

# Tutorial: OpenRefine



By

Atima

Han Zhuang

Ishita Vedvyas

Rishikesh Dole

# Tutorial: OpenRefine

## INDEX

1. Introduction	3
2. Big Idea	3
3. Why OpenRefine?	4
4. Background	4
5. Dataset	4
6. Key Features	
a. Importing Data	5
b. Filtering/ Faceting	7
c. Editing cells/columns	15
d. Reconciliation	19
e. Exporting Data	22
f. Undo/ Redo	30
7. Strengths and Weaknesses	34
8. FAQ	34
9. Installation	34
10. Resources	35

# 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
<ol style="list-style-type: none"><li>1. Batch editing of rows and columns possible</li><li>2. Used for exploring and transforming data.</li><li>3. No Schema Required</li><li>4. Data is always visible at each step of editing.</li><li>5. More interactive and visual.</li></ol>	<ol style="list-style-type: none"><li>1. Editing of one cell at a time.</li><li>2. Used for entering data and performing calculations, functions.</li><li>3. No Schema Required</li><li>4. Data is always visible</li><li>5. Visual is not impressive.</li></ol>	<ol style="list-style-type: none"><li>1. Schema and programming language required for editing.</li><li>2. Data is out of sight unless script is run to view it.</li></ol>

## 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/>"<sup>1</sup>

Following are some of the examples of sample sets used.

# Tutorial: OpenRefine

1)



2)



3)



## 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**

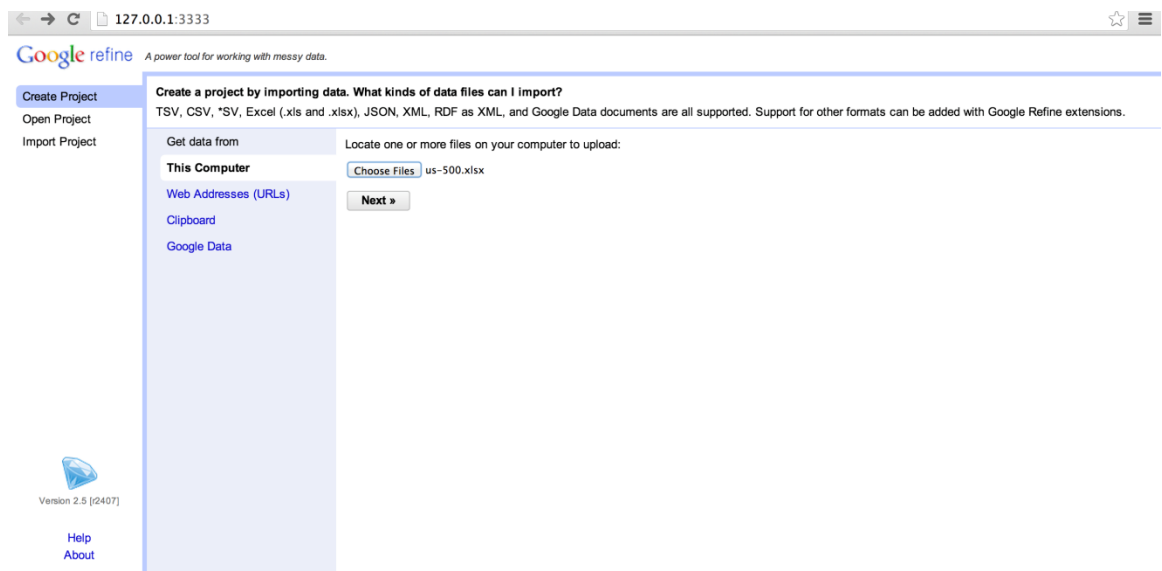
# Tutorial: OpenRefine

## 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.

## Parsing Data

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

	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email
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
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
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
4.	Lenna	Paprocki	Feltz Printing Service	639 Main St	Anchorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com
5.	Donette	Foller	Printing Dimensions	34 Center St	Hamilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.net
6.	Simona	Morasca	Chapman, Ross E Esq	3 McAuley Dr	Ashland	Ashland	OH	44805	419-503-2484	419-800-6759	simona@morasca.com
7.	Mitsue	Tolner	Morlong Associates	7 Eads St	Chicago	Cook	IL	60632	773-573-6914	773-924-8565	mitsue_tolner@yahoo.com
8.	Leota	Dillard	Commercial Press	7 W Jackson Blvd	San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com
9.	Sage	Wieser	Truhlar And Truhlar Atlys	5 Boston Ave #88	Sioux Falls	Minnehaha	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.net
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

**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 split data sorted by their counts.

# Tutorial: OpenRefine

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.

The screenshot shows the OpenRefine interface with a data table of 500 rows. A dropdown menu is open over the 'state' column, showing various facet options. The table data includes columns for first\_name, last\_name, company\_name, address, city, county, state, zip, phone1, phone2, email, and web. The 'state' column values include New Orleans, Livingston, Gloucester, Anchorage, Butler, Ashland, Cook, Santa Clara, and Minnesota.

- 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

Facets-Clusters - Google Refine

127.0.0.1:3333/project?project=1877640074305

Google Refine Facets-Clusters Permalink

Open... Export Help

Facet / Filter Undo / Redo

Refresh Reset All Remove All

42 matching rows (500 total) 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	we
38.	Rozella	Ostrosky	Parkway Company	17 Morena Blvd	Camarillo	Ventura	CA	93012	805-832-6163	805-609-1531	rozella.ostrosky@ostrosky.com	http://w
48.	Kanisha	Waycott	Schroer, Gene E Esq	5 Tomahawk Dr	Los Angeles	Los Angeles	CA	90006	323-453-2780	323-315-7314	kanisha_waycott@yahoo.com	http://w
71.	Kallie	Blackwood	Rowley Schlingen Inc	701 S Harrison Rd	San Francisco	San Francisco	CA	94104	415-315-2761	415-604-7609	kallie.blackwood@gmail.com	http://w
73.	Bobbye	Rhym	Smits, Patricia Garty	30 W 80th St #1995	San Carlos	San Mateo	CA	94070	650-528-5783	650-811-9032	brhym@rhym.com	http://w
74.	Micaela	Rhymes	H Lee Leonard Attorney At Law	20932 Hedley St	Concord	Contra Costa	CA	94520	925-647-3298	925-522-7798	micaela_rhymes@gmail.com	http://w
84.	Dominque	Dickerson	E A I Electronic Assocs Inc	69 Marquette Ave	Hayward	Alameda	CA	94545	510-993-3758	510-901-7640	dominique.dickerson@dickerson.org	http://w
91.	Tamara	Wardrip	Jewel My Shop Inc	4800 Black Horse Pike	Burlingame	San Mateo	CA	94010	650-803-1936	650-216-5075	twardrip@cox.net	http://w
92.	Cory	Gibes	Chinese Translation Resources	83649 W Belmont Ave	San Gabriel	Los Angeles	CA	91776	626-572-1096	626-696-2777	cory.gibes@gmail.com	http://w
95.	Elvera	Benimadho	Tree Musketeers	99385 Charity St #840	San Jose	Santa Clara	CA	95110	408-703-8505	408-440-8447	elvera.benimadho@cox.net	http://w
96.	Carma	Vanheusen	Springfield Div Oh Edison Co	68556 Central Hwy	San Leandro	Alameda	CA	94577	510-503-7169	510-452-4835	carma@cox.net	http://w

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.

Facets-Clusters - Google Refine

127.0.0.1:3333/project?project=1877640074305

Google Refine Facets-Clusters Permalink

Open... Export Help

Facet / Filter Undo / Redo

Refresh Reset All Remove All

55 matching rows (500 total) 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	we
38.	Rozella	Ostrosky	Parkway Company	17 Morena Blvd	Camarillo	Ventura	CA	93012	805-832-6163	805-609-1531	rozella.ostrosky@ostrosky.com	http://w
48.	Kanisha	Waycott	Schroer, Gene E Esq	5 Tomahawk Dr	Los Angeles	Los Angeles	CA	90006	323-453-2780	323-315-7314	kanisha_waycott@yahoo.com	http://w
71.	Kallie	Blackwood	Rowley Schlingen Inc	701 S Harrison Rd	San Francisco	San Francisco	CA	94104	415-315-2761	415-604-7609	kallie.blackwood@gmail.com	http://w
73.	Bobbye	Rhym	Smits, Patricia Garty	30 W 80th St #1995	San Carlos	San Mateo	CA	94070	650-528-5783	650-811-9032	brhym@rhym.com	http://w
74.	Micaela	Rhymes	H Lee Leonard Attorney At Law	20932 Hedley St	Concord	Contra Costa	CA	94520	925-647-3298	925-522-7798	micaela_rhymes@gmail.com	http://w
84.	Dominque	Dickerson	E A I Electronic Assocs Inc	69 Marquette Ave	Hayward	Alameda	CA	94545	510-993-3758	510-901-7640	dominique.dickerson@dickerson.org	http://w
91.	Tamara	Wardrip	Jewel My Shop Inc	4800 Black Horse Pike	Burlingame	San Mateo	CA	94010	650-803-1936	650-216-5075	twardrip@cox.net	http://w
92.	Cory	Gibes	Chinese Translation Resources	83649 W Belmont Ave	San Gabriel	Los Angeles	CA	91776	626-572-1096	626-696-2777	cory.gibes@gmail.com	http://w
95.	Elvera	Benimadho	Tree Musketeers	99385 Charity St #840	San Jose	Santa Clara	CA	95110	408-703-8505	408-440-8447	elvera.benimadho@cox.net	http://w
96.	Carma	Vanheusen	Springfield Div Oh Edison Co	68556 Central Hwy	San Leandro	Alameda	CA	94577	510-503-7169	510-452-4835	carma@cox.net	http://w

California

Apply Cancel

Enter Esc

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

# Tutorial: OpenRefine

The screenshot shows the OpenRefine interface for the 'Facets-Clusters' project. The main table displays 55 matching rows for the 'state' facet, sorted by name count. The table columns include first\_name, last\_name, company\_name, address, city, county, state, zip, phone1, phone2, and email. The 'state' facet on the left shows 53 choices, with 'CA' having the highest count at 65.

id	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	url
22	Veronika	Inouye	C 4 Network Inc	6 Greenleaf Ave	San Jose	Santa Clara	CA	95111	408-540-1765	408-813-4592	vinouye@aol.com	http://w
38	Rozella	Ostrosky	Parkway Company	17 Morena Blvd	Camarillo	Ventura	CA	93012	805-832-6163	805-609-1531	rozella.ostrosky@ostrosky.com	http://w
48	Kanisha	Waycott	Schroer, Gene E Esq	5 Tomahawk Dr	Los Angeles	Los Angeles	CA	90006	323-453-2780	323-315-7314	kanisha_waycott@yahoo.com	http://w
58	Shenika	Seewald	East Coast Marketing	4 Otis St	Van Nuys	Los Angeles	CA	91405	818-423-4007	818-749-8650	shenika@gmail.com	http://w
71	Kallie	Blackwood	Rowley Schlingen Inc	701 S Harrison Rd	San Francisco	San Francisco	CA	94104	415-315-2761	415-604-7609	kallie.blackwood@gmail.com	http://w
73	Bobbye	Rhym	Smts, Patricia Garity	30 W 80th St #1995	San Carlos	San Mateo	CA	94070	650-528-5783	650-811-9032	brhym@rhym.com	http://w
74	Micaela	Rhymes	H Lee Leonard Attorney At Law	20932 Hedley St	Concord	Contra Costa	CA	94520	925-647-3298	925-522-7798	micaela_rhymes@gmail.com	http://w
84	Dominque	Dickerson	E A I Electronic Assocs Inc	69 Marquette Ave	Hayward	Alameda	CA	94545	510-993-3758	510-901-7640	dominque.dickerson@dickerson.org	http://w
87	Stephaine	Barfield	Beutelschies & Company	47154 Whipple Ave Nw	Gardena	Los Angeles	CA	90247	310-774-7643	310-968-1219	stephaine@barfield.com	http://w
91	Tammara	Wardrip	Jewel My Shop Inc	4800 Black Horse Pike	Burlingame	San Mateo	CA	94010	650-803-1936	650-216-5075	twardrip@cox.net	http://w

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

The screenshot shows the OpenRefine interface for the 'Faceting-Clustering' project. The main table displays 67 matching rows for the 'city' facet, sorted by name count. The table columns include first\_name, last\_name, company\_name, address, city, county, state, zip, phone1, phone2, and email. The 'city' facet on the left shows 51 choices, with 'CA' having the highest count at 67.

id	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	url
8	Leota	Dillard	Commercial Press	7 W Jackson Blvd	San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com	http://w
13	Kiley	Caldarera	Feiner Bros	25 E 75th St #69	Los Angeles	Los Angeles	CA	90034	310-498-5651	310-254-3084	kiley.caldarera@aol.com	http://w
22	Veronika	Inouye	C 4 Network Inc	6 Greenleaf Ave	San Jose	Santa Clara	CA	95111	408-540-1765	408-813-4592	vinouye@aol.com	http://w
38	Rozella	Ostrosky	Parkway Company	17 Morena Blvd	Camarillo	Ventura	CA	93012	805-832-6163	805-609-1531	rozella.ostrosky@ostrosky.com	http://w
48	Kanisha	Waycott	Schroer, Gene E Esq	5 Tomahawk Dr	Los Angeles	Los Angeles	CA	90006	323-453-2780	323-315-7314	kanisha_waycott@yahoo.com	http://w
58	Shenika	Seewald	East Coast Marketing	4 Otis St	Van Nuys	Los Angeles	CA	91405	818-423-4007	818-749-8650	shenika@gmail.com	http://w
71	Kallie	Blackwood	Rowley Schlingen Inc	701 S Harrison Rd	San Francisco	San Francisco	CA	94104	415-315-2761	415-604-7609	kallie.blackwood@gmail.com	http://w
73	Bobbye	Rhym	Smts, Patricia Garity	30 W 80th St #1995	San Carlos	San Mateo	CA	94070	650-528-5783	650-811-9032	brhym@rhym.com	http://w
74	Micaela	Rhymes	H Lee Leonard Attorney At Law	20932 Hedley St	Concord	Contra Costa	CA	94520	925-647-3298	925-522-7798	micaela_rhymes@gmail.com	http://w
84	Dominque	Dickerson	E A I Electronic Assocs Inc	69 Marquette Ave	Hayward	Alameda	CA	94545	510-993-3758	510-901-7640	dominque.dickerson@dickerson.org	http://w

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

# Tutorial: OpenRefine

The screenshot shows the OpenRefine interface with a data table of 10 rows. The left sidebar displays two faceted filters: 'state' with 53 choices and 'city' with 342 choices. The main table has columns for first\_name, last\_name, company\_name, address, city, county, state, zip, phone1, phone2, email, and web. The first row is James Butt, Benton, John B Jr, 6649 N Blue Gum St, New Orleans, Orleans, LA, 70116, 504-821-8927, 504-845-1427, jbutt@gmail.com, http://www. The second row is Josephine Darakty, Chanay, Jeffrey A Esq, 4 B Blue Ridge Blvd, Brighton, Livingston, MI, 48116, 810-292-9388, 810-374-9640, josephine\_darakty@darakty.org, http://www. The third row is Art Venere, Ornel, James L, 8 W Cerritos Ave #54, Bridgeport, Gloucester, NJ, 8014, 856-636-8749, 856-264-4130, art@venere.org, http://www. The fourth row is Lenna Paprocki, Feltz Printing Service, 639 Main St, Anchorage, Anchorage, AK, 99501, 907-385-4412, 907-921-2010, lpaprocki@hotmail.com, http://www. The fifth row is Donette Foller, Printing Dimensions, 34 Center St, Hamilton, Butler, Ohio, 45011, 513-570-1893, 513-549-4561, donette.foller@cox.net, http://www. The sixth row is Simona Morasca, Chapman, Ross E Esq, 3 McAuley Dr, Ashland, Ashland, Ohio(oh), 44805, 419-503-2484, 419-800-6759, simona@morasca.com, http://www. The seventh row is Mitsue Tollner, Morlong Associates, 7 Eads St, Chicago, Cook, IL, 60632, 773-573-6914, 773-924-8565, mitsue\_tollner@yahoo.com, http://www. The eighth row is Leota Dillard, Commercial Press, 7 W Jackson Blvd, San Jose, Santa Clara, Cal-CA, 95111, 408-752-3500, 408-813-1105, leota@hotmail.com, http://www. The ninth row is Sage Wieser, Truhlar And Truhlar Atlys, 5 Boston Ave #68, Sioux Falls, Minnehaha, SD, 57105, 605-414-2147, 605-794-4895, sage\_wieser@cox.net, http://www. The tenth row is Kris Marrier, King, Christopher A Esq, 228 Runamuck Pl #2808, Baltimore, Baltimore, MD, 21224, 410-855-8723, 410-804-4694, kris@gmail.com, http://www.

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

Google Refine us 500.xlsx Permalink

Open... Export Help

Facet / Filter Undo / Redo

500 rows Show as: rows records Show: 5 10 25 50 rows Extensions: Freebase

Refresh Reset All Remove All

county 209 choices Sort by: name count Cluster

All	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	web
1	James	Butt	Facet				LA	70116	504-621-8927	504-945-1427	jbutt@gmail.com	http://www
2	Josephine	Darakij	Text filter				MI	48116	810-292-9388	810-374-9840	josephine_darakij@darakij.org	http://www
3	Art	Venere	Edit cells				NJ	8014	856-638-8749	856-264-4130	art@venere.org	http://www
4	Lenna	Paprocki	Edit column				AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com	http://www
5	Donette	Foller	Transpose				OH	45011	513-570-4412	513-549-4561	donette.foller@cox.net	http://www
6	Simona	Morasca	Sort...				OH	44805	419-503-2484	419-800-6759	simona@morasca.com	http://www
7	Mitsue	Tolner	View				IL	60632	773-573-6914	773-924-8565	mitsue_tolner@yahoo.com	http://www
8	Leota	Dillard	Reconcile	sads St	Chicago	Cook	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com	http://www
9	Sage	Wieser	Commercial Press	7 W Jackson Blvd	San Jose	Santa Clara	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.net	http://www
10	Kris	Marrier	Truhlar And Truhlar Atlys	5 Boston Ave #88	Sioux Falls	Minnehaha	MD	21224	410-655-6723	410-804-4694	kris@gmail.com	http://www

javascript()

### Custom Facet on column company\_name

Expression Language Google Refine Expression Language (GREL) ▼

```
value.split(" ")[0]
```

No syntax error.

Preview History Starred Help

row	value	value.split(" ")[0]
1.	Benton, John B Jr	Benton,
2.	Chanay, Jeffrey A Esq	Chanay,
3.	Chemel, James L Cpa	Chemel,
4.	Feltz Printing Service	Feltz
5.	Printing Dimensions	Printing
6.	Chapman, Ross E Esq	Chapman,

OK Cancel

# Tutorial: OpenRefine

The screenshot shows the OpenRefine interface with a table of 2 matching rows. The left sidebar shows facets for 'county' and 'company\_name'. The 'county' facet is expanded, showing 2 choices: Lackawanna and Maricopa. The 'company\_name' facet is also expanded, showing 472 choices, with 'Bailey' selected. The main table displays the following data:

	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	web
126	Rory	Papasergi	Bailey Crtl Co Div Babcock	83 County Road 437 #8581	Clarke Summit	Lackawanna	PA	18411	570-867-7489	570-469-8401	rpapasergi@cox.net	http://www.baileycr.com
347	Helene	Rodenberger	Bailey Transportation Prod Inc	347 Chestnut St	Peoria	Maricopa	AZ	85381	623-461-8551	623-426-4507	helene@aol.com	http://www.baileycr.com

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

The screenshot shows the OpenRefine interface with a table of 500 rows. The left sidebar shows facets for 'county' and 'company\_name'. The 'county' facet is expanded, showing 209 choices. The 'company\_name' facet is also expanded, showing 472 choices. The main table displays the following data:

	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	web
1.	James	Butt	Benton, John B Jr	8649 N Blue Gum St	New Orleans	Orleans	LA					http://www...
2.	Josephine	Darakjy	Chanay, Jeffrey A Esq	4 B Blue Ridge Blvd	Brighton	Livingston	MI					http://www...
3.	Art	Venere	Chemel, James L	8 W Cerritos Ave #54	Bridgeport	Gloucester	NJ					http://www...
4.	Lenna	Paprocki	Feltz Printing Service	639 Main St	Anchorage	Anchorage	AK					http://www...
5.	Donette	Foller	Printing Dimensions	34 Center St	Hamilton	Butler	OH					http://www...
6.	Simona	Morasca	Chapman, Ross E Esq	3 Mcauley Dr	Ashland	Ashland	OH					http://www...
7.	Mitsue	Tollner	Morlong Associates	7 Eads St	Chicago	Cook	IL					http://www...
8.	Leota	Dilliar	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 Atlys	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	MD	21224	410-855-8723	410-804-4694	kris@gmail.com	http://www...

A context menu is open over the 'zip' column, showing options: Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile. The 'Facet' option is selected, and a sub-menu is open showing: Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet..., Custom numeric facet..., and Customized facets.

# Tutorial: OpenRefine

Google refine us 500.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 500 rows Extensions: Freebase

Refresh Reset All Remove All Show as: rows records Show: 5 10 25 50 rows < first < previous 1 - 10 next > last

Allegany 1

Allen 1

Anchorage 4

Anne Arundel 2

Arapahoe 1

Ashland 1

Atlantic 4

---

**company\_name** change

472 choices Sort by: name count

20 1

A 3

Abc 1

Accurel 1

Ace 1

Acme 1

Acqua 1

Admiral 1

Advantage 1

Affiliated 1

Alabama 1

---

**zip** change reset

1,000.00 — 100,000.00

All	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	we
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://w
2.	Josephine	Darakly	Chanay, Jeffrey A Esq	4 B Blue Ridge Blvd	Brighton	Livingston	MI	48116	810-292-9388	810-374-9840	josephine_darakly@darakly.org	http://w
3.	Art	Venere	Chemel, James L Cpe	8 W Cerritos Ave #54	Bridgeport	Gloucester	NJ	8014	856-636-8749	856-264-4130	art@venere.org	http://w
4.	Lenna	Paprocki	Feltz Printing Service	639 Main St	Anchorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com	http://w
5.	Donette	Foller	Printing Dimensions	34 Center St	Hamilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.net	http://w
6.	Simona	Morasca	Chapman, Ross E Esq	3 McAuley Dr	Ashland	Ashland	OH	44805	419-503-2484	419-800-6759	simona@morasca.com	http://w
7.	Mitsue	Tolner	Morlong Associates	7 Eads St	Chicago	Cook	IL	60632	773-573-6914	773-924-8565	mitsue_tolner@yahoo.com	http://w
8.	Leota	Dillard	Commercial Press	7 W Jackson Blvd	San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com	http://w
9.	Sage	Wieser	Truhlar And Truhlar Allys	5 Boston Ave #60	Sioux Falls	Minnehaha	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.net	http://w
10.	Kris	Marrier	King, Christopher A Esq	228 Runamuck Pl #2808	Baltimore	Baltimore City	MD	21224	410-855-8723	410-304-4694	kris@gmail.com	http://w

# 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 use 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

Google refine Transform\_data.txt Permalink

Open... Export Help

Facet / Filter Undo / Redo

22 rows Extensions: Freebase

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

« first < previous 1 - 10 next > last »

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

▼ All	▼ Column 1
1.	**1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbjr.com
2.	**1985 Josephine Darakijy Chanay, Jeffrey A Esq 4 B Blue Ridge Blvd Brighton Livingston MI 48116 810-292-9388 810-374-9840 josephine_darakijy@darakijy.org http://www.chanayjeffreyaesq.com
3.	**1990 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.chemejameslcpa.com
4.	**1988 Lenna Paprocki Feltz Printing Service 639 Main St Anchorage Anchorage AK 99501 907-385-4412 907-921-2010 lpaprocki@hotmail.com http://www.feltzprintingservice.com
5.	**1988 Margaret Lopez
6.	**1988 Donette Foller Printing Dimensions 34 Center St Hamilton Butler OH 45011 513-570-1893 513-549-4561 donette_foller@cox.net http://www.printingdimensions.com
7.	**1990 Simona Morasca Chapman, Ross E Esq 3 McAuley Dr Ashland Ashland OH 44805 419-503-2484 419-800-6759 simona@morasca.com http://www.chapmanroseseesq.com
8.	**1993 Mitsue Tollner Morlong Associates 7 Eads St Chicago Cook IL 60632 773-573-6914 773-924-8565 mitsue_tollner@yahoo.com http://www.morlongassociates.com
9.	**1983 Leota Dillard Commercial Press 7 W Jackson Blvd San Jose Santa Clara CA 95111 408-752-3500 408-813-1105 leota@hotmail.com http://www.commercialpress.com
10.	**1984 Sage Wieser Truhlar And Truhlar Attys 5 Boston Ave #68 Sioux Falls Minnehaha SD 57105 605-414-2147 605-794-4895 sage_wieser@cox.net http://www.truhlarandtruhlarattys.com

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

Transform\_data.txt - Google x

127.0.0.1:3333/project?project=2152684130791

Google refine Transform\_data.txt Permalink

Open... Export Help

Facet / Filter Undo / Redo

22 rows Extensions: Freebase

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

« first < previous 1 - 10 next > last »

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

1. Facet

2. Text filter

3. Edit cells

4. Edit column

5. Transpose

6. Sort...

7. View

8. Reconcile

9. Transform...

10. Cluster and edit...

Custom text transform on column Column 1

Expression Language Google Refine Expression Language (GREL)

value.replace("\*\*", "") No syntax error.

Preview History Starred Help

row	value	value.replace("**", "")
1.	**1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbjr.com	1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbjr.com
2.	**1985 Josephine Darakijy Chanay, Jeffrey A Esq 4 B Blue Ridge Blvd Brighton Livingston MI 48116 810-292-9388 810-374-9840 josephine_darakijy@darakijy.org http://www.chanayjeffreyaesq.com	1985 Josephine Darakijy Chanay, Jeffrey A Esq 4 B Blue Ridge Blvd Brighton Livingston MI 48116 810-292-9388 810-374-9840 josephine_darakijy@darakijy.org http://www.chanayjeffreyaesq.com

On error  keep original  set to blank  store error  Re-transform up to 10 times until no change

OK Cancel



# Tutorial: OpenRefine

The screenshot shows the OpenRefine interface with a text transformation operation applied to a column. The operation is highlighted in yellow and reads: `text transform on 20 cells in column Column 1: grel:value.replace("****", "")`. The interface shows 22 rows of data, with the first row being: 1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 70111 504-621-8927 504-845-1427 jbuttl@gmail.com http://www.bentonjohnbjr.com

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

The screenshot shows the OpenRefine interface with the 'Edit column' menu open. The menu options are: Split into several columns..., Add column based on this column..., Add column by fetching URLs..., Add columns from Freebase..., Rename this column, Remove this column, Move column to beginning, Move column to end, Move column left, and Move column right. The 'Add column based on this column...' option is highlighted.

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

# Tutorial: OpenRefine

**Add column based on column Column 1**

New column name:

On error:  set to blank  store error  copy value from original column

Expression:  Language:  No syntax error.

Preview History Starred Help

row	value	value[0,5]
1.	1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbr.com	1988
2.	1985 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.chanayjeffreyaesq.com	1985
3.	1990 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.chemelijameslcpa.com	1990

OK Cancel

Google refine Transform\_data.txt Permalink Create new column Year based on column Column 1 by filling 22 rows with grel:value[0,5] Undo Open... Export Help

Facet / Filter Undo / Redo 2 22 rows Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows < first < previous 1 - 10 next > last >

All Column 1 Year

1.	1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbr.com	1988
2.	1985 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	1985
3.	1990 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.chemelijameslcpa.com	1990
4.	1988 Lenna Paprocki Feltz Printing Service 639 Main St Anchorage Anchorage AK 99501 907-385-4412 907-921-2010 lpaprocki@hotmail.com http://www.feltzprintingservice.com	1988
5.	1988 Margaret Lopez	1988
6.	1988 Donette Foller Printing Dimensions 34 Center St Hamilton Butler OH 45011 513-570-1893 513-548-4561 donette.foller@cox.net http://www.printingdimensions.com	1988
7.	1990 Simona Morasca Chapman, Ross E Esq 3 McAuley Dr Ashland Ashland OH 44805 419-503-2484 419-800-6759 simona@morasca.com http://www.chapmanrossesq.com	1990
8.	1993 Mitsue Tolner Morlong Associates 7 Eads St Chicago Cook IL 60632 773-573-8914 773-924-8565 mitsue_tolner@yahoo.com http://www.morlongassociates.com	1993
9.	1983 Lesta Dillard Commercial Press 7 W Jackson Blvd San Jose Santa Clara CA 95111 408-752-3500 408-813-1105 lesta@hotmail.com http://www.commercialpress.com	1983
10.	1984 Sage Wieser Truhlar And Truhlar Atlys 5 Boston Ave #88 Sioux Falls Minnehaha SD 57105 605-414-2147 605-794-4895 sage_wieser@cox.net http://www.truhlarandtruhlaratlys.com	1984

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

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.

**Custom text transform on column Column 1**

Expression:  Language:  No syntax error.

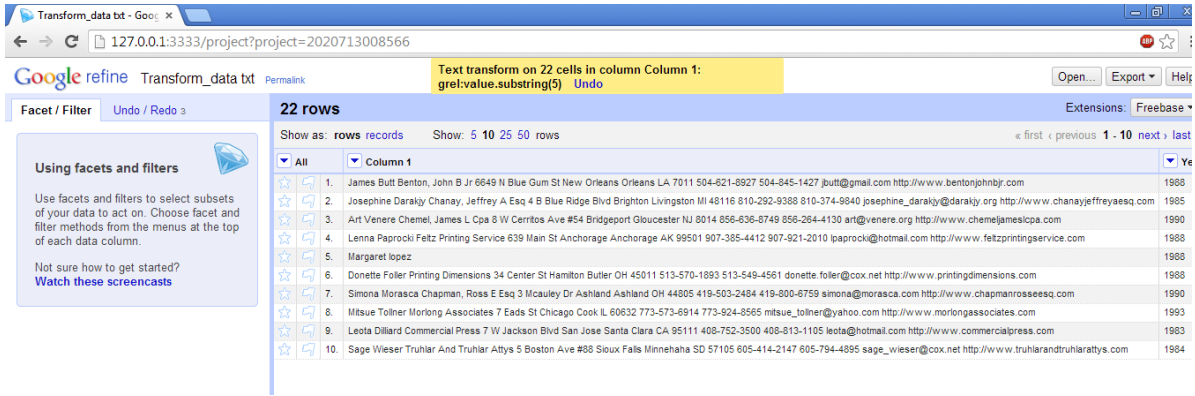
Preview History Starred Help

row	value	value.substring(5)
1.	1988 James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbr.com	James Butt Benton, John B Jr 6649 N Blue Gum St New Orleans Orleans LA 7011 504-621-8927 504-845-1427 jbutt@gmail.com http://www.bentonjohnbr.com
2.	1985 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.chanayjeffreyaesq.com	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.chanayjeffreyaesq.com

On error:  Keep original  Re-transform up to  times until no change  set to blank  store error

OK Cancel

# 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.

Google refine us 500.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 0 **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
1.	James		Butt, John B Jr		6649 N Blue Gum St	New Orleans	Orleans	LA	70116	504-621-8927	504-845-1427	jbutt@gmail.com
2.	Josephine		Darakjy, Jeffrey A		4 B Blue	Brighton	Livingston	MI	48116	810-292-9388	810-374-9840	josephine_darakjy@
3.	Art		Venere			Edgport	Gloucester	NJ	8014	856-636-8749	856-264-4130	art@venere.org
4.	Lenna		Paprocki			Chorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.c
5.	Donette		Foller			Smilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.n
6.	Simona		Morasca			hland	Ashland	OH	44805	419-503-2484	419-800-6759	simona@morasca.c
7.	Mitsue		Tollner			Chicago	Cook	IL	60632	773-573-6914	773-924-8565	mitsue_tollner@yah
8.	Leota		Butt			San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com
9.	Sage		Wieser		Truhlar And Truhlar Allys	Sioux Falls	Minnehaha	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.n
10.	Kris		Marrier		King, Christopher A Esq	Baltimore	Baltimore City	MD	21224	410-655-8723	410-804-4694	kris@gmail.com

Function 1: value + "(string)"

500 rows

Custom text transform on column last\_name

Expression Language Google Refine Expression Language (GREL)

"Mr." + value No syntax error.

Preview History Starred Help

row	value	"Mr." + value
1.	Butt	Mr.Butt
2.	Darakjy	Mr.Darakjy
3.	Venere	Mr.Venere
4.	Paprocki	Mr.Paprocki
5.	Foller	Mr.Foller
6.	Morasca	Mr.Morasca

On error  keep original  set to blank  store error  Re-transform up to 10 times until no change

OK Cancel

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

# Tutorial: OpenRefine

**Custom text transform on column last\_name**

Expression Language Google Refine Expression Language (GREL)

```
value.trim().length()
```

No syntax error.

**Preview** | History | Starred | Help

row	value	value.trim().length()
1.	Butt	4
2.	Darakjy	7
3.	Venere	6
4.	Paprocki	8
5.	Foller	6
6.	Morasca	7

On error  keep original  Re-transform up to  times until no change  
 set to blank  
 store error

OK Cancel

## 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

# Tutorial: OpenRefine

## 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

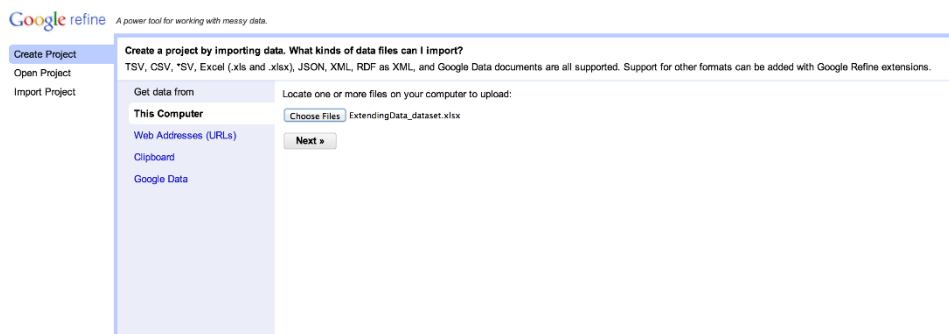
Edit Columns → Add Column by fetching URLs

#### 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

Google refine ExtendingData\_dataset.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 0 14 rows Extensions: Freebase

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

All	Name	Facebook ID	Fans
1.	Michael Jackson	michaeljackson	66918318
2.	Lady Gaga	ladygaga	60891376
3.	Harry Potter	harrypottermovie	66857389
4.	Will Smith	WillSmith	55227066
5.	Oprah Winfrey	oprahwinfrey	9097710
6.	Muse	muse	16935374
7.	Nirvana	Nirvana	24395990
8.	The Beatles	thebeatles	37898688
9.	Wayne Rooney	WayneRooney	15496640
10.	Mr. Bean	MrBean	49713035

Using facets and filters  
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.  
Not sure how to get started? [Watch these screencasts](#)

Step 3:

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

Google refine ExtendingData\_dataset.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 0 14 rows Extensions: Freebase

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

All	Name	Facebook ID	Fans
1.	Michael Jackson	Facet	3318
2.	Lady Gaga	Text filter	1376
3.	Harry Potter	Edit cells	7389
4.	Will Smith	Edit column	7066
5.	Oprah Winfrey	Edit column	7710
6.	Muse	Transpose	5374
7.	Nirvana	Sort...	5990
8.	The Beatles	View	8688
9.	Wayne Rooney	View	8640
10.	Mr. Bean	Reconcile	3035

Using facets and filters  
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.  
Not sure how to get started? [Watch these screencasts](#)

Step 4:

Select edit column -> Add column by fetching URLs.

Google refine ExtendingData\_dataset.xlsx Permalink Open... Export Help

Facet / Filter Undo / Redo 0 14 rows Extensions: Freebase

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

All	Name	Facebook ID	Fans
1.	Michael Jackson	Facet	3318
2.	Lady Gaga	Text filter	1376
3.	Harry Potter	Edit cells	7389
4.	Will Smith	Edit column	7066
5.	Oprah Winfrey	Edit column	7710
6.	Muse	Transpose	5374
7.	Nirvana	Sort...	5990
8.	The Beatles	View	8688
9.	Wayne Rooney	View	8640
10.	Mr. Bean	Reconcile	3035

Using facets and filters  
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.  
Not sure how to get started? [Watch these screencasts](#)

Step 5:

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

# Tutorial: OpenRefine

### Add column by fetching URLs based on column Facebook ID

New column name  Throttle delay  milliseconds

On error  set to blank  store error

**Formulate the URLs to fetch:**

Expression  Language  No syntax error.

**Preview** History Starred Help

row	value	"http://graph.facebook.com/"+value
1.	michaeljackson	http://graph.facebook.com/michaeljackson
2.	ladygaga	http://graph.facebook.com/ladygaga
3.	harrypottermovie	http://graph.facebook.com/harrypottermovie
4.	WillSmith	http://graph.facebook.com/WillSmith
5.	oprahwinfrey	http://graph.facebook.com/oprahwinfrey
6.	muse	http://graph.facebook.com/muse
7.	Mi...	http://graph.facebook.com/Mi...

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

ExtendingData\_dataset.xlsx

127.0.0.1:3333/project?project=1722706264111

Google refine ExtendingData\_dataset.xlsx

Open... Export Help

Facet / Filter Undo / Redo 14 rows

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

All	Name	Facebook ID	URLs
1.	Michael Jackson	michaeljackson	["about:~King of Pop~Michael Jackson.com", "bio:~Michael Jackson, one of the most widely beloved entertainers and profoundly influential artists..."]
2.	Lady Gaga	ladygaga	["about:~ARTPOP is now available worldwide! Get it here: http://www.artsartpop.com..."]
3.	Harry Potter	harrypottermovie	["about:~Harry Potter Wizard's Collection - http://www.harrypotter.com..."]

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.

ExtendingData\_dataset.xlsx

127.0.0.1:3333/project?project=1722706264111

Google refine ExtendingData\_dataset.xlsx

Open... Export Help

Facet / Filter Undo / Redo 14 rows

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

All	Name	Facebook ID	URLs
1.	Michael Jackson	michaeljackson	["about:~King of Pop~Michael Jackson.com", "bio:~Michael Jackson, one of the most widely beloved entertainers and profoundly influential artists..."]
2.	Lady Gaga	ladygaga	["about:~ARTPOP is now available worldwide! Get it here: http://www.artsartpop.com..."]
3.	Harry Potter	harrypottermovie	["about:~Harry Potter Wizard's Collection - http://www.harrypotter.com..."]

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Add column based on this column...

Add column by fetching URLs...

Add columns from Freebase...

Rename this column

Remove this column

Move column to beginning

Move column to end

Move column left

Move column right

# Tutorial: OpenRefine

Step 9:

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

### Add column based on column URLs

New column name

On error  set to blank  store error  copy value from original column

Expression  Language  No syntax error.

**Preview** History Starred Help

row	value	value.parseJson() [\"link\"]
1.	{\"about\": \"King of Pop\\nwww.MichaelJackson.com\", \"bio\": \"Michael Jackson, one of the most widely beloved entertainers and profoundly influential artists of all-time, leaves an indelible imprint on popular music and culture.\\n\\nFive of Jackson's solo albums - \\\"Off the Wall,\\\" \\\"Thriller,\\\" \\\"Bad,\\\" \\\"Dangerous\\\" and \\\"HIStory,\\\" all with Epic Records - are among the top-sellers of all time and \\u201cThriller\\u201d holds the distinction as the larargest selling album worldwide in the history	<a href="http://www.facebook.com/michaeljackson">http://www.facebook.com/michaeljackson</a>

OK Cancel

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.



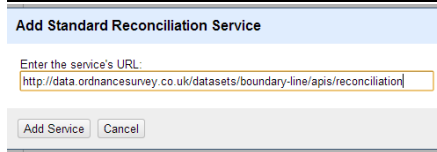
# Tutorial: OpenRefine

## **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>.



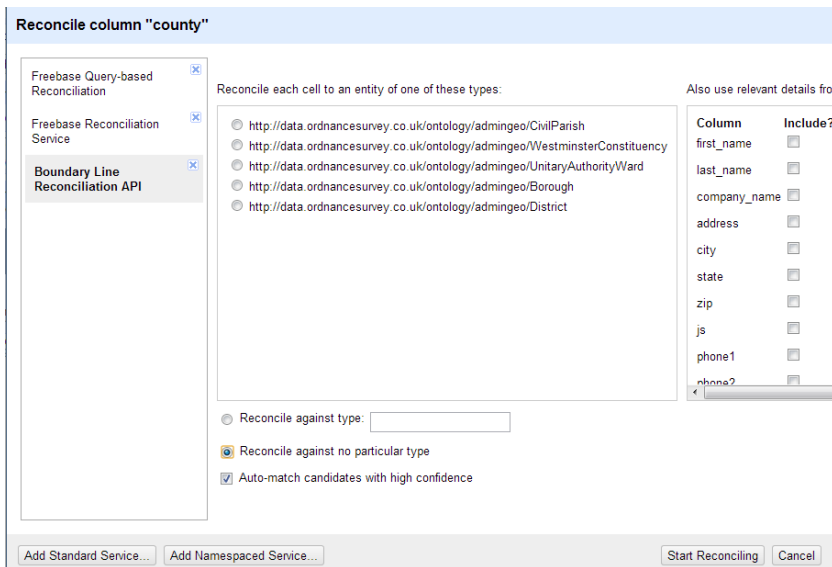
Add Standard Reconciliation Service

Enter the service's URL:

# Tutorial: OpenRefine

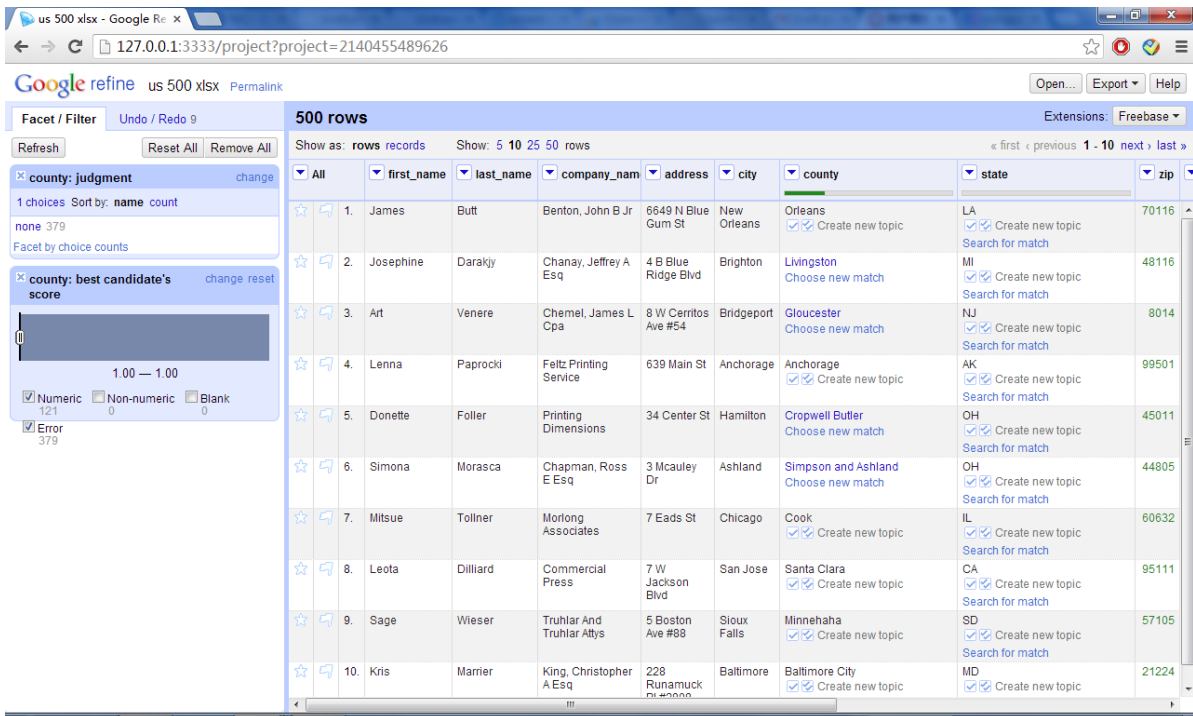
## Step 2

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



## Step 3

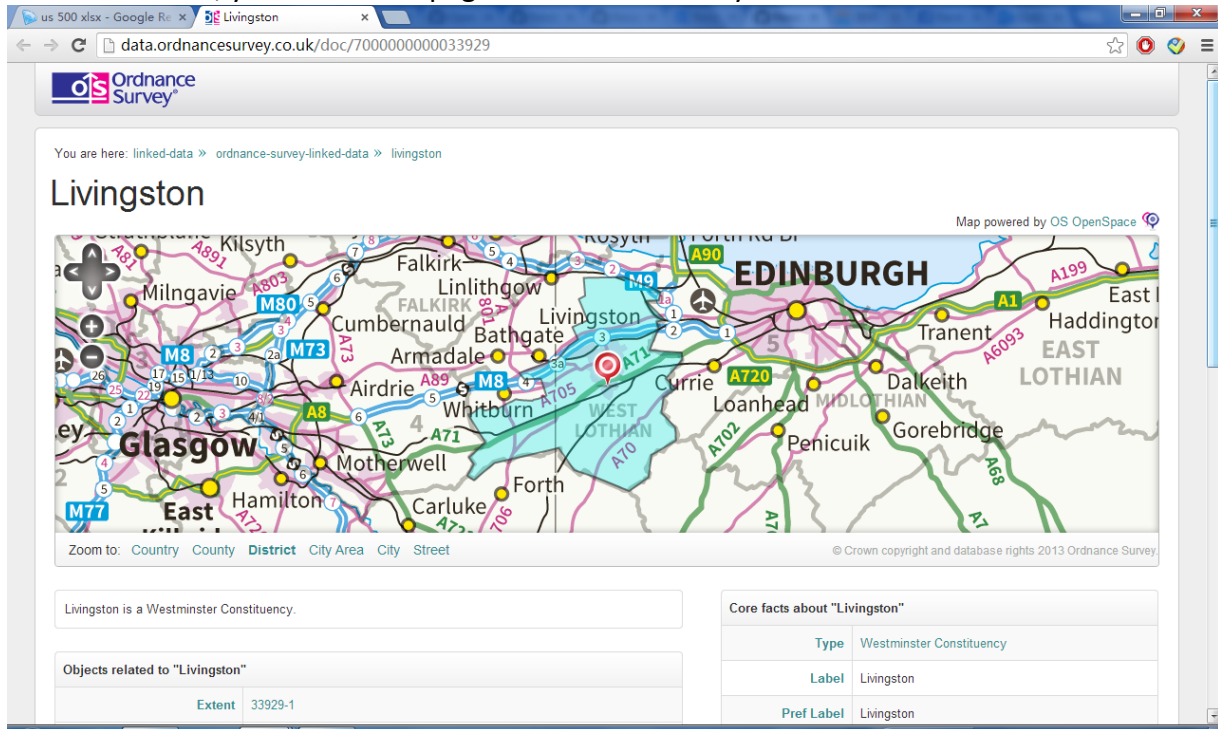
You should now see that most of the text labels have turned to hyperlinks.





# Tutorial: OpenRefine

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



The screenshot shows a web browser window displaying the Ordnance Survey website. The URL is [data.ordnancesurvey.co.uk/doc/7000000000033929](http://data.ordnancesurvey.co.uk/doc/7000000000033929). The page title is "Livingston". Below the title is a map of the Livingston area, showing roads, rivers, and surrounding towns like Glasgow and Edinburgh. The map is powered by OS OpenSpace. Below the map, there is a section for "Core facts about 'Livingston'" with a table:

Type	Westminster Constituency
Label	Livingston
Pref Label	Livingston

## 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.

# Tutorial: OpenRefine

## (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.

The screenshot shows the OpenRefine interface. At the top, it says "Google refine us 500 Copy xlsx Permi". Below this, there are two tabs: "Facet / Filter" and "Undo / Redo 4". The "Undo / Redo" tab is active. Below the tabs, there are two buttons: "Extract..." and "Apply...". Below the buttons, there is a "Filter:" label followed by an empty text input field. Below the input field, there is a list of five steps:

0. Create project
1. Text transform on 0 cells in column last\_name: value.trim()
2. Text transform on 12 cells in column county: value.toTitlecase()
3. Text transform on 500 cells in column last\_name: value.toUppercase()
4. Text transform on 500 cells in column first\_name: value.toLowerCase()
5. Text transform on 68 cells in column email: grel:value.split("@gmail.com").join("")

Step 4 is highlighted in blue, and step 5 is greyed out.

# Tutorial: OpenRefine

nalink Open... Export Help

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	504-821-8927	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	VERERE	Chemel, James L Cpa	8 W Cerritos Ave #54	Bridgeport	Gloucester	NJ	8014	856-636-6749	856-264-4130	art@venere.org	http://www
4.	lenna	PAPROCKI	Fetz 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 Atyhs	5 Boston Ave #88	Sioux Falls	Minnehaha	SD	57105	605-414-2147	605-794-4895	sage_wieser@cox.net	http://www
10.	kris	MARRER	King, Christopher A Esq	228 Runamuck Pl #2808	Baltimore	Baltimore City	MD	21224	410-855-8723	410-804-4694	kris@gmail.com	http://www

Windows taskbar: 12:43 20-11-2013

## (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.

Google refine us 500 Copy xlsx Perr

Facet / Filter **Undo / Redo 3** Extract... Apply...

Filter:

0. Create project
1. Text transform on 0 cells in column last\_name: value.trim()
2. Text transform on 12 cells in column county: value.toTitlecase()
3. Text transform on 500 cells in column last\_name: value.toUppercase()
4. Text transform on 500 cells in column first\_name: value.toLowerCase()
5. Text transform on 68 cells in column email: grel:value.split("@gmail.com").join("")



# Tutorial: OpenRefine

main

Open... Export Help

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	<a href="http://www">http://www</a>
	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	<a href="http://www">http://www</a>
	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	<a href="http://www">http://www</a>
	4.	Lenna	PAPROCKI	Feltz Printing Service 639 Main St	Anchorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com	<a href="http://www">http://www</a>
	5.	Donette	FOLLER	Printing Dimensions 34 Center St	Hamilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.net	<a href="http://www">http://www</a>
	6.	Simona	MORASCA	Chapman, Ross E Esq 3 Mcauley Dr	Ashland	Ashland	OH	44805	419-503-2484	419-800-6759	simona@morasca.com	<a href="http://www">http://www</a>
	7.	Mitsue	TOLLNER	Morlong Associates 7 Eads St	Chicago	Cook	IL	60632	773-573-6914	773-924-8565	mitsue_tollner@yahoo.com	<a href="http://www">http://www</a>
	8.	Leota	DILLIARD	Commercial Press 7 W Jackson Blvd	San Jose	Santa Clara	CA	95111	408-752-3500	408-813-1105	leota@hotmail.com	<a href="http://www">http://www</a>
	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	<a href="http://www">http://www</a>
	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	<a href="http://www">http://www</a>

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>