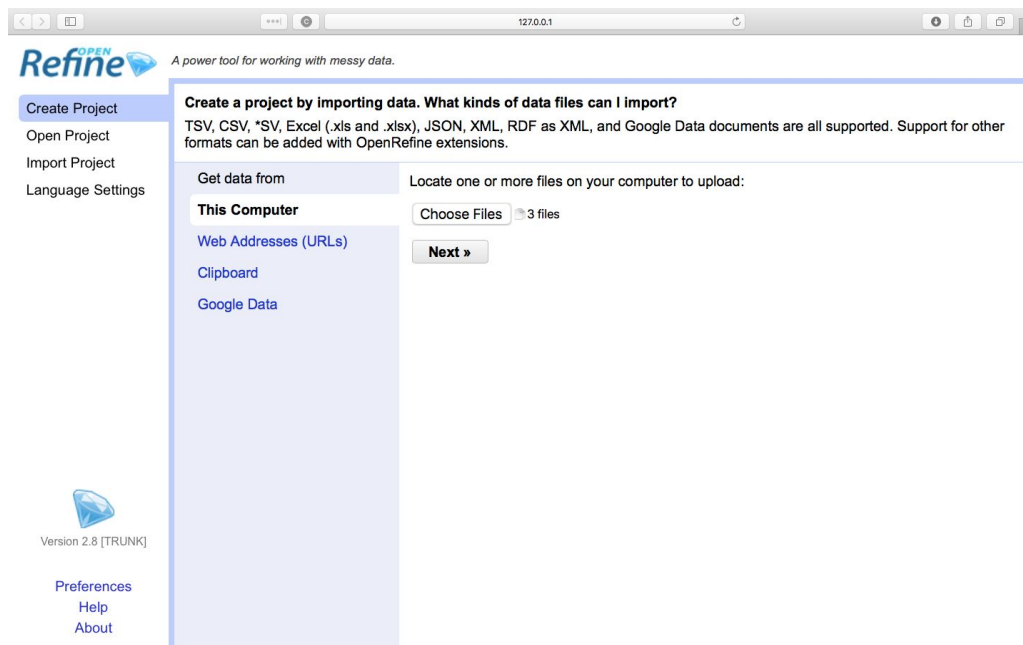


Figure 1: OpenRefine - ready to create project based on “Choose Files” from “This Computer” (here: Tables A,B,C), press “Next” (file upload occurs).

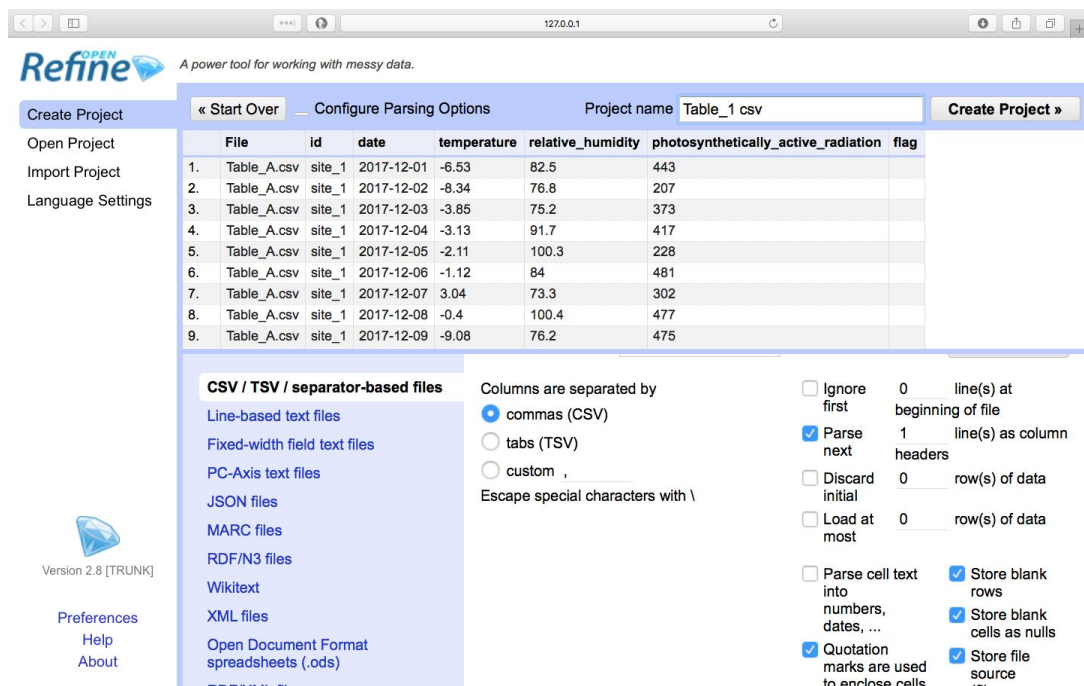


Figure 2: OpenRefine - edit parsing options (here: Tables A,B,C), give “Project name”, press “Create Project”.

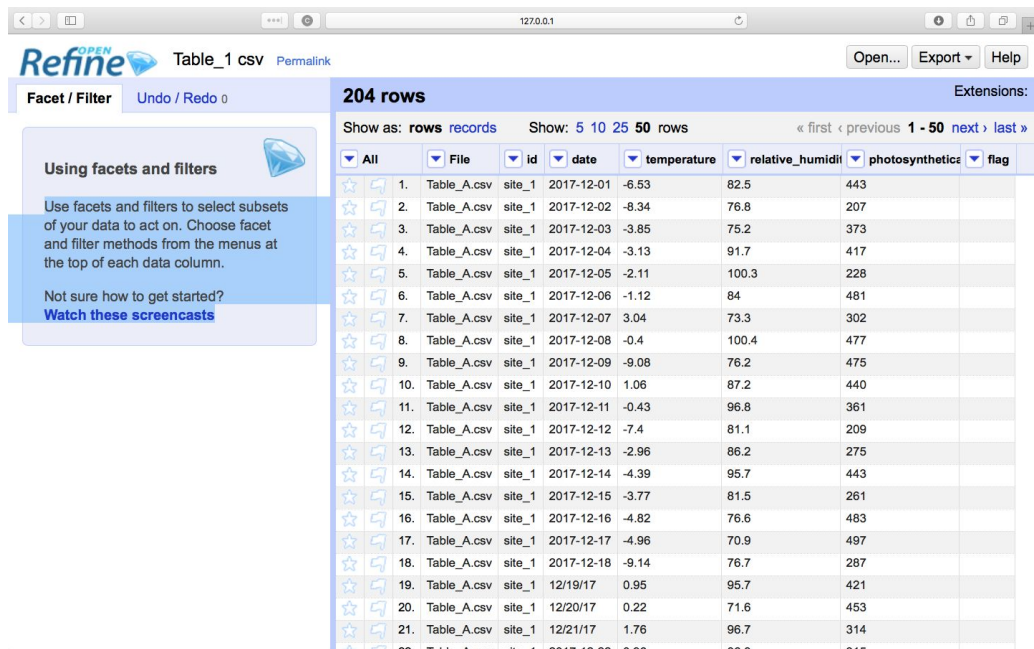


Figure 3: OpenRefine -Table\_1 concatenated from Tables A,B,C with additional column “File”.

**Column issues:** Table 1 contains a series of common data value problems:

- Replace blank cells with "NA".
- Identify and clear leading and prevailing "white spaces".
- Identify multiple missing value codes and replace with "NA".
- Identify characters in columns with numerical values.
- Identify and remove duplicate rows.
- Ensure correct precision of values.

**Solution:** Check values for **each individual column** by means of using the **facet/filter option** in left hand column of screen after choosing appropriate "facet" option in **column menu**.

1. Replace "blank cells" with "NA" using "Text facet" from column menus. Figures 4-6 show this process for column "flag". Repeat for all individual columns
2. Identify and clear leading and prevailing "white spaces": from each! column main menu choose "cells" -> "common transformations" -> "trim leading and prevailing white spaces" (Figure 7).
3. Identify multiple missing value codes and replace with "NA"; characters in cells with numerical values, For example column temperature shows missing values of "NaN, NA, -9999". The latter appears as outlier in numerical values.
  - a. Column menu "cells" -> "common transforms" -> "to number"
  - b. "Facet" -> "numeric facet". In **facet/filter option** in left hand column of screen data are classified into "numeric", "non-numeric", "blank", "error". Choose subset to edit (Figure 8) according to your standards. Attention: in column "photosynthetic radiation" missing data are characterized as "sensor fail" and "dog has eaten ...". Change to "NA" values and move explanation in flag column.
4. Use "Facet" -> "custom facet" to find matching rows.
5. Use "Facet" -> "text facet" on "Date" column. Edit 3 cells with format MM/DD/YYYY to YYYY-MM-DD (Figure 9). Other possibility is to convert column to format "Date" by means of choosing "cells" -> "common transformations" -> "to date" from main column menu.

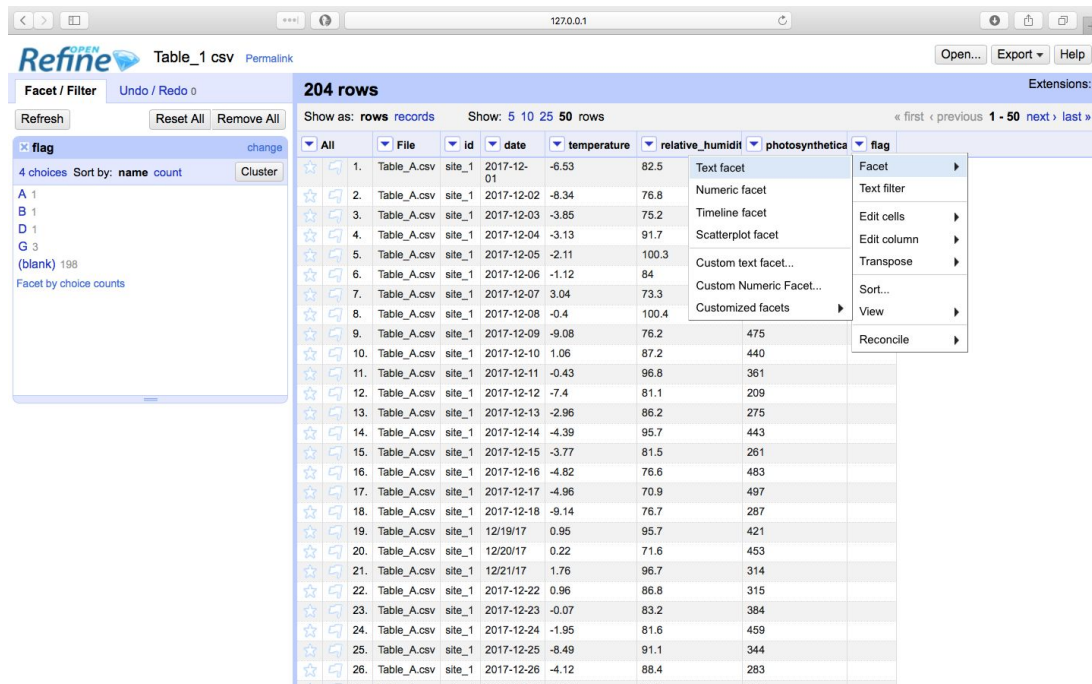


Figure 4: Choose from column “flag’s” menu “Facet” -> ”Text facet”. In left hand “Facet/Filter” window see the list of “flag” values. Hover over “blank” and click on “edit”.

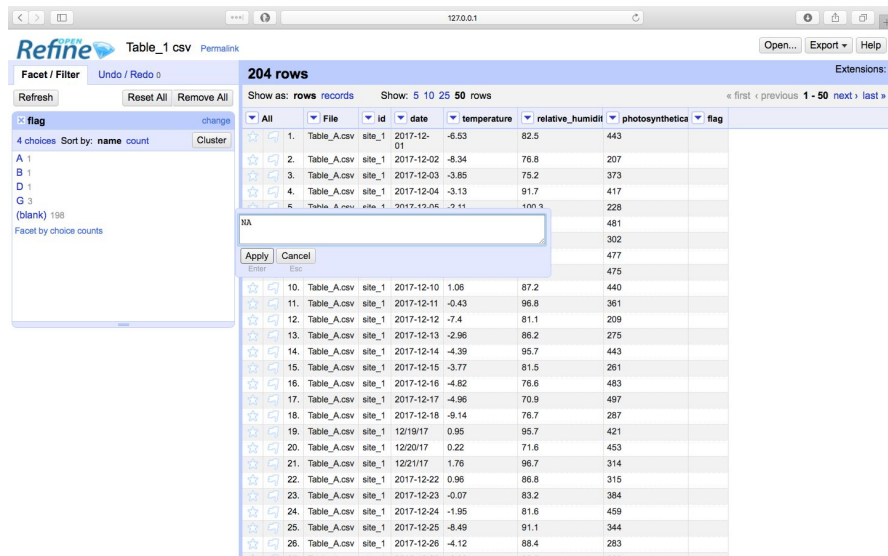


Figure 5: In edit form replace “(blank)” with “NA” and click “apply”. This changes all “flag” cells with the value “(blank)” to “NA”. See result of operation in Figure 6.

Refine Table\_1.csv Permalink

Facet / Filter Undo / Redo 1

Refresh Reset All Remove All

204 rows

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 50 next > last »

Extensions:

× flag change

5 choices Sort by: name count Cluster

A 1  
B 1  
D 1  
G 3  
NA 198

Facet by choice counts

	File	id	date	temperature	relative_humidity	photosyntheticity	flag
1.	Table_A.csv	site_1	2017-12-01	-6.53	82.5	443	NA
2.	Table_A.csv	site_1	2017-12-02	-8.34	76.8	207	NA
3.	Table_A.csv	site_1	2017-12-03	-3.85	75.2	373	NA
4.	Table_A.csv	site_1	2017-12-04	-3.13	91.7	417	NA
5.	Table_A.csv	site_1	2017-12-05	-2.11	100.3	228	NA
6.	Table_A.csv	site_1	2017-12-06	-1.12	84	481	NA
7.	Table_A.csv	site_1	2017-12-07	3.04	73.3	302	NA
8.	Table_A.csv	site_1	2017-12-08	-0.4	100.4	477	NA
9.	Table_A.csv	site_1	2017-12-09	-9.08	76.2	475	NA
10.	Table_A.csv	site_1	2017-12-10	1.06	87.2	440	NA
11.	Table_A.csv	site_1	2017-12-11	-0.43	96.8	361	NA
12.	Table_A.csv	site_1	2017-12-12	-7.4	81.1	209	NA
13.	Table_A.csv	site_1	2017-12-13	-2.96	86.2	275	NA
14.	Table_A.csv	site_1	2017-12-14	-4.39	95.7	443	NA
15.	Table_A.csv	site_1	2017-12-15	-3.77	81.5	261	NA
16.	Table_A.csv	site_1	2017-12-16	-4.82	76.6	483	NA
17.	Table_A.csv	site_1	2017-12-17	-4.96	70.9	497	NA
18.	Table_A.csv	site_1	2017-12-18	-9.14	76.7	287	NA
19.	Table_A.csv	site_1	12/19/17	0.95	95.7	421	NA
20.	Table_A.csv	site_1	12/20/17	0.22	71.6	453	NA
21.	Table_A.csv	site_1	12/21/17	1.76	96.7	314	NA
22.	Table_A.csv	site_1	2017-12-22	0.96	86.8	315	NA
23.	Table_A.csv	site_1	2017-12-23	-0.07	83.2	384	NA
24.	Table_A.csv	site_1	2017-12-24	-1.95	81.6	459	NA
25.	Table_A.csv	site_1	2017-12-25	-8.49	91.1	344	NA
26.	Table_A.csv	site_1	2017-12-26	-4.12	88.4	283	NA
27.	Table_A.csv	site_1	2017-12-27	-6.33	86.2	228	NA

Figure 6: Results of text facet operations on blank values of column “flags”.

Refine Table\_1.csv Permalink

Facet / Filter Undo / Redo 1

Refresh Reset All Remove All

204 rows

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 50 next > last »

Extensions:

× flag change

5 choices Sort by: name count Cluster

A 1  
B 1  
D 1  
G 3  
NA 198

Facet by choice counts

	File	id	date	temperature	relative_humidity	photosyntheticity	flag
1.	Table_A.csv	site_1	2017-12-01	-6.53	82.5	443	NA
2.	Table_A.csv	site_1	2017-12-02	-8.34	76.8	207	NA
3.	Table_A.csv	site_1	2017-12-03	-3.85	75.2	373	NA
4.	Table_A.csv	site_1	2017-12-04	-3.13	91.7	417	NA
5.	Table_A.csv	site_1	2017-12-05	-2.11	100.3	228	NA
6.	Table_A.csv	site_1	2017-12-06	-1.12	84	481	NA
7.	Table_A.csv	site_1	2017-12-07	3.04	73.3	302	NA
8.	Table_A.csv	site_1	2017-12-08	-0.4	100.4	477	NA
9.	Table_A.csv	site_1	2017-12-09	-9.08	76.2	475	NA
10.	Table_A.csv	site_1	2017-12-10	1.06	87.2	440	NA
11.	Table_A.csv	site_1	2017-12-11	-0.43	96.8	361	NA
12.	Table_A.csv	site_1	2017-12-12	-7.4	81.1	209	NA
13.	Table_A.csv	site_1	2017-12-13	-2.96	86.2	275	NA
14.	Table_A.csv	site_1	2017-12-14	-4.39	95.7	443	NA
15.	Table_A.csv	site_1	2017-12-15	-3.77	81.5	261	NA
16.	Table_A.csv	site_1	2017-12-16	-4.82	76.6	483	NA
17.	Table_A.csv	site_1	2017-12-17	-4.96	70.9	497	NA
18.	Table_A.csv	site_1	2017-12-18	-9.14	76.7	287	NA
19.	Table_A.csv	site_1	12/19/17	0.95	95.7	421	NA
20.	Table_A.csv	site_1	12/20/17	0.22	71.6	453	NA
21.	Table_A.csv	site_1	12/21/17	1.76	96.7	314	NA
22.	Table_A.csv	site_1	2017-12-22	0.96	86.8	315	NA
23.	Table_A.csv	site_1	2017-12-23	-0.07	83.2	384	NA
24.	Table_A.csv	site_1	2017-12-24	-1.95	81.6	459	NA
25.	Table_A.csv	site_1	2017-12-25	-8.49	91.1	344	NA
26.	Table_A.csv	site_1	2017-12-26	-4.12	88.4	283	NA
27.	Table_A.csv	site_1	2017-12-27	-6.33	86.2	228	NA

Figure 7: “Trim leading and prevailing white spaces” from each column using “cells” from main column menu, then “Common Transforms”.



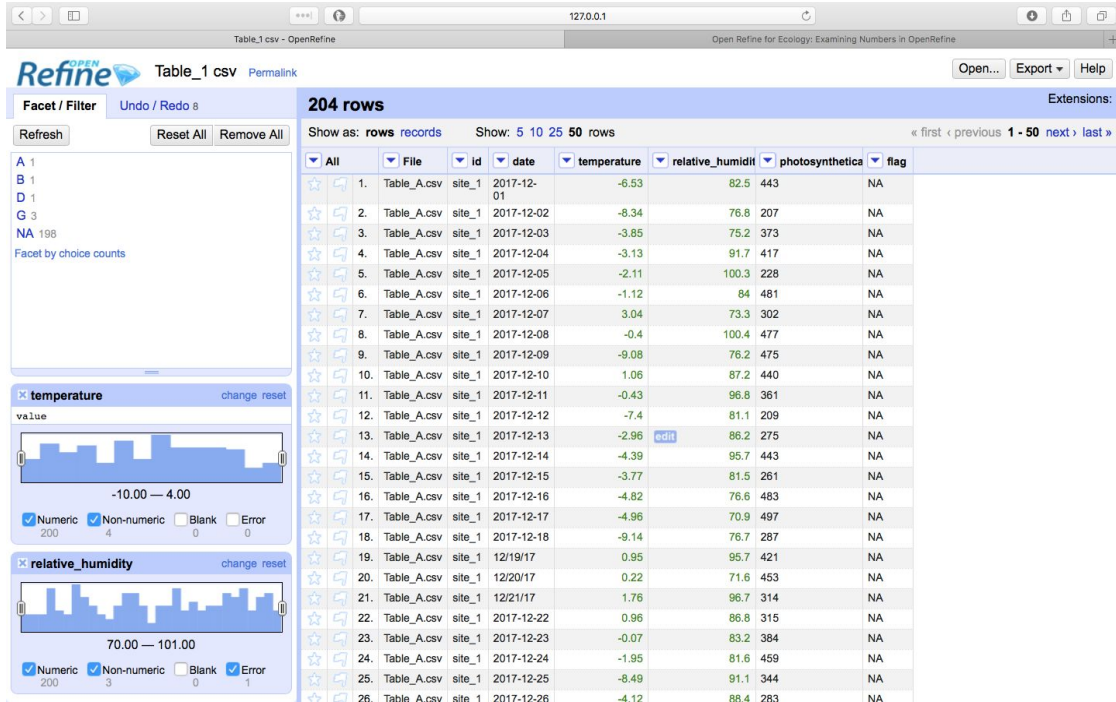


Figure 8: Choose “Facet” -> “Numeric” from each column menu. Analyse display on left hand side column of screen “Facet/Filter”. Select and edit “Non-numeric”, “Blank” and “Error” values.

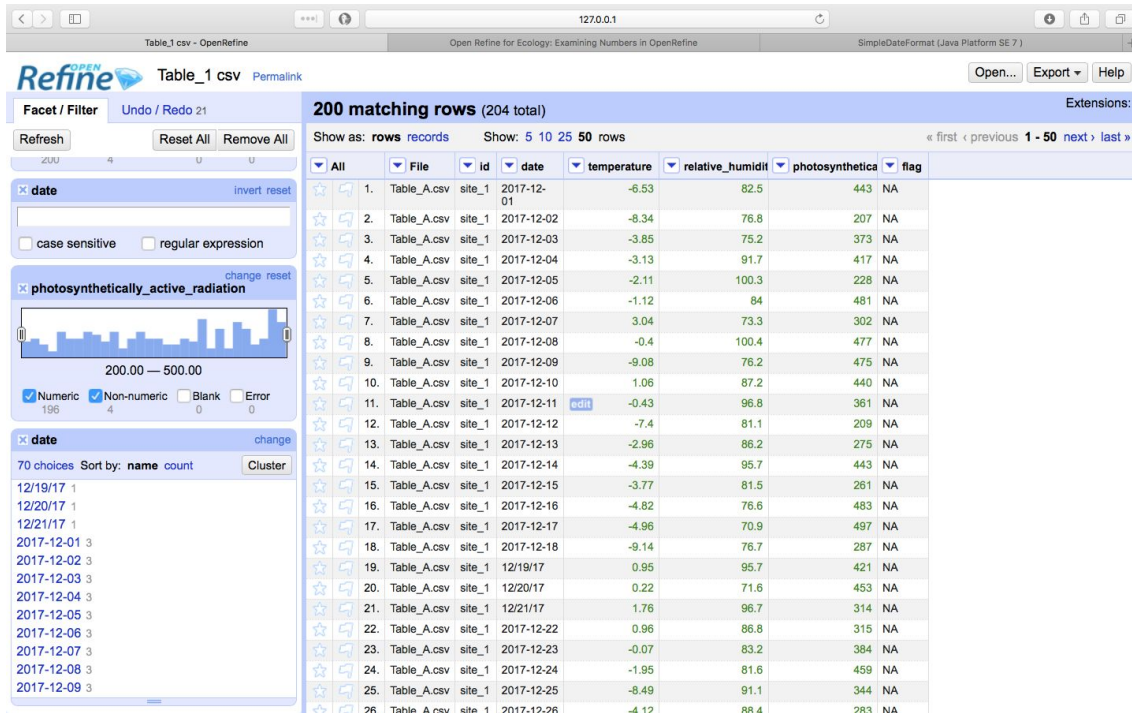


Figure 9: To unify “Date” format: Choose “Facet” -> “text” from “date” column menu. Analyse display on left hand side column of screen “Facet/Filter”. Modify cells with outlier format.