Tutorial: OpenRefine

By

Atima

Han Zhuang

Ishita Vedvyas

Rishikesh Dole

Tutorial: OpenRefine

INDEX

Tutorial: OpenRefine

1. INTRODUCTION

Openrefine is a data manipulation tool which cleans, reshapes and intelligently edit batch messy, and unstructured data. It is an open source tool and its code can be reused in other projects too. Openrefine offers many features like faceting, clustering, editing cells, reconciling, extending web services, which helps to clean and transform data effectively. Openrefine is easy as excel and powerful like access database. It makes many common tasks easy to do. It helps to analyse the data through filtering, faceting and converts the data into a more structured format.

2. BIG IDEA

The big idea behind choosing OpenRefine as our tool is to provide a tutorial by which users can have a free and an open source tool to manipulate their data sets. OpenRefine provides the flexibility to choose from a variety of data set functionalities, which makes it even more user friendly. Users can use this tool to get a big view of their data in terms of statistically curved graphs. They can play with messy data without worrying about risks, since they can undo their activity at any time. Cleaning, transforming and fetching URLs for a dataset can be easily done by simply having the application downloaded in the system.

What –A messy, unstructured, inconsistent dataset can be explored using open refine. In general, it will be very difficult to explore data through redundancies and inconsistencies. But, OpenRefine gives several functions through which one can filter the data, edit the inconsistencies, and view the data. It's a tool to clean the data.

Why- Spreadsheets can also refine a dataset but they are not the best tool for it as Openrefine cleans data in a more systematic controlled manner. While using historical data, we come across issues like blank fields, duplicate records, inconsistent formats and using Openrefine tool can help to resolve such issues.

When-Now data analysis play an important role in business. Data analysts improve decision making, cut costs and identify new business opportunities. Analysis of data is a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. So, to ensure the accuracy of our analysis, we have to clean our data

Tutorial: OpenRefine

3. Why OpenRefine is a better tool?

Google refine	Spreadsheets	Databases
1. Batch editing of rows and columns possible 2. Used for exploring and transforming data. 3. No Schema Required 4. Data is always visible at each step of editing. 5. More interactive and visual.	1. Editing of one cell at a time. 2. Used for entering data and performing calculations, functions. 3. No Schema Required 4. Data is always visible 5. Visual is not impressive.	1. Schema and programming language required for editing. 2. Data is out of sight unless script is run to view it.

4. BACKGROUND

- Google Refine finds its root in the Freebase Gridworks solution developed by Metaweb Technologies, Inc. in May 2010.
- Initially it was a tool designed to support the Freebase database and community for data cleaning, reconciliation and upload.
- This historical link with Freebase is still present in Google Refine, as the solution supports reconciliation against Freebase database.
- In July 2010 Google acquired Metaweb and by extension, Freebase and Gridworks. Freebase Gridworks has been renamed Google Refine and the code and documentation moved to a code.google.com instance.
- The freshly renamed Google Refine continued to be an open source project for data cleaning.

5. DATASET

There are various types of datasets used .The datasets used are either downloaded from the internet or prepared on our own to suit the functions and the situations.

Retrieved from "http://www.briandunning.com/"[1]

Following are some of the examples of sample sets used.

Tutorial: OpenRefine

1)

us-500.xlsx

2)

Facet.xlsx

3)

ExtendingData_dataset.xlsx

6. KEY FEATURES

There are many features in OpenRefine. We have focussed on the most used and important features of OpenRefine. They are listed as follows:

a) Importing Data

b) Filtering/ Faceting

c) Editing cells/columns

d) Reconciliation

e) Exporting Data

f) Undo/ Redo

A brief explanation of the features

a) Importing Data: - The importing data is used to get the data from various external sources. It comprises of two parts; namely Creation of Project and Parsing Data.

Creation of Project

Tutorial: OpenRefine

A) Navigation:-

OpenRefine→Click on 'Create project'→ Select from the list where you want to get the data→Click on Choose File→Next→Create Project (after checking parsing for the data)

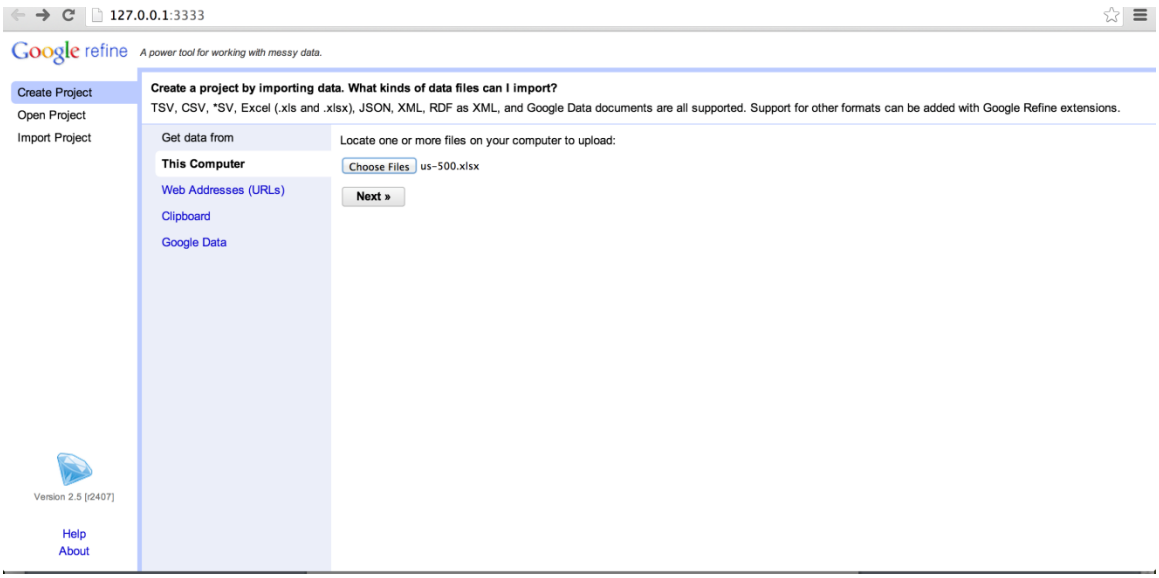
B) Steps

1) Open Google Refine

2) Click on Choose File to browse through your documents and to select the particular document to play with.

3) You can even choose a website, or from your clipboard, or even google data.

4) We'll be showing an example through a file in the computer.



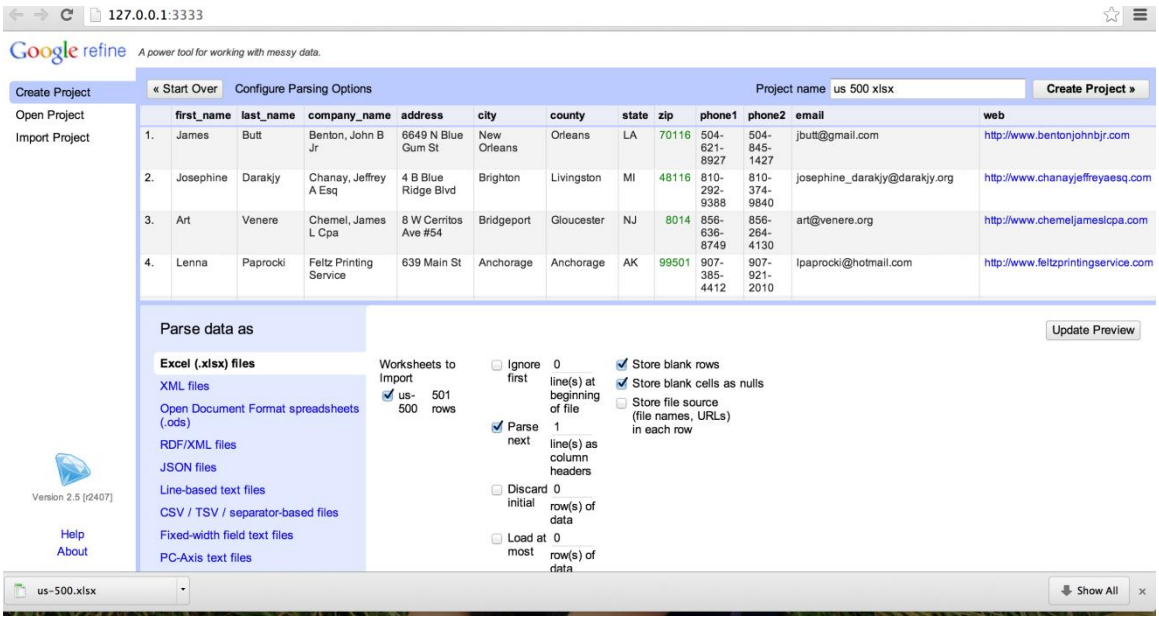
We have selected a file Refine_Excel from our computer. It is in the xlsx format.
Note: - File formats supported by Open refine includes TSV, CSV, *SV, .xls, .xlsx, JSON, XML, RDF as XML and google documents.

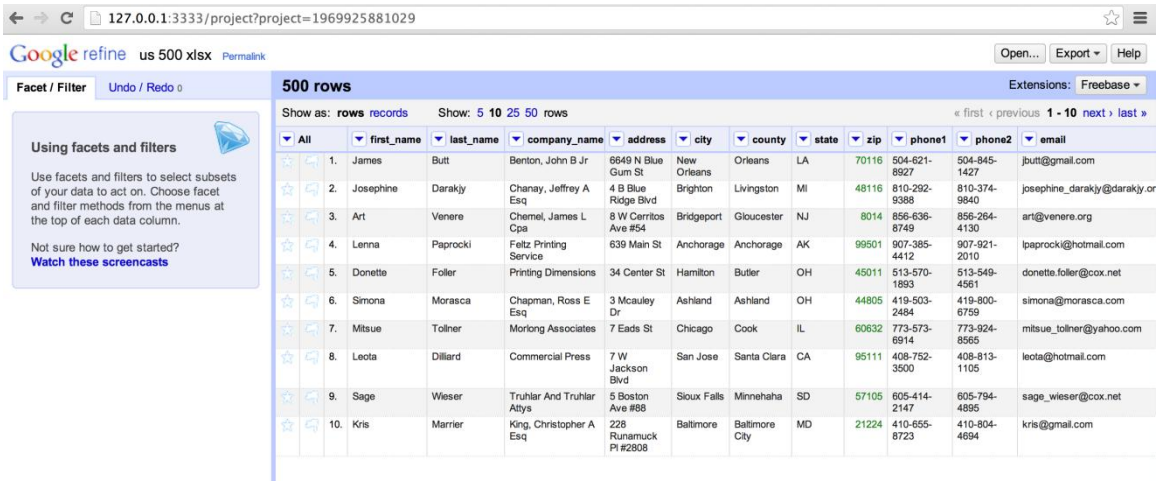
5) Click on Next after selecting the file. This begins the uploading process.
6) The file is uploaded.
7) At this step, give a name to the project, and click on create project. You can even open an existing project, or import it from somewhere. We have given the name "Sample" to the project. This begins the project creation.

<u>Parsing Data</u>

As it is shown in the image below, the bottom part displays the details of the document for parsing the Data such as, the number of rows, etc.

Tutorial: OpenRefine





b) Filtering /Faceting Data: - It is a method to filter data into subsets for ease of use. It can be done for text, number and dates.

Types of Facets

1. Text: - This facet filters the same set of data in groups which helps to easily edit the data in groups. It shows number of rows for each group and gives a larger picture of data. Text facets can be applied on several columns.
2. Numeric: - This facet groups numbers into numeric range bins. Then we can select any range for use showing consecutive numbers.
3. Custom Text Facet: - This is a text facet in which you can split the column data using expression (value.split(“”) [0]) without creating new column. Groups will be made according to spilt data sorted by their counts.

4. Custom Numerical Facets: - This facet allows you to customize the numeric facets. The numeric values can be grouped by their logs, modulus, length of string etc.
5. Customized Facets: - There are various types of customizable facets. They include Word Facet, Duplicate Facets, Numeric log facet, 1- bounded numeric log facet, text length facet, Log of text length facet, Unicode char-code facet, Facet by error, Facet by Blank.

Example:-

Suppose we want to set a Filter data for the 'State' column, but we are finding certain discrepancies in the columns. So, we make use of the 'Facet' feature.

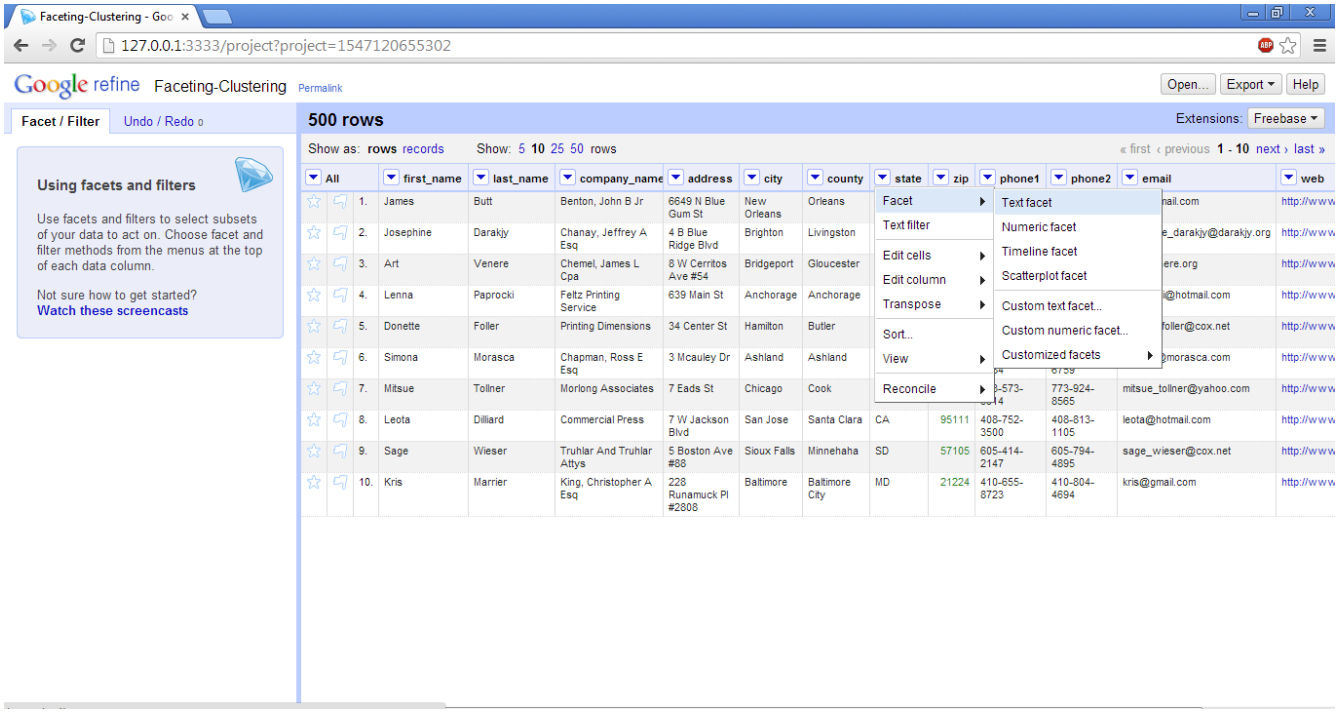
A) Navigation

Facet→Text Filter

B) Steps

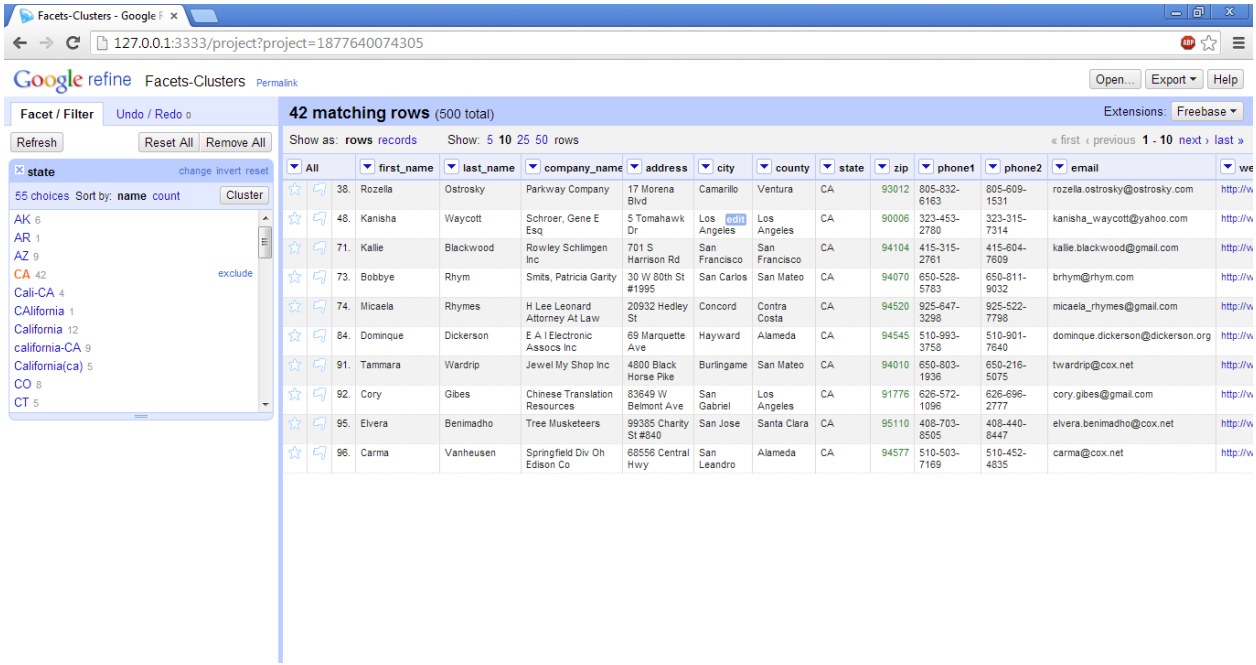
1) So Select Facet option in dropdown

2) Select text facet, it will club same items into one group and shows number of lines for each group.

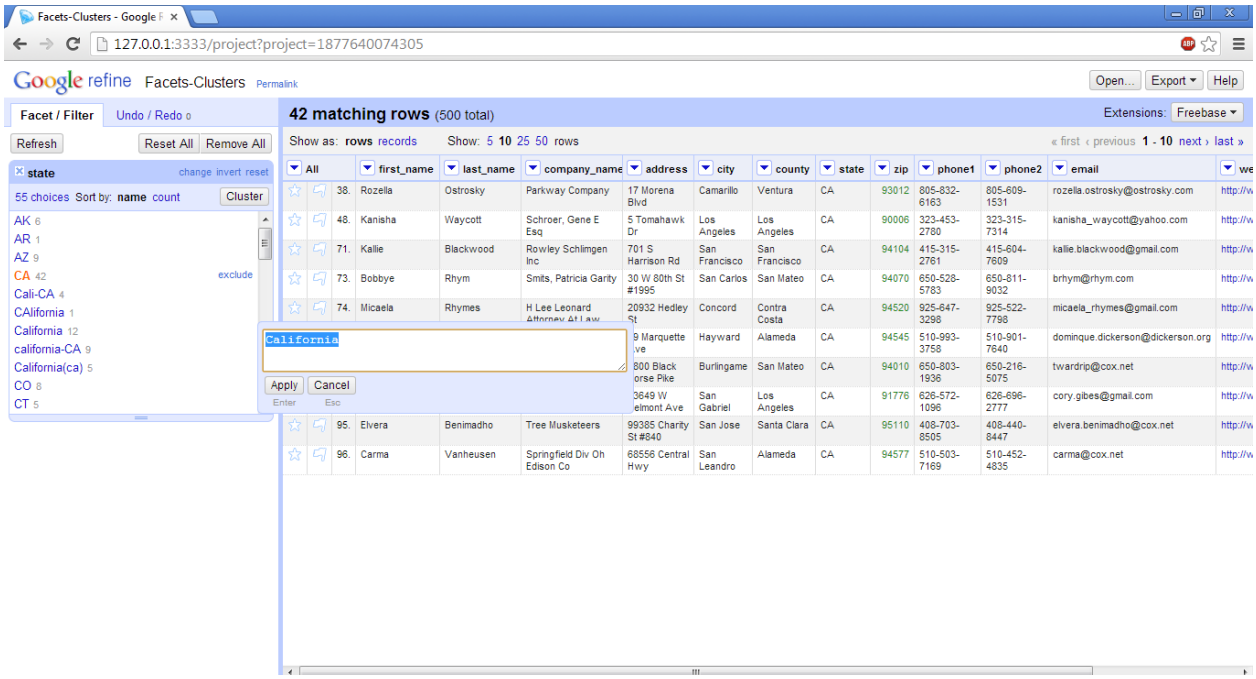


3) Applied text facet on "State" column and it grouped data in subsets of each state like below. Groups of state California, CA, California(ca) have been made separately which shows inconsistency in data.
4) We can club this data under one state by editing the group name and clustering.

Tutorial: OpenRefine

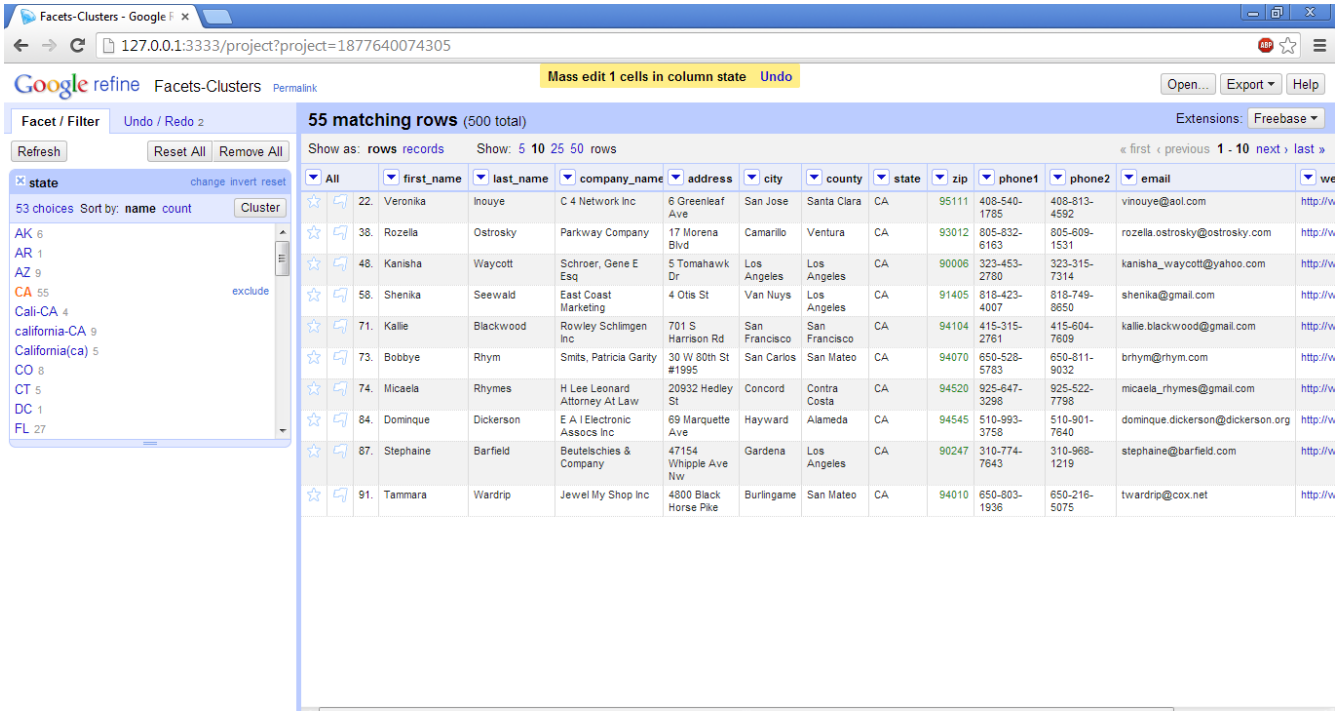


5) There are two groups CAlifornia and California and we can edit their names to CA which will merge them to group CA , for making the data consistent. Before merging CA group has 42 rows.

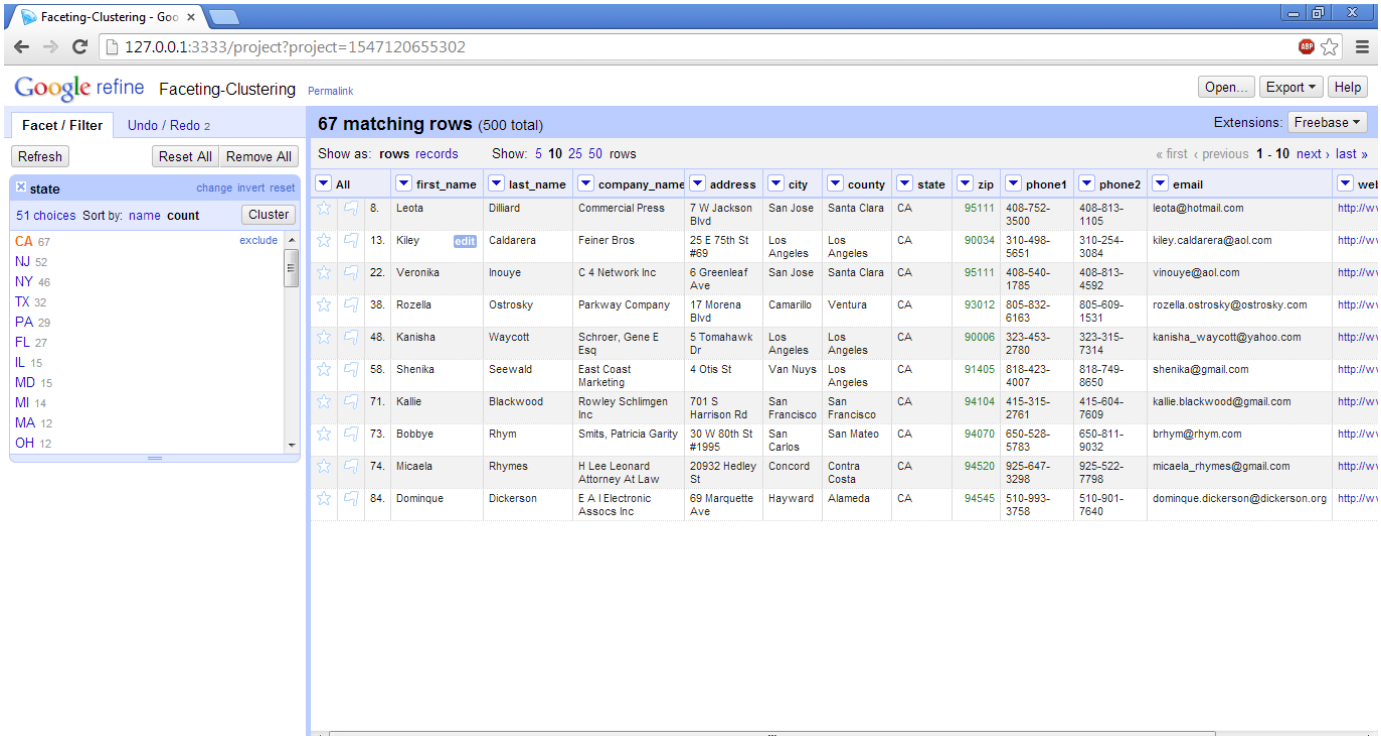


6) After editing the group names to CA, the total number of rows are now 55 as shown below.

Tutorial: OpenRefine

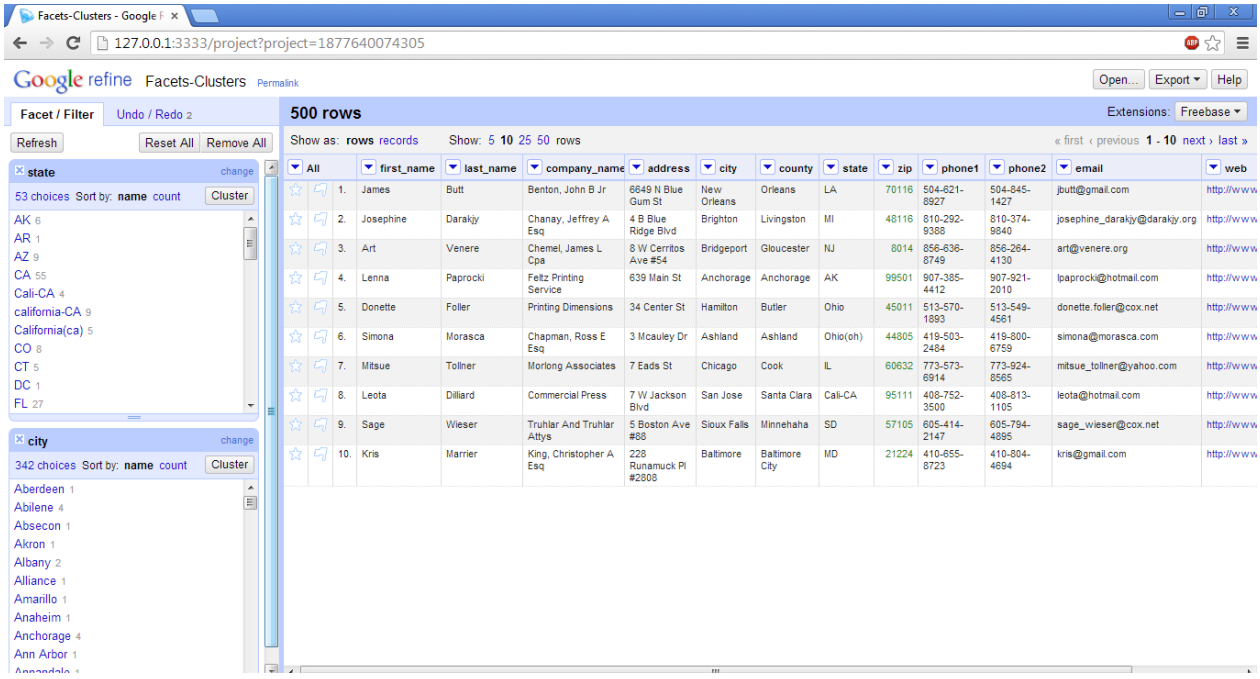


7) We can sort the groups by name, count to find the biggest groups as below.



8) You can create multiple facets. I created facet on second column “city” as below.

Tutorial: OpenRefine



9) Also, group, CA , California(ca) , Cali-CA, California-CA belong to same family but we can merge them together to make one group with clustering feature to make data more consistent. We can do this by clustering:

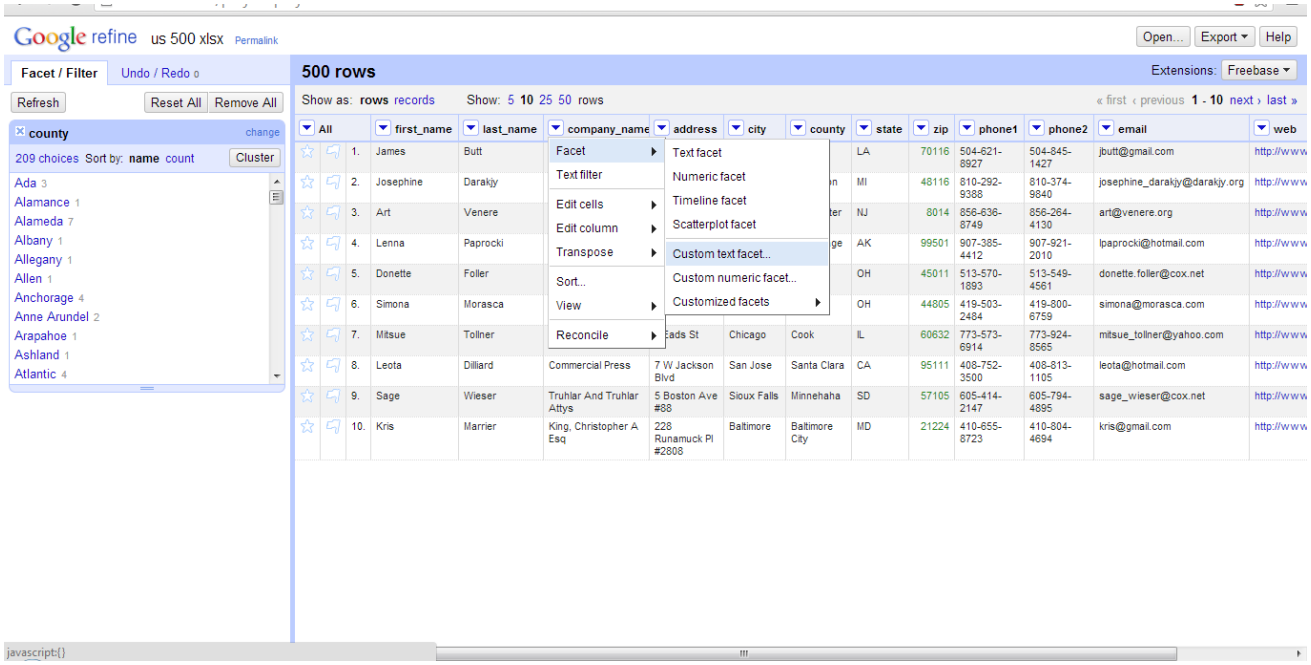
Custom text facet

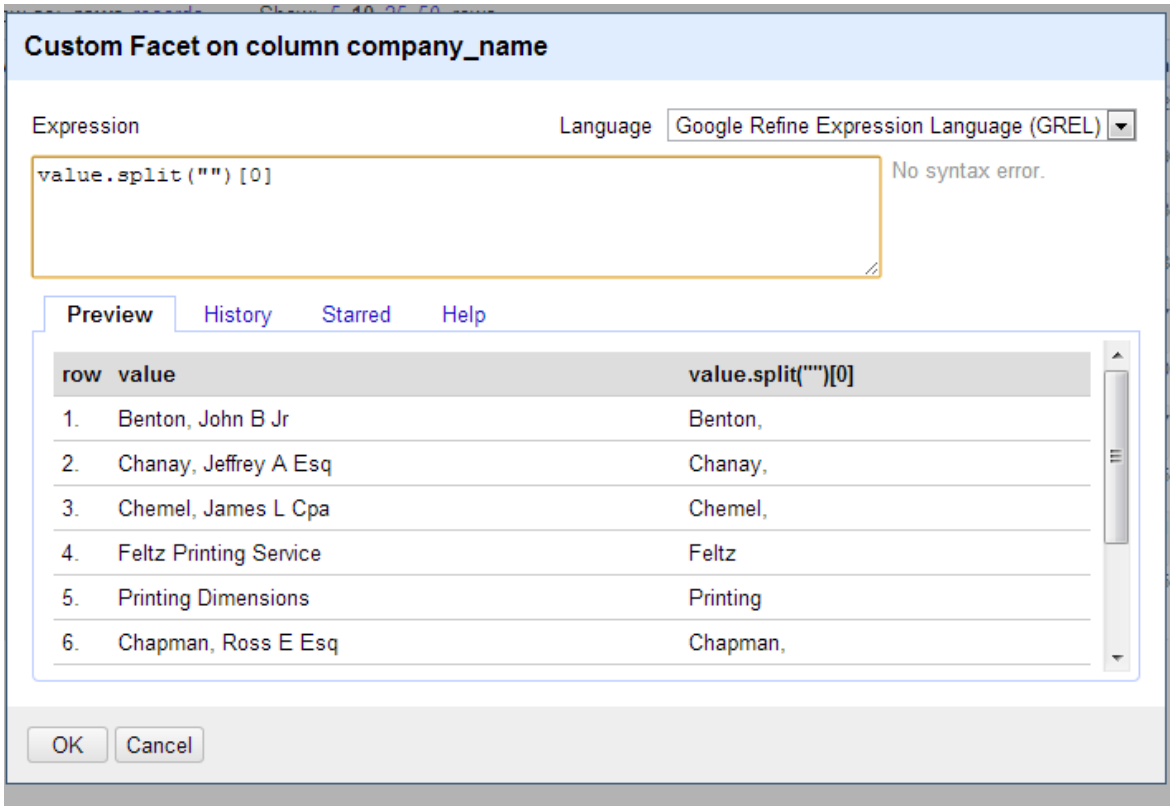
This is used when you want to edit a cell like extracting only first name. You need to put value

Value.split(“ “)[0]

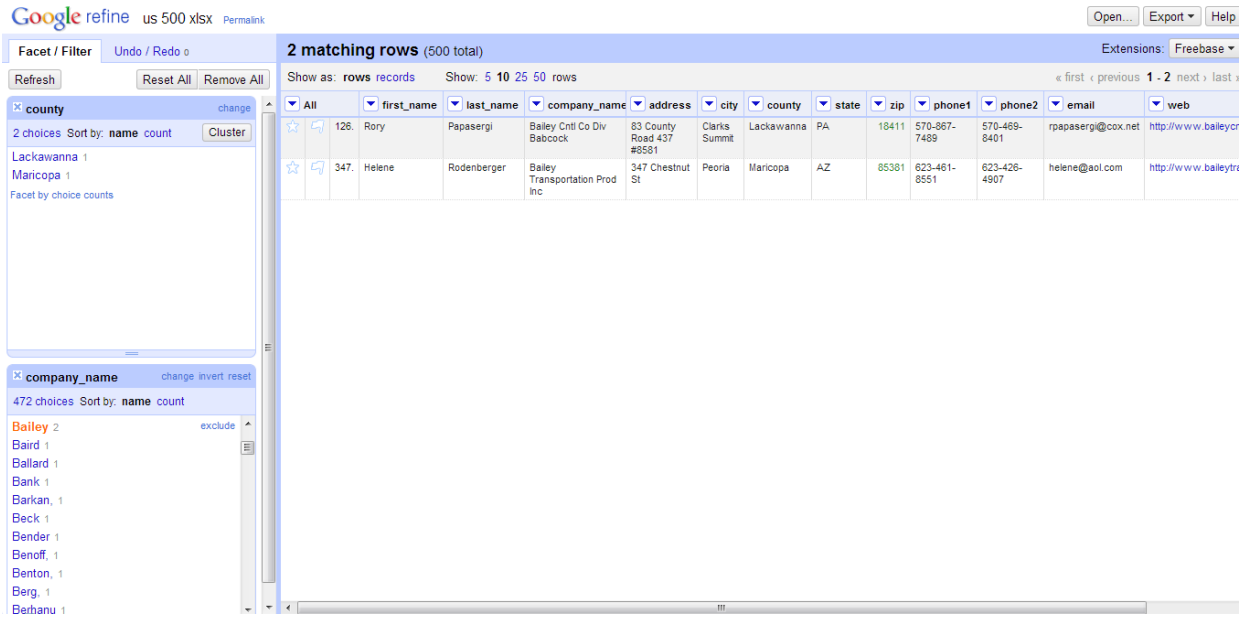
I applied custom text facet on “country_name” as below

Tutorial: OpenRefine

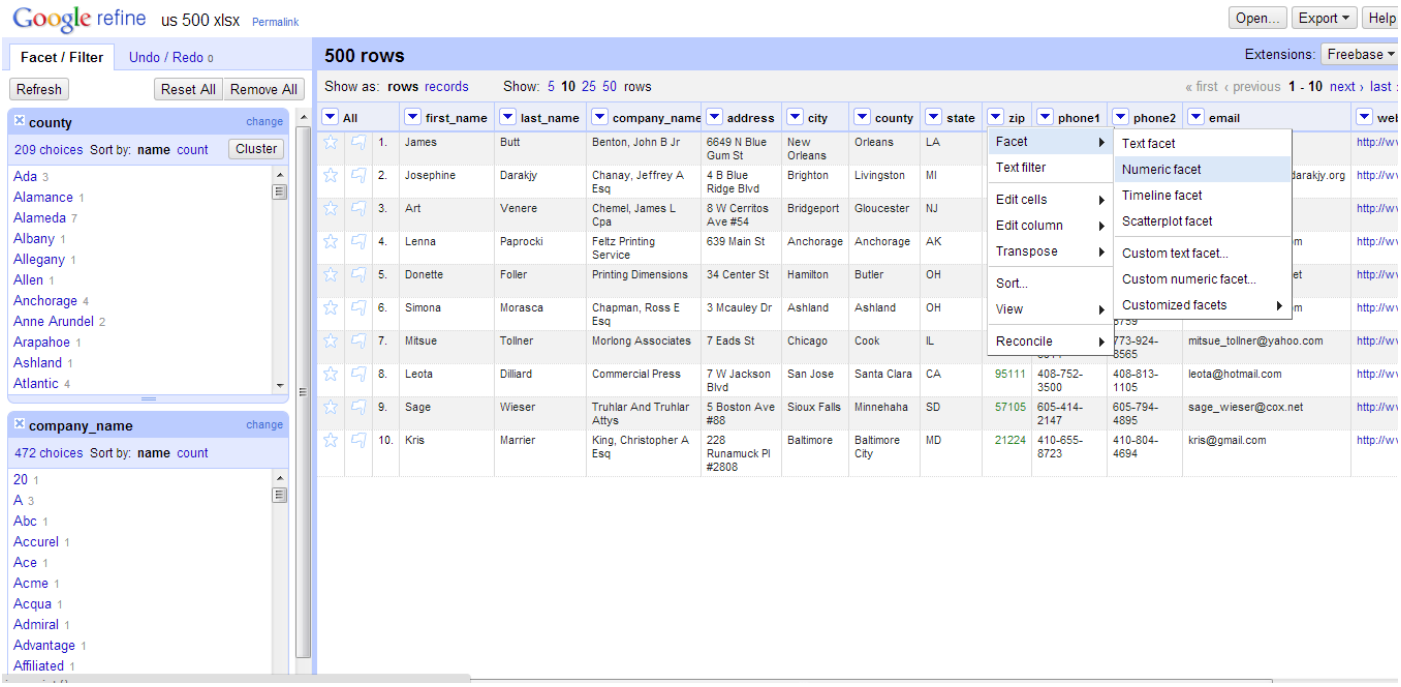




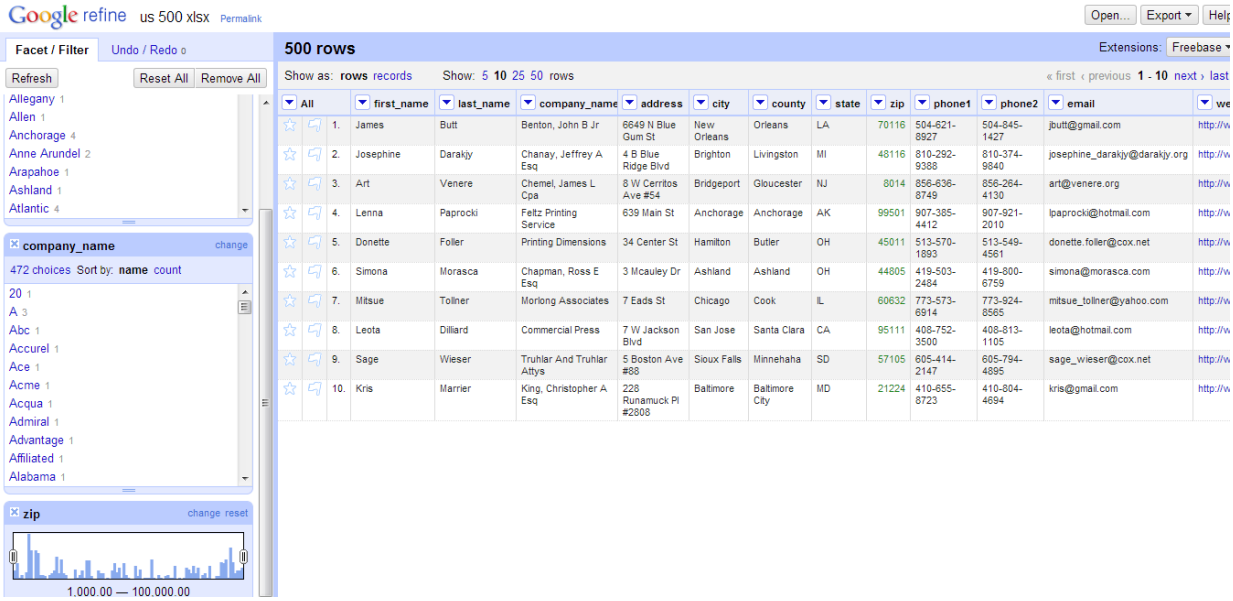
Tutorial: OpenRefine



Numeric facets to sort numbers which put numbers in numeric range bins.



Tutorial: OpenRefine



Tutorial: OpenRefine

c) Editing Cells/Columns/ Rows

(i) Editing cells by using common Transforms: - These are a few functions used to transform text cell values in a batch. The navigation for these functions is **Editing Cells→Common Transforms**. Some of them are for removing whitespaces, applying capitalization styles and to convert the data into the desired data type.

1) Trim leading and trailing whitespace: - This function is used to remove the whitespaces or the blank spaces in a word in the column

2) Collapse consecutive whitespace: - it is used to reduce the consecutive whitespace which occur back to back in a column.

3) Un-escape HTML identities: - There are certain entities which get attached while getting things from the web browser. So, to remove those, we can used to remove HTML identities.

4) to titlecase: - It is used to capitalize all the first alphabet of all the words.

5) to uppercase: - It is used to capitalize all the alphabets of the words.

6) to lowercase: - It is used to lower case all the alphabets of the words.

7) to number: - convert to number format

8) to date :- convert to date format

9) to text: - convert to text format

Transforming Data:

Suppose if we have uploaded text file for all students of UMD consisting of their courses, year, age etc. then we can extract year or any other data to make a new column for easy usage. New column can be made from existing columns

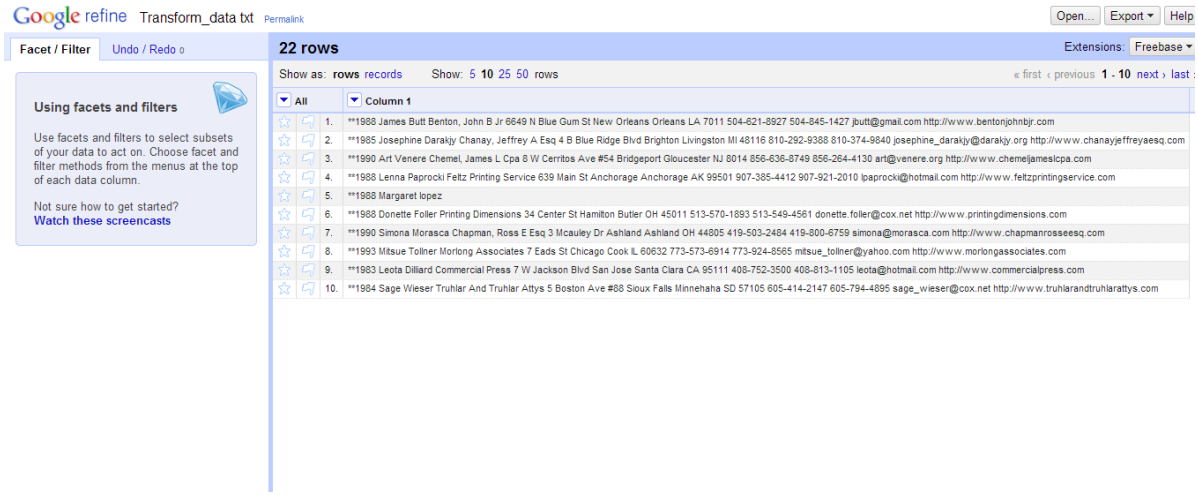
A) Navigation

Edit Cells→Transform

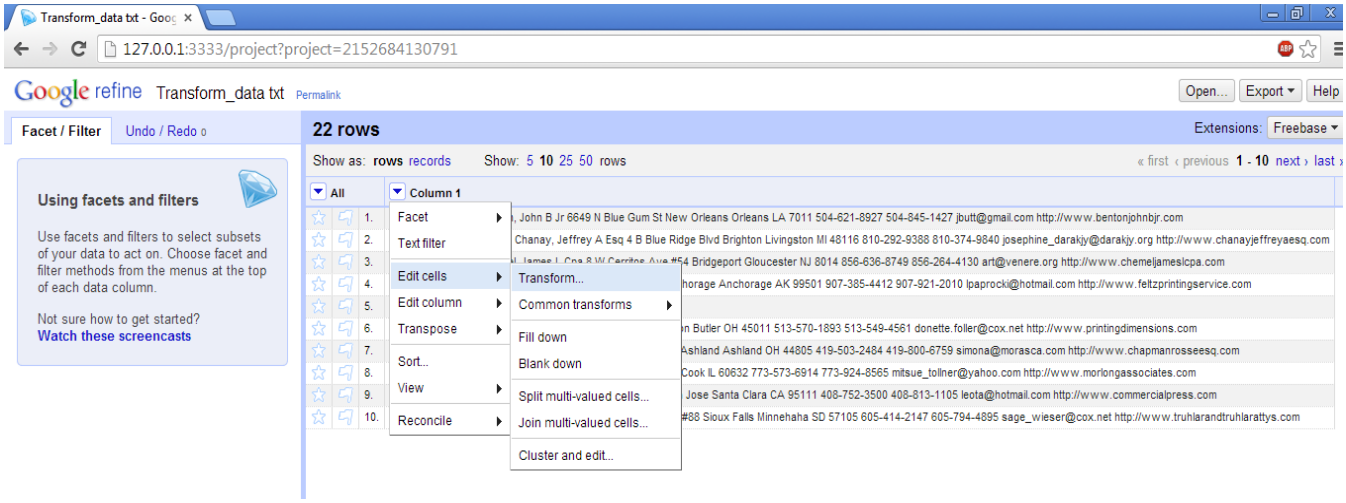
B) Steps

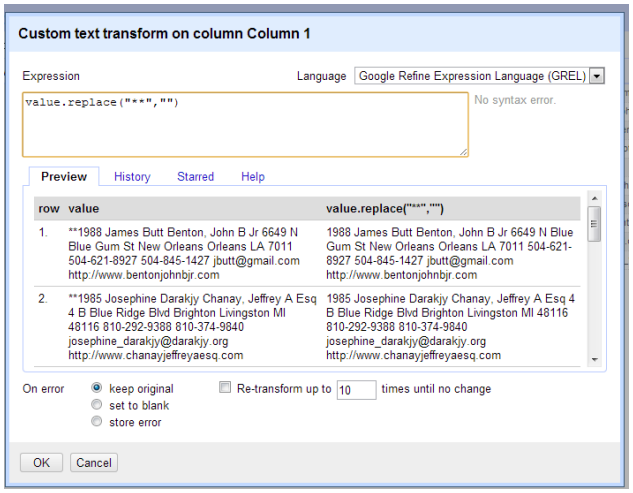
We can remove space or any other unwanted symbols from the data too. Imported text file. The text file had some blank rows but while importing text file, we can select option not to store blank rows and it filters the data while creating project only.

Tutorial: OpenRefine

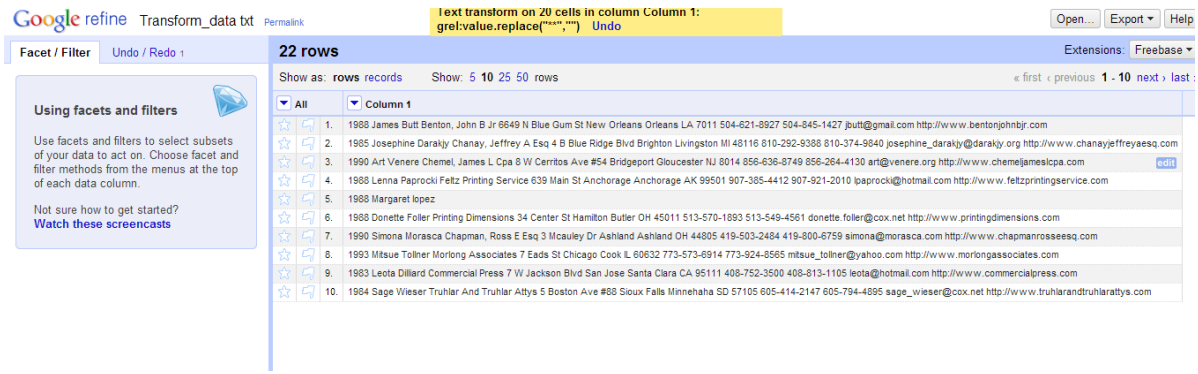


1) Using the expression value.replace("**","") to remove double stars.

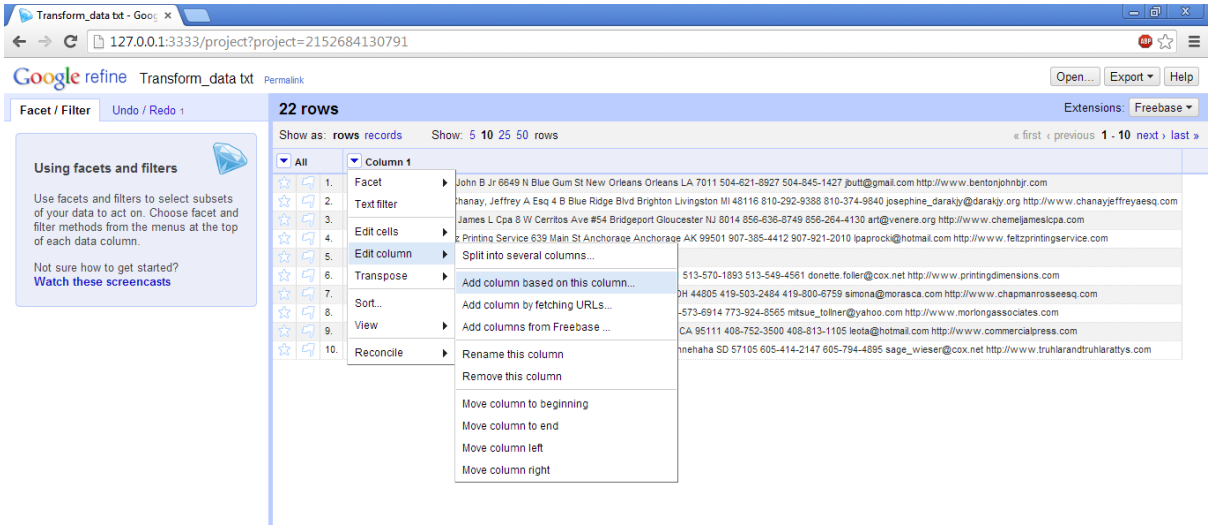




Tutorial: OpenRefine

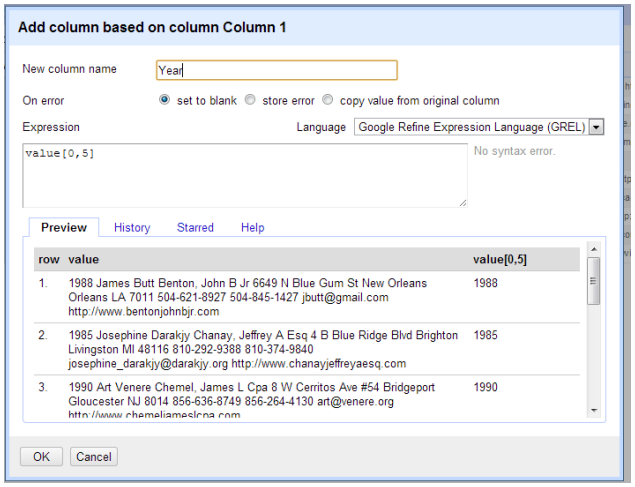


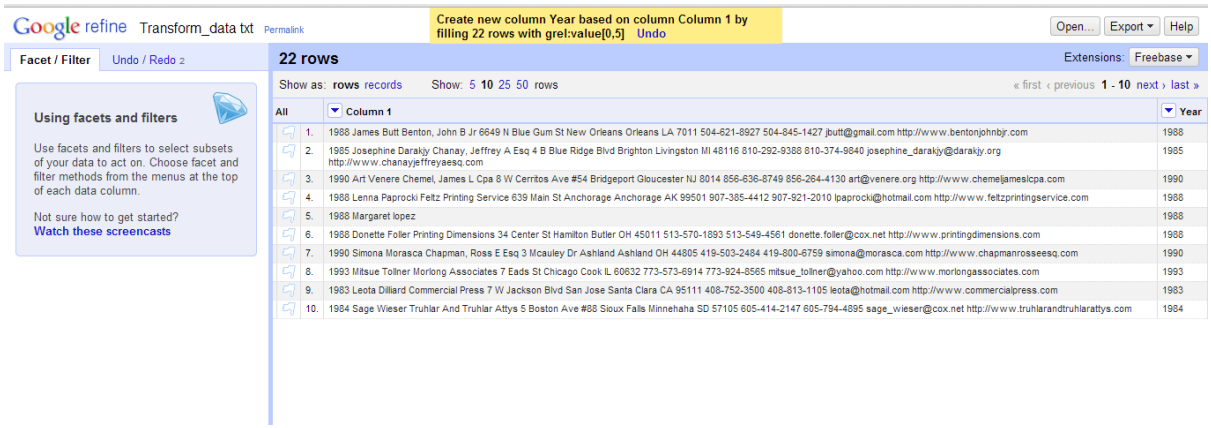
2) To remove year prefixed, select option “add column based on this column”



3) Use expression value[1,5] which specifies the character range to separate years as below

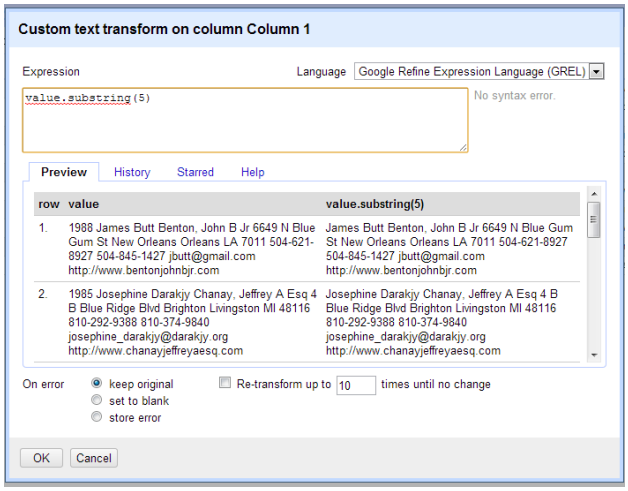
Tutorial: OpenRefine



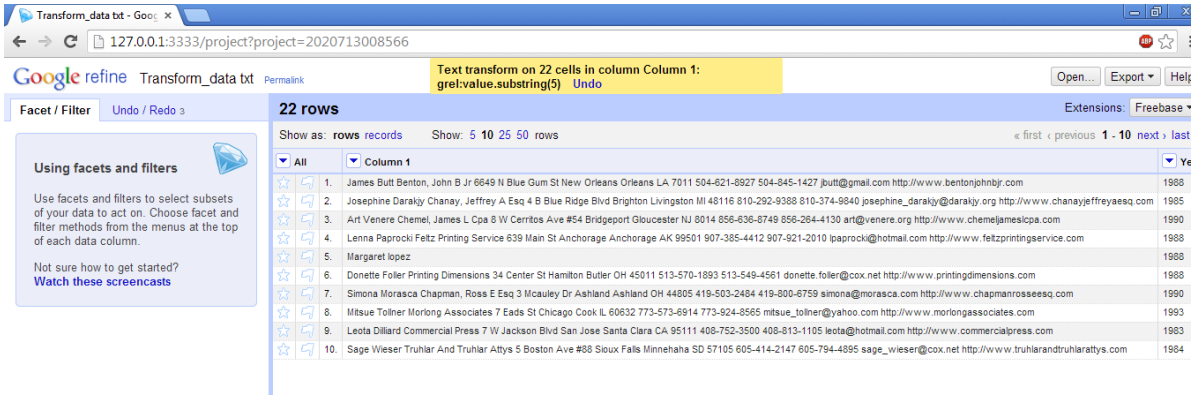


4) We can remove the year from original column by “Transform” command

5) Using character string “value.substring(5)” which displays the data excluding year.



Tutorial: OpenRefine



(ii) Understanding Expressions: - OpenRefine support 'Expressions'. And these are used to transform existing data or create new data based on existing data .This sounds similar to the 'Formula' which we used to have in Excel. But there is a big difference between them. The 'Formula' in the Excel can only be used to store various formulae for each cell for that specific column.

Whereas, in Expressions, here by making use of "GREL", we can access every row and column and can set up conditions according to that.

A) Navigation

Edit Cells→Transform

B) Steps

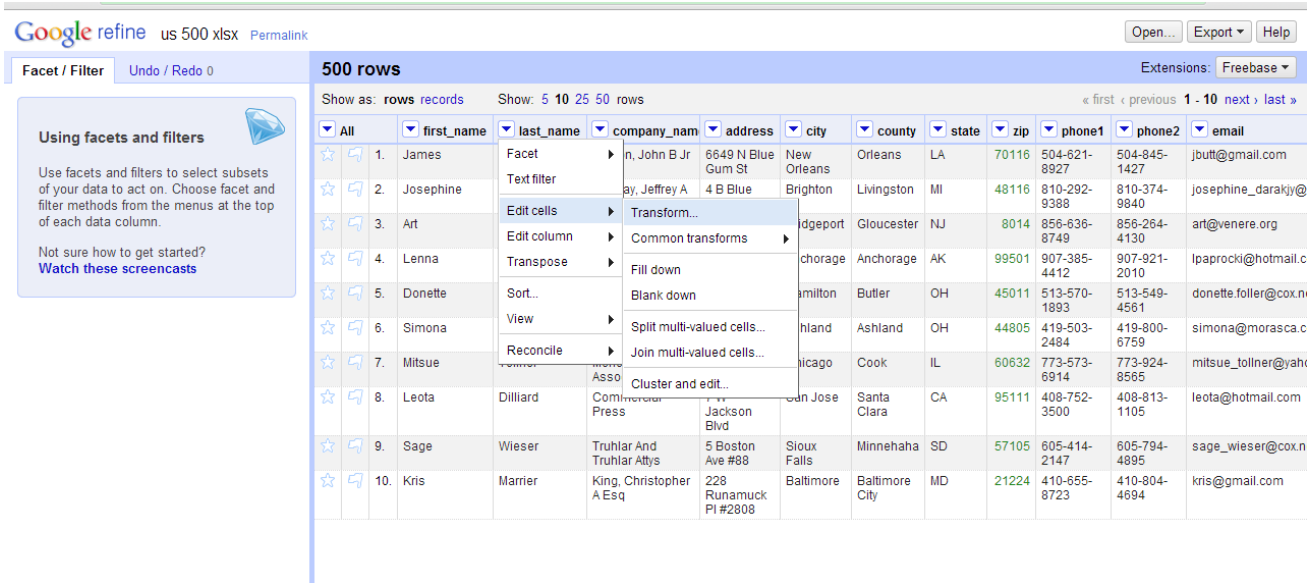
Dataset

first_name	last_name	company_name	address
James	Butt	Benton, John B Jr	6649 N Blue Gum St
Josephine	Darakjy	Chanay, Jeffrey A Esq	4 B Blue Ridge Blvd
Art	Venere	Chemel, James L Cpa	8 W Cerritos Ave #54

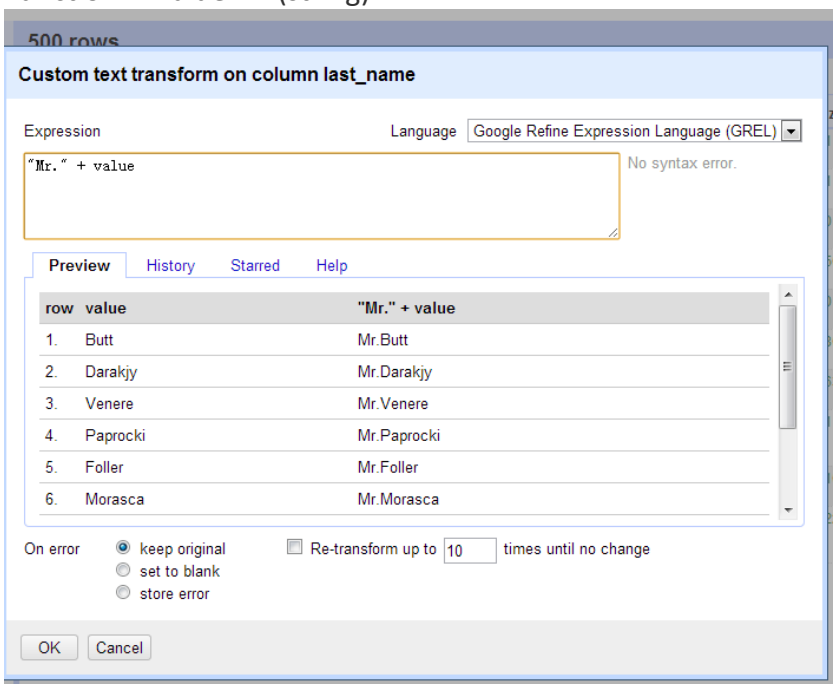
When you invoke the Transform command on, say, column "friend" and enter an expression, OpenRefine will go through each row in the data (matching facets and filters, if any), and evaluate that expression for that row in order to obtain a result for that row.

Tutorial: OpenRefine

To transform data, we should choose one row and click Edit Cells and then Transform.

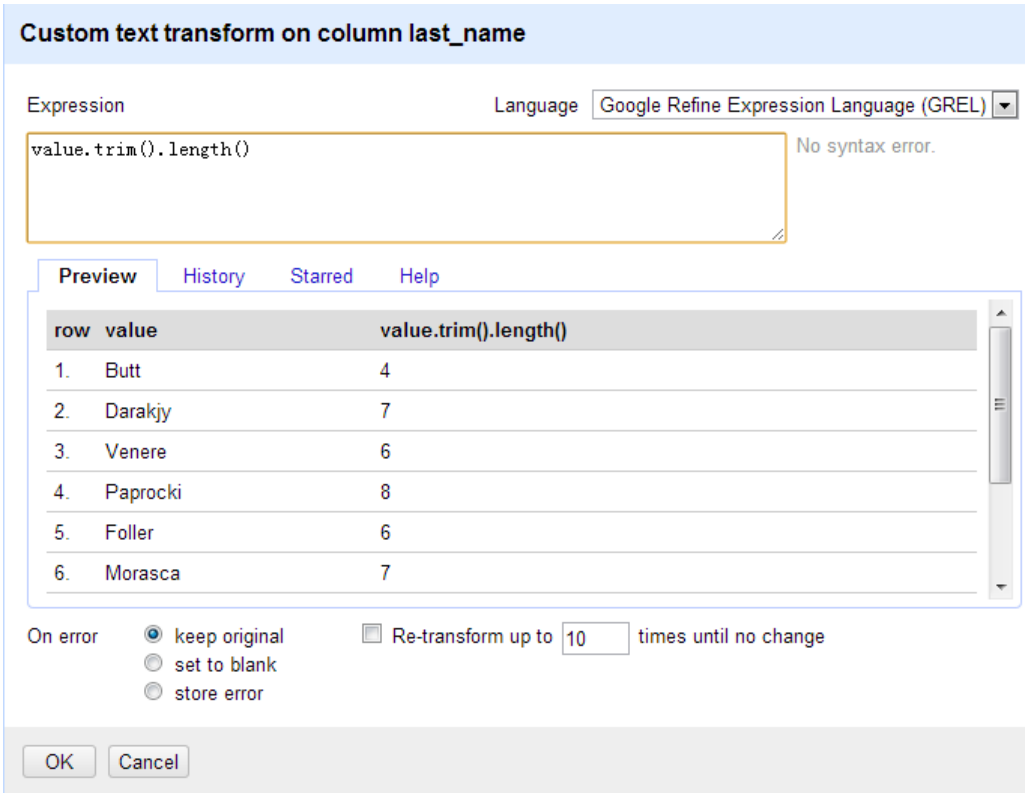


Function 1: value + " (string)"



Function 2: value.trim().length()

Tutorial: OpenRefine



Some Functions at a Glance

1) Syntax: - “Mr.” + value

Explanation: - used to concatenate two strings

2) Syntax: - value + 3.9

Explanation:-add two numbers, and if the value other than numbers, then it concatenates string

3) Syntax: - value.trim().length()

Explanation:-trims the leading and trailing whitespace

4) Syntax: - value.substring(0,3)

Explanation: - take the substring of value from character index 0 up to and excluding character index 3

d. RECONCILIATION

The Reconciliation feature is used to link text names or values in the columns of your data to database identifiers in various database ID spaces. It helps you to develop Metadata to your data.

There are various methods by which Reconciliation is achieved. One of them includes

1) Extending Data by Calling Web Services

Open Refine has a very useful functionality of extending data. For example, there may be a column off addresses with street names, zip code etc. However, you are interested in finding the longitude and latitude as well. In such a scenario, you can call a web service based on that column, and create a new column out of it with latitude and longitude. The web service in such a case can be a Google Geocoding API Web Service, which basically gives the latitude and longitude based on the address. Similarly, you can fetch URLs.

A) Navigation

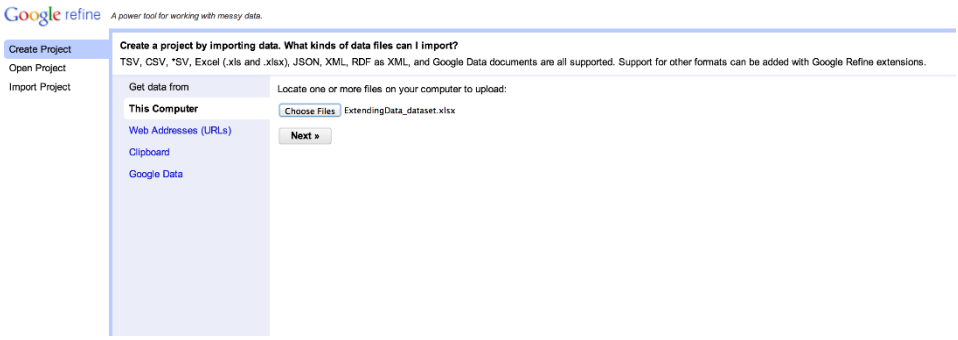
Edit Columns→Add Column by fetching URLs

B) Steps

To explain this feature, we are presenting a scenario of Facebook pages. So basically, with just the ID, as the column, we will fetch the URLs of each page.

Step 1:

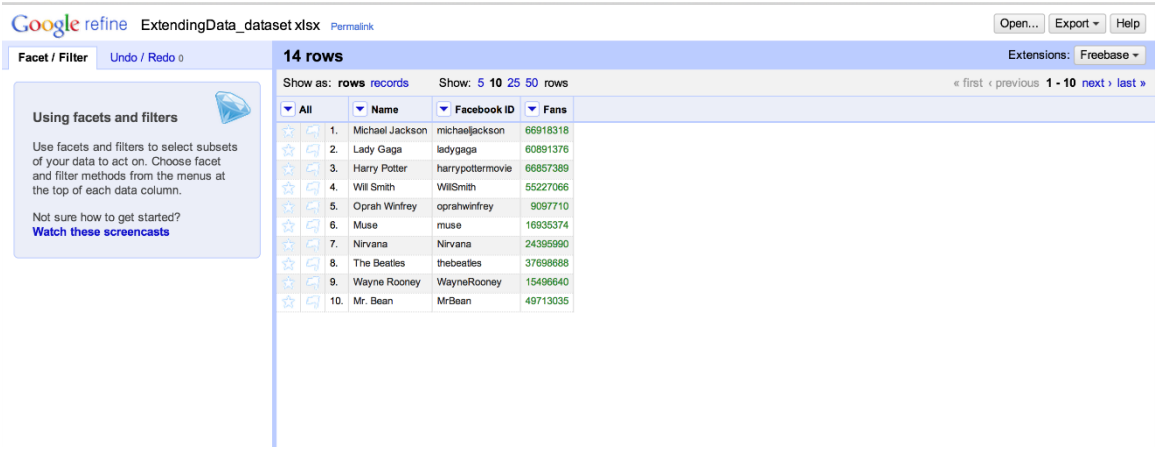
Import the dataset to OpenRefine.



Step 2:

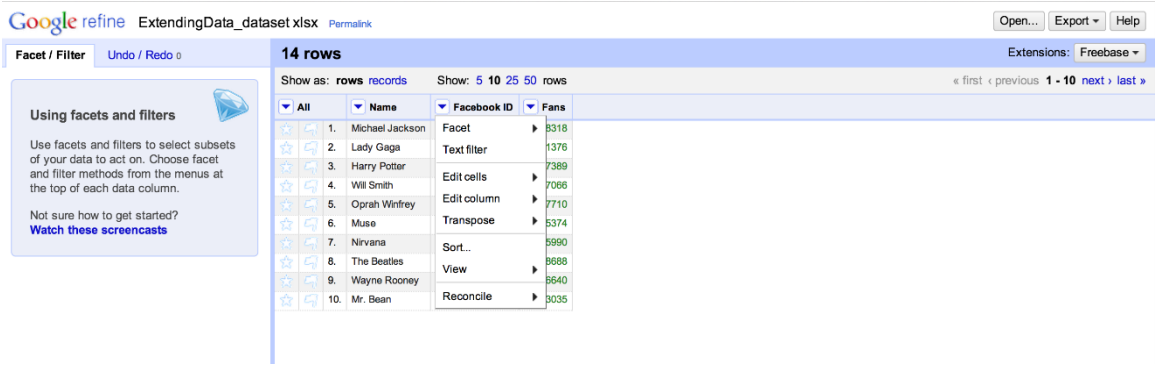
Create Project

Tutorial: OpenRefine



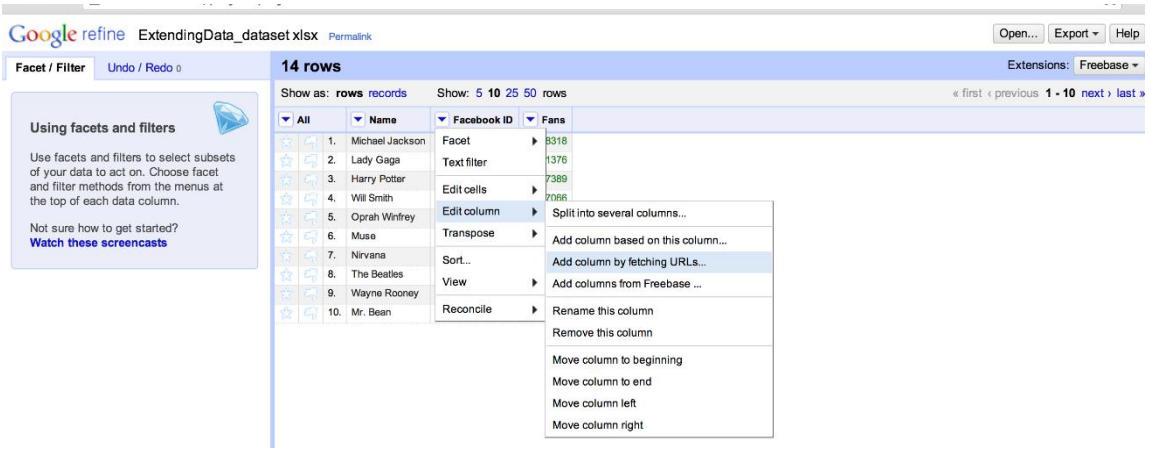
Step 3:

Click the drop drown button in the required column – in this case – Facebook ID.



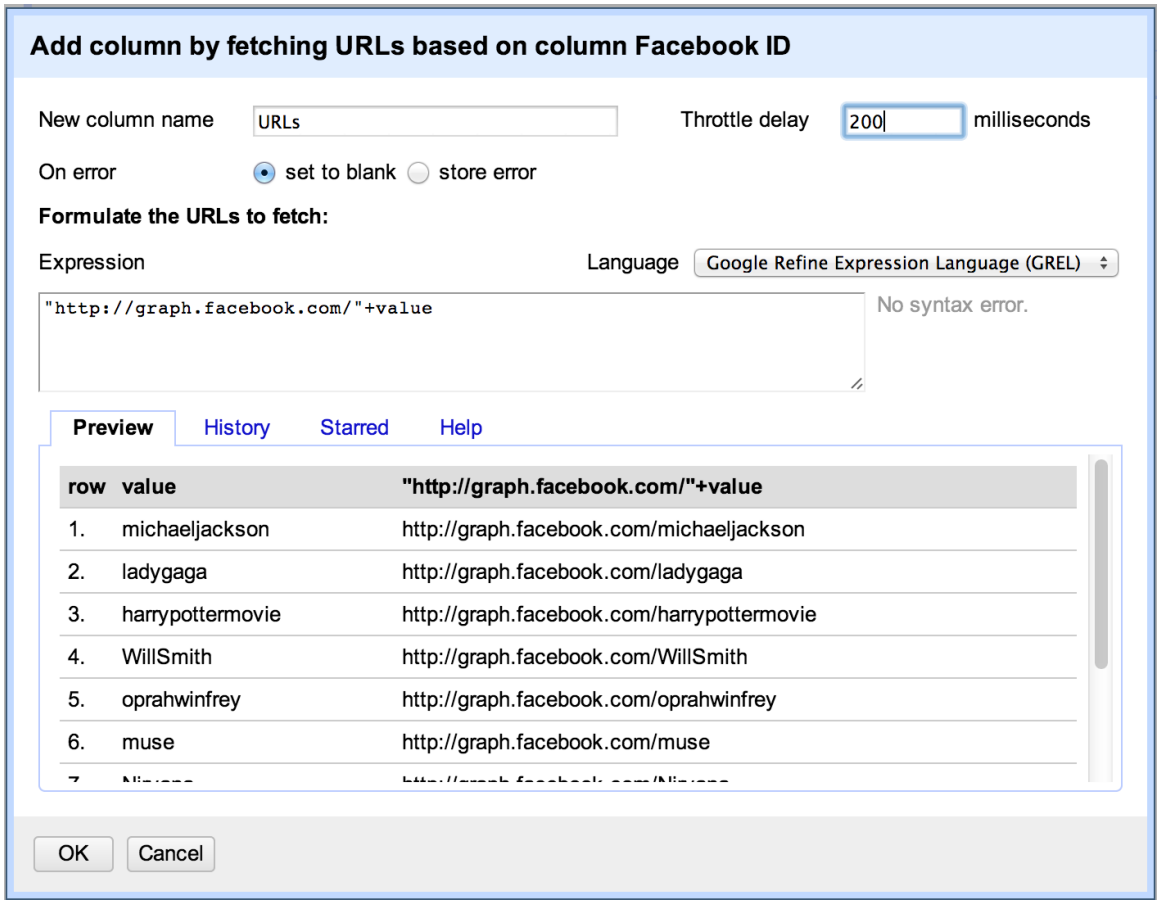
Step 4:

Select edit column -> Add column by fetching URLs.



Step 5:

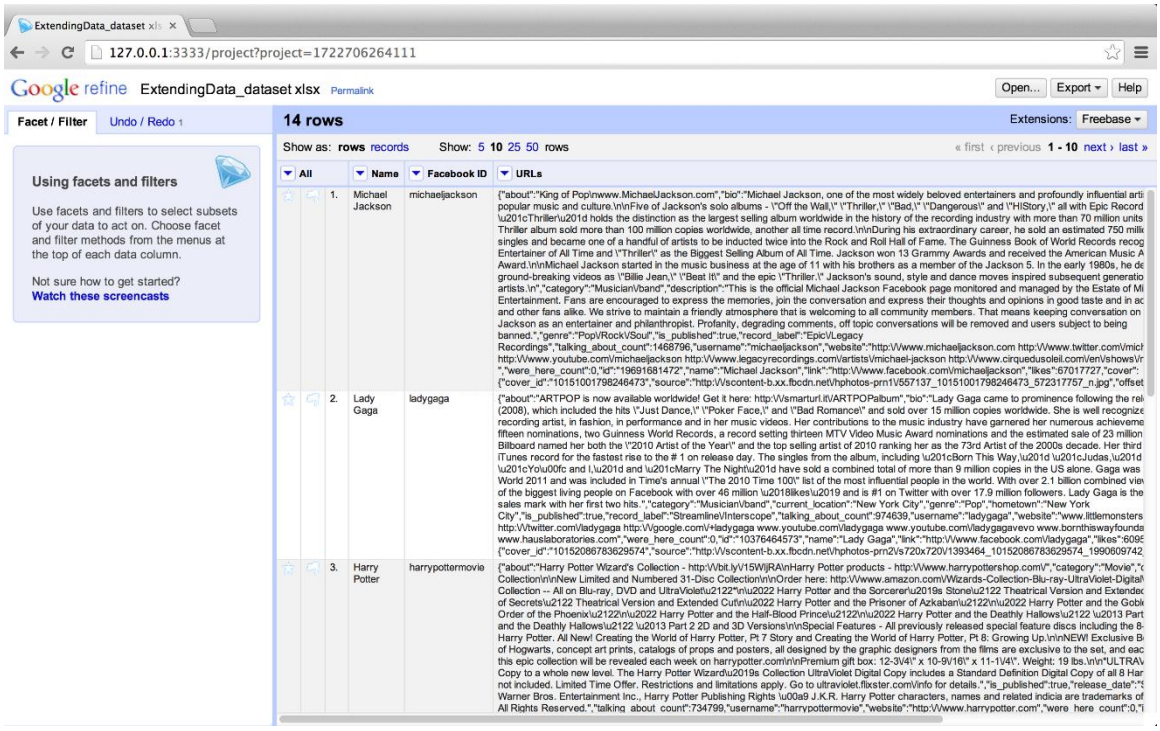
In the expression bar, put the URL of the web service for fetching the URLs.



Step 6:

Give a time limit for the throttle delay, and a column name. Click OK. Throttle delay tells OpenRefine to wait the specified number of milliseconds between each URL requests. Now, result will be stored into the cell in the new column on the same row as the original cell. The result is in JSON, thus it will have to be parsed, i.e., syntactically analysed.

Tutorial: OpenRefine

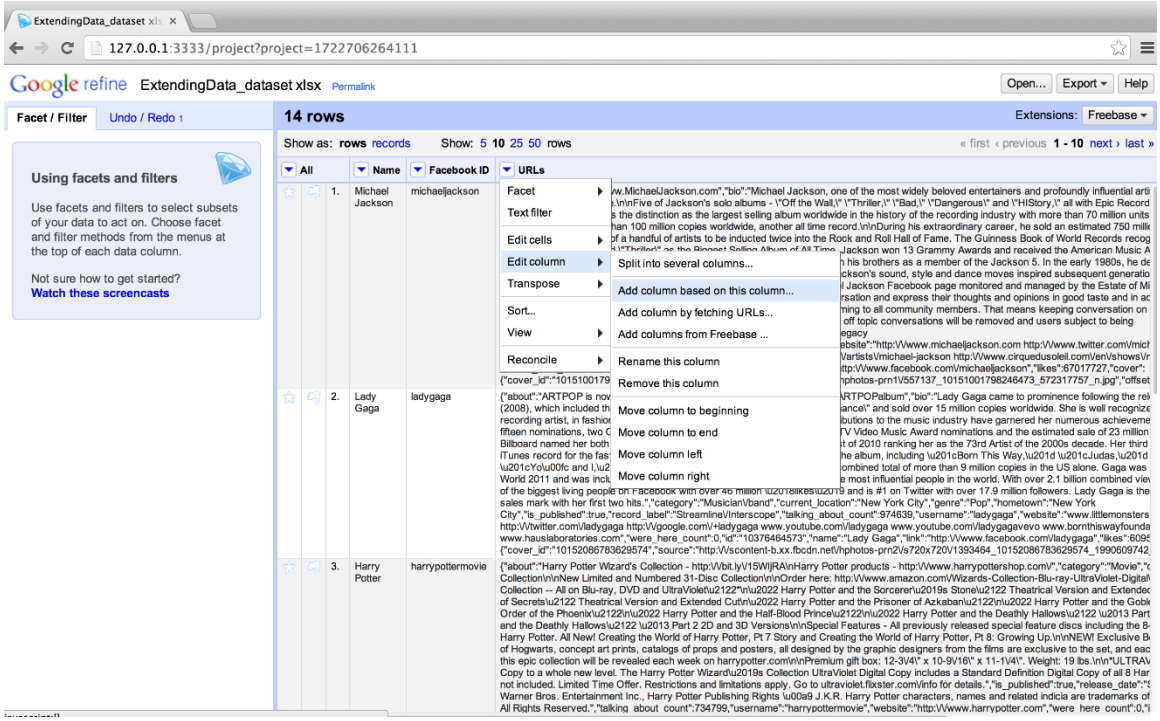


Step 7:

Click the drop down button in the newly created column, in this case, URLs.

Step 8:

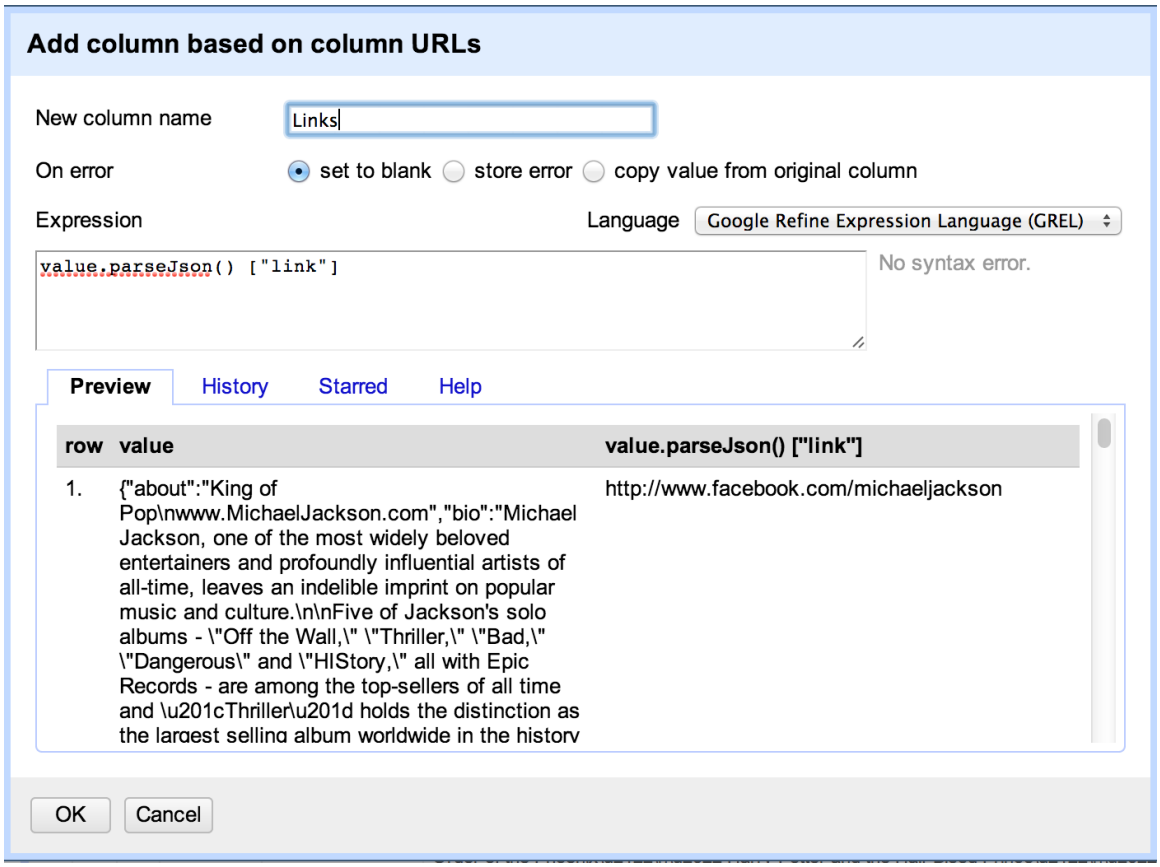
Select edit column -> Add column based on this column.



Tutorial: OpenRefine

Step 9:

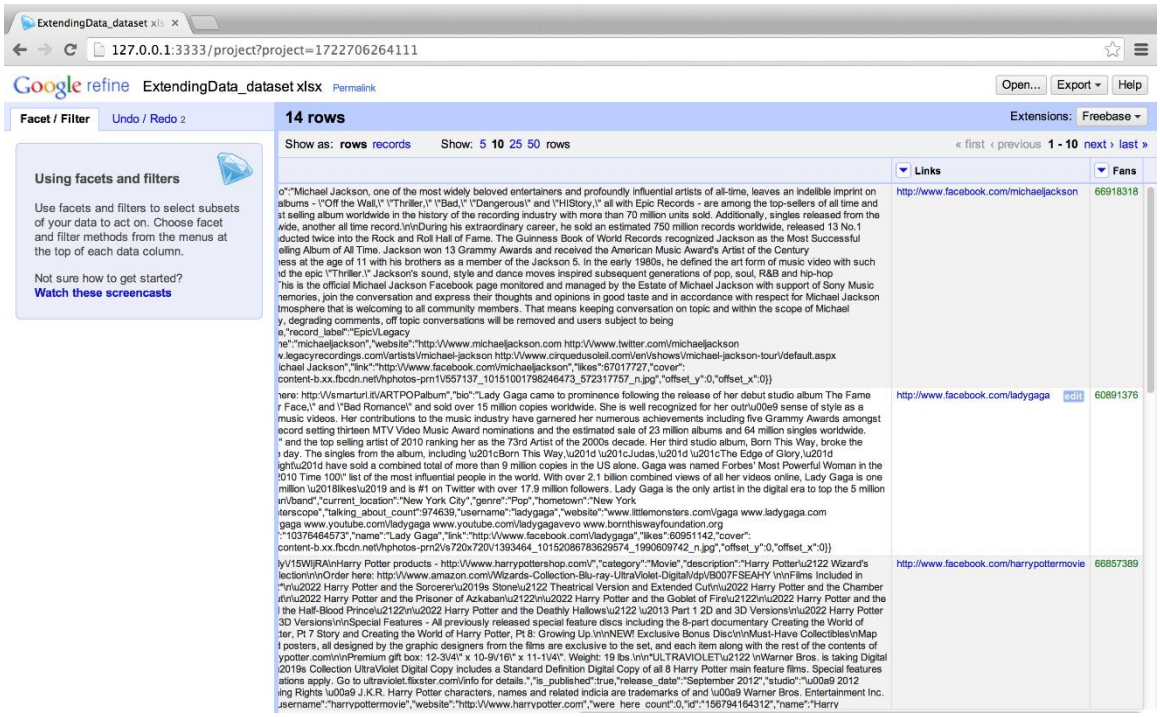
In the expression bar, type: value.parseJson() [“link”]. Give a column name and click OK.



We write [“link”], so that only the links are parsed and fetched from the entire list of information retrieved.

A new column with the links of each of the pages will be displayed.

Tutorial: OpenRefine



2) Extending data by using Database API

We can use the "reconciliation" feature to link text names in our data to the database. By connecting your data with other databases, we get more value out of our data. Technically, it is a semi-automated process of matching text names to database IDs (keys). It helps you to find Metadata for your data.
Enlisting a few links of API databases to connect to:

1) http://data.ordnancesurvey.co.uk/datasets/boundary-line/apis/reconciliation[2]

2) http://opencorporates.com/reconcile[3]

A) Navigation

Column→Reconciliation→Start Reconciling
Then, add standard service→Enter the API database in which you want to search for→Start Reconciling

B) Example

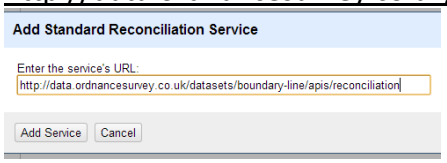
Suppose there is a list of alumni students of a university and we have to find their details about the location of county in the map is, so as to where these counties are located. Openrefine give us a strong feature to link to databases. As we want to know the location information, we choose the API of Ordnance Survey Linked Data, which allows users to turn text into URIs by matching against the Ordnance Survey linked data.

Tutorial: OpenRefine

C) Steps

Step 1

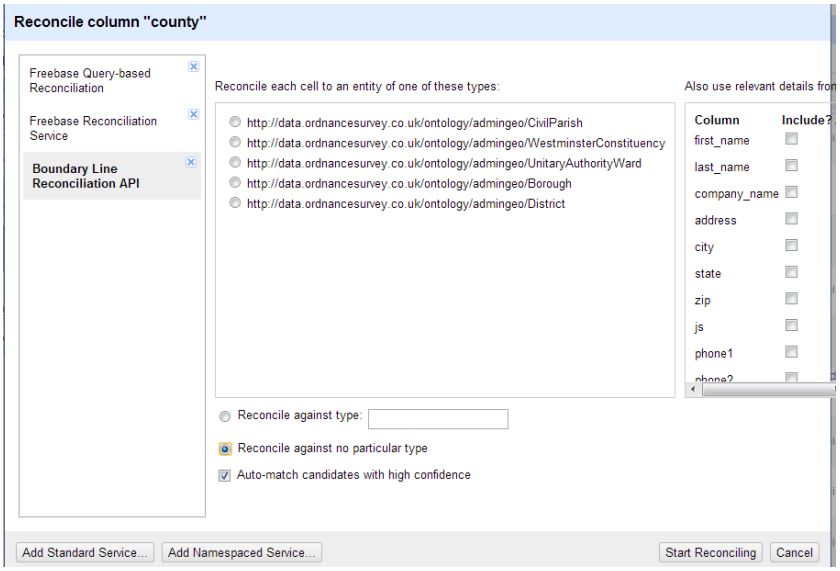
Click the down arrow of county column and select reconcile -> start reconciling. Now click ‘Add Standard Service’ and add the following URL: http://data.ordnancesurvey.co.uk/datasets/boundary-line/apis/reconciliation.



Tutorial: OpenRefine

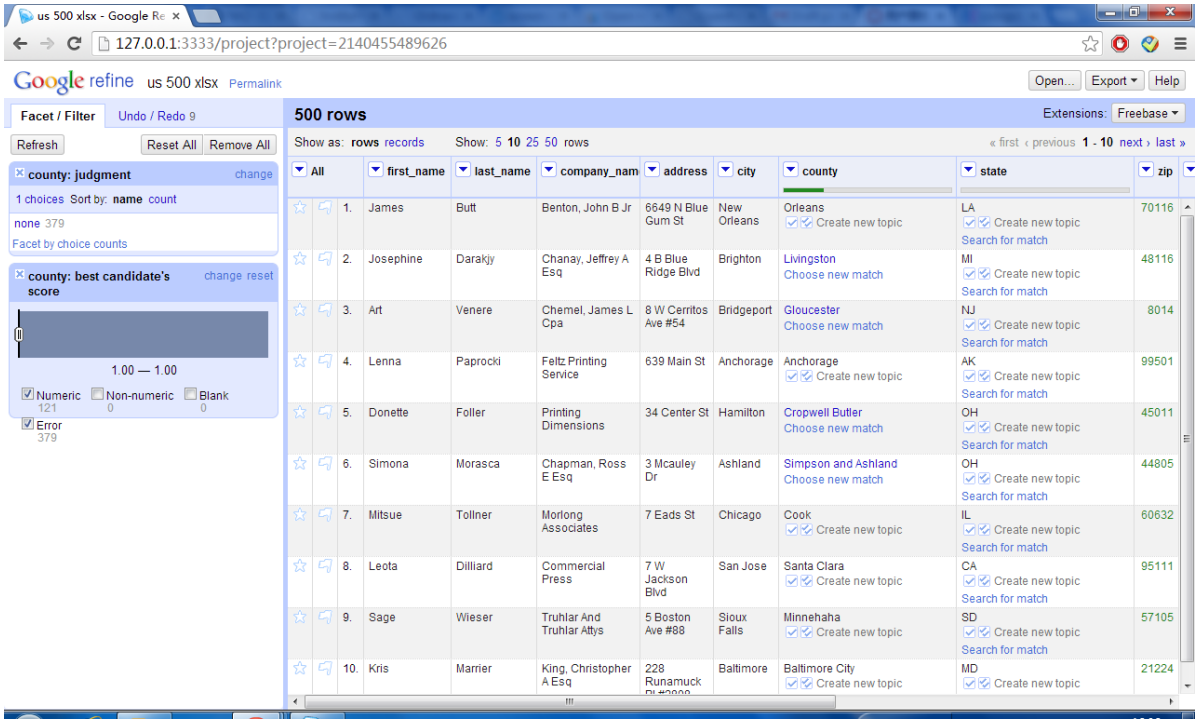
Step 2

As the ‘county’ column will contain a mixture of types select the ‘reconcile against no particular type’ option and click ‘start reconciling’.

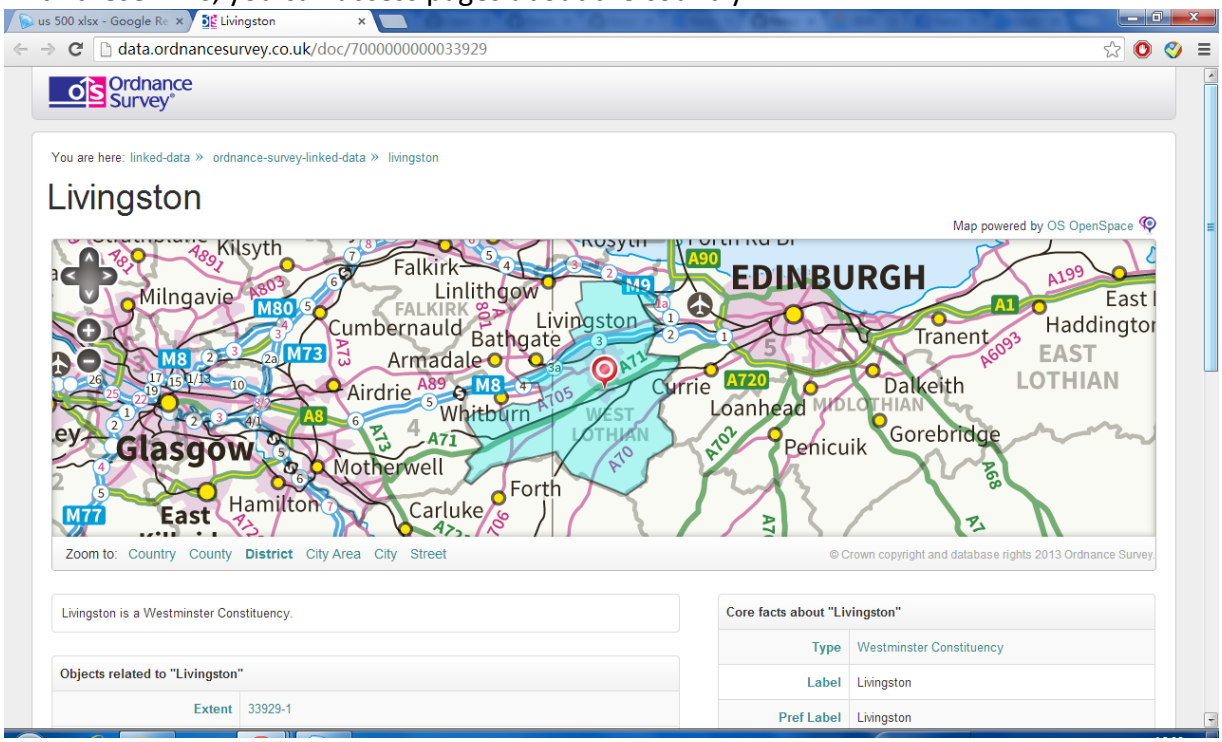


Step 3

You should now see that most of the text labels have turned to hyperlinks.



With these links, you can access pages about the country.



e. EXPORTING DATA

You can export data from an existing OpenRefine project in several formats:

1) Tab separated values (TSV)

2) Comma separated values (CSV)

3) Excel

4) HTML table

To export from a project, we just need to click on the Export button at the top right corner and select the format we want.

f. UNDO/ REDO

Undo/Redo function: - The Undo/Redo function of this tool gives you the flexibility to make mistakes and to rectify them. It gives you an opportunity to make a lot of trials and error on your data.

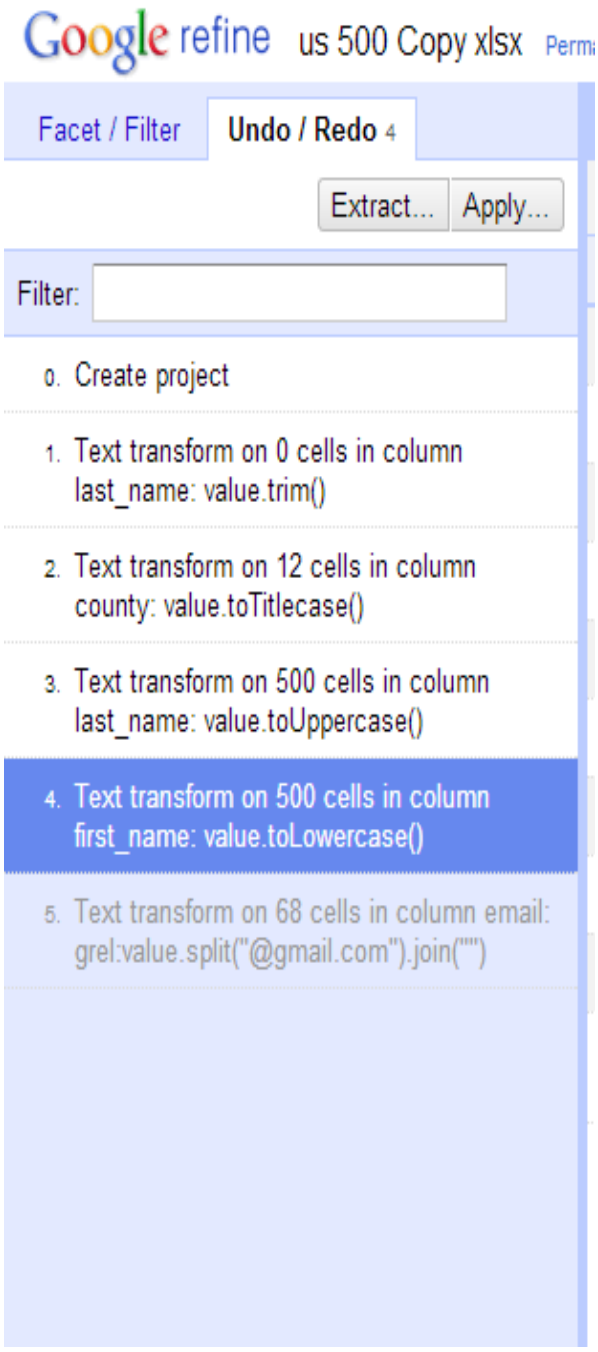
To undo several actions at the same time, select the actions in the list that you want to undo, and then click on the step number you would like to go back to. The tool will automatically take you to the earlier step undoing all the actions you performed in the middle disappear. All of the actions that you selected can be undone or reversed.

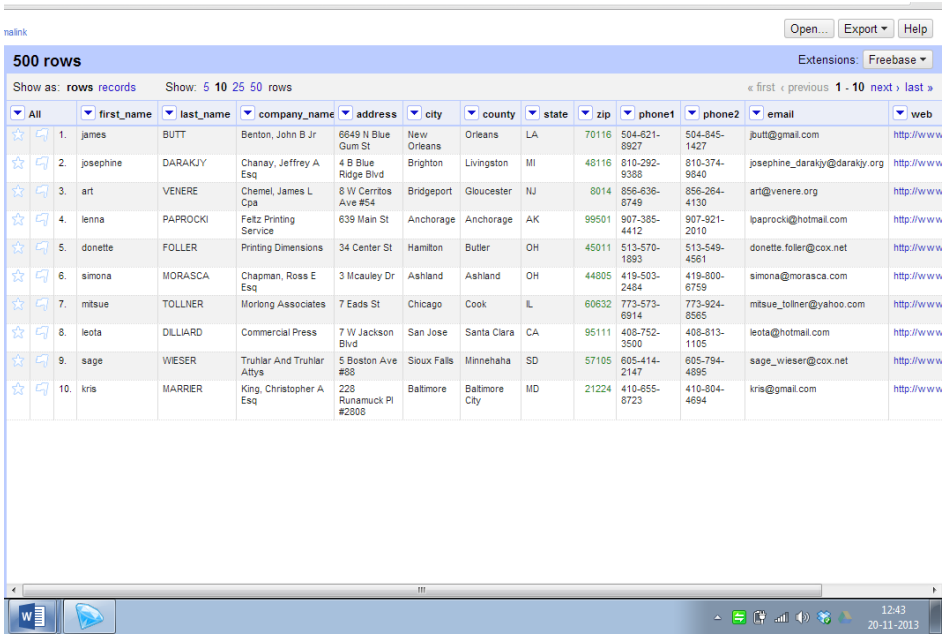
Even the redo functions let you go to the step where you want to go.

(i) Before Undo/Redo

Here, you can see that on the left side of the panel, there are five steps which have been performed by the user .Out of which, the last one is greyed. The greyed step number 5 explains that Undo has already been performed once. To prove this point, we perform 'Undo' once again.

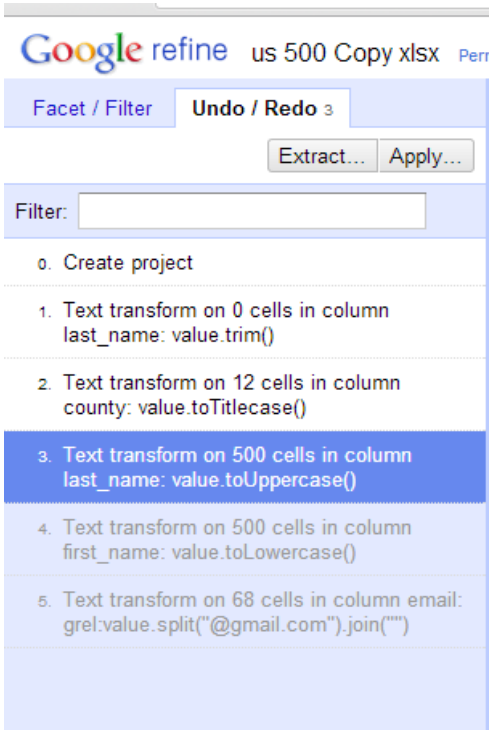
So, following is the screenshot before we perform the Undo Function and along with it is the Data which is present.

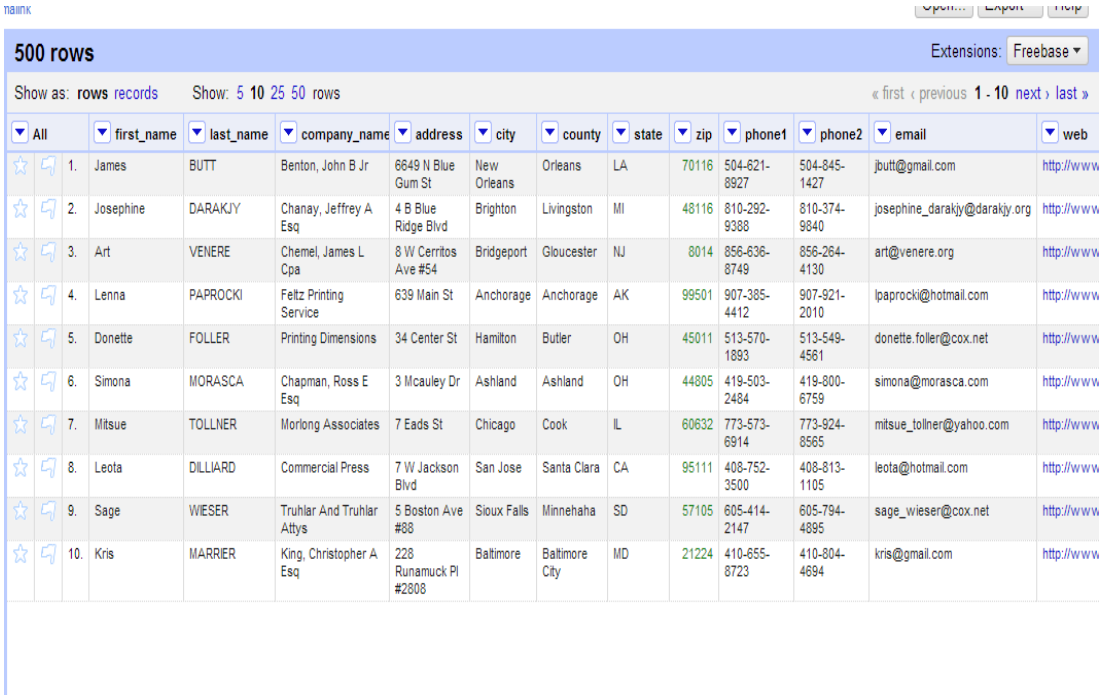




(ii) After Undo/Redo

Here, you can see the same five steps with two greyed steps: step number 4 and step number 5.Here, we can see that the Undo function is performed and you can also see the data being altered to the previous step; which is making the first alphabet of the ‘first_name ‘ column capital.





500 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All		first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	web
	1.	James	BUTT	Benton, John B Jr	6649 N Blue Gum St	New Orleans	Orleans	LA	70116	504-621-8927	504-845-1427	jbutt@gmail.com	http://www
	2.	Josephine	DARAKJY	Chanay, Jeffrey A Esq	4 B Blue Ridge Blvd	Brighton	Livingston	MI	48116	810-292-9388	810-374-9840	josephine_darakjy@darakjy.org	http://www
	3.	Art	VENERE	Chemel, James L Cpa	8 W Cerritos Ave #54	Bridgeport	Gloucester	NJ	8014	856-636-8749	856-264-4130	art@venere.org	http://www
	4.	Lenna	PAPROCKI	Feltz Printing Service	639 Main St	Anchorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com	http://www
	5.	Donette	FOLLER	Printing Dimensions	34 Center St	Hamilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.net	http://www
	6.	Simona	MORASCA	Chapman, Ross E Esq	3 Mcauley Dr	Ashland	Ashland	OH	44805	419-503-2484	419-800-6759	simona@morasca.com	http://www
	7.	Mitsue	TOLLNER	Morlong Associates	7 Eads St	Chicago	Cook	IL	60632	773-573-6914	773-924-8565	mitsue_tollner@yahoo.com	http://www
	8.	Leota	DILLIARD	Commercial Press	7 W Jackson Blvd	San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com	http://www
	9.	Sage	WIESER	Truhlar And Truhlar Attys	5 Boston Ave #88	Sioux Falls	Minnehaha	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.net	http://www
	10.	Kris	MARRIER	King, Christopher A Esq	228 Runamuck Pl #2808	Baltimore	Baltimore City	MD	21224	410-655-8723	410-804-4694	kris@gmail.com	http://www

Similarly, the ‘Redo’ function is also performed. The ‘Redo’ function is used to redo an action that you undid.

Tutorial: OpenRefine

7. Strengths and Weaknesses

Strengths

1. OpenRefine is a desktop application. It opens in the browser as a Local Webserver. So, the data is safe and it doesn't get uploaded to the Google server.
2. It has facets which is used to filter the data into subsets and these clusters can be customized and organised into meaningful data.
3. It has a Browser based interface, and so can handle more data efficiently.
4. Openrefine has a strong feature in extending data -- user can use it to find Meta Data and it can be used to correlate with it.

Weakness

1. The UI of Openrefine is not user friendly. Although the features and functions are strong, the UI make Openrefine looks boring. Besides, in the visualization, the function is not scalable. For instance, Openrefine give user a view of data, but the image is not big enough to figure out complex distribution.
2. Unfortunately Google has removed support for this tool, making few of its features redundant.

8. FAQ

1) OpenRefine opens on my browser. So do I require internet to run it?
Ans. No. OpenRefine doesn't require internet for running. It's a normal application, however runs on a local server.

2) What is the index of the first characteristic of one cell?
Ans: It should be 0. So if you want the first two characteristic, you should write as value[0,1].

3) Is OpenRefine only for Windows Operating System?
Ans. OpenRefine is compatible with both Windows Operating System and Mac OS as well.

4) Is my data safe in Openrefine?
Ans: Yes it is safe because it runs on local server and data is stored on your computer's directory.

9. Installation

It is an open source tool and you can go to the website "http://openrefine.org/download.html" and download and unzip it. Then run google-refine.exe and it will automatically open a browser window to start Openrefine.

Tutorial: OpenRefine

10. Resources

1. http://www.briandunning.com/
2. http://data.ordnancesurvey.co.uk/datasets/boundary-line/apis/reconciliation
3. http://opencorporates.com/reconcile