

Enriching Data

In this exercise we are going to use data from Open Corporates to enrich a government financial transactions dataset. The purpose of this exercise is to reveal the power of linked data and how you can go about creating linked data from tabular, transactional data.

The dataset we are going to be using is the Foreign and Commonwealth Office spend over £25,000 dataset.

<http://data.gov.uk/dataset/financial-transactions-data-fco>

In the making of the exercise the **April 2010** dataset was used.

The question we are going to try and answer is:

“How many of the companies the FCO dealt with in are still active, how many are dissolved and how many have been liquidated?”

Step 1 - Download and Pivot

For this exercise we do not need all the detail of the transactions. What is required is the total amount spent with each company. To filter and group the data we can use a pivot table, available via google docs (spreadsheet).

Once you have downloaded the data, upload it to google docs and create a google sheet.

Once loaded, from the **data** menu select **pivot table report**.

To create the pivot table which will group the data by company name, select the options specified on the right.

Once done export this pivot table as csv.

A	B	C
AAR ENVIRONN	28,000.00	
AL-RAMI MOTO	46,529.11	
ARTICLE 19	58,681.00	
ASSOC OF COM	34,948.03	
BRITISH COUNI	142,903.99	
BULLET MARKI	62,947.10	
BUYING SOLUT	112,740.63	
CABLE AND WI	133,585.75	
CAPGEMINI UP	193,098.09	
COFFEY INTER	213,173.91	
COGENT ELEC'	101,617.71	
CONCERTO CO	66,083.17	
DESIGN IT SOL	62,479.45	
DTZ CONSULTI	151,189.42	
ELEMENT ENEI	92,237.50	
ESSENT TRADI	740,400.00	
FISCAL CRIME	89,822.37	
FUJITSU SERVI	296,976.99	

Report Editor

'spend-april-10'1A1:I495 -
Edit range...

Rows - Add field

Group by: Supplier Name ×

Order: Ascending ▾

Sort by: Supplier Name ▾

Show totals

Columns - Add field

Values - Add field

Display: Amount (↕) ×

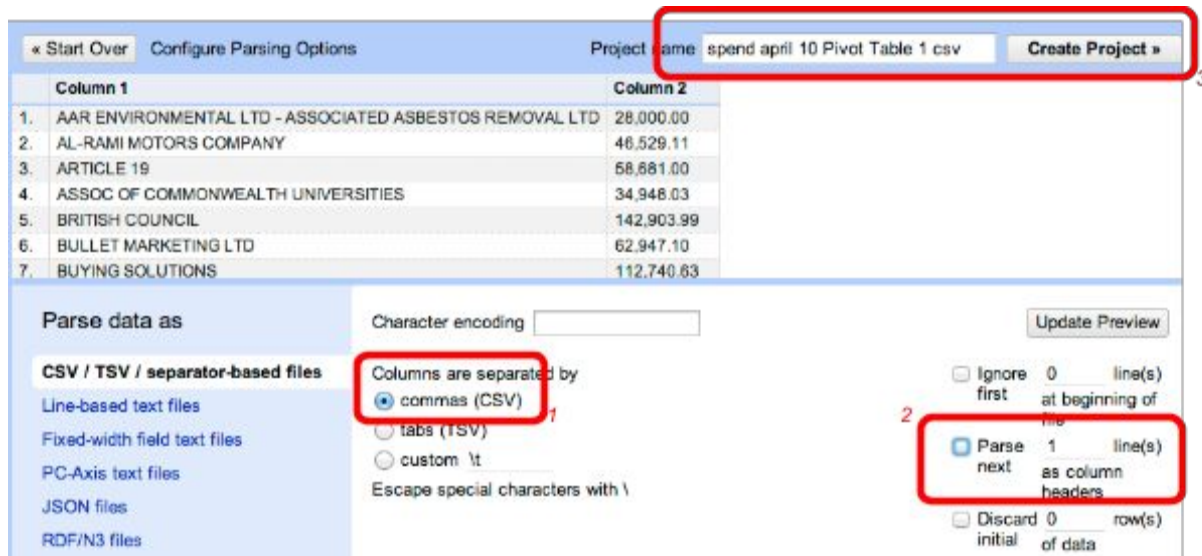
Summarise by: SUM ▾

Step 2 - Import into Refine

Having obtained an aggregated version of the data with many fewer rows, import this data into Google/Open Refine.

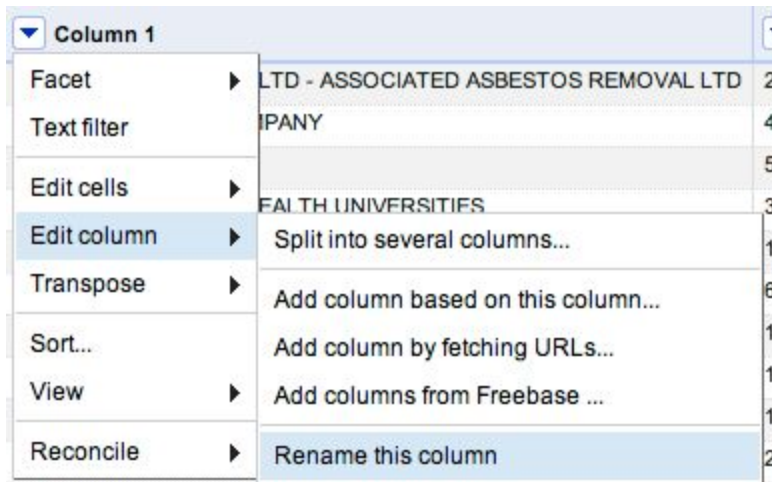


When you import the data, ensure that you do not use the first row as the column headers. The exported pivot table from google docs won't have any.



Step 3 - Reconcile and link

With the data now imported we have a column listing company names in plain text and amounts spent with each. Before we continue you might want to change the column titles to reflect the data.



In order to link the data to publicly available data we are going to use the Open Corporates reconciliation API. From the company name column drop down select the **Start Reconcile** option from the **Reconcile** menu. You will note that Open Corporates is not yet available as an option so a new **standard reconciliation service** will need to be added. The service URL is:

`https://opencorporates.com/reconcile`

Add Standard Reconciliation Service

Enter the service's URL:

Once done, select this service and click **Start Reconciling**.

When complete you will notice that each company may have several matches in Open Corporates, with each scored differently. You can click on each option to see a brief overview of that company and then tick which you believe is the correct option.

☆	🗨	3.	ARTICLE 19 <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> ARTICLE 19 (91) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> ARTICLE 19 FILMS, L.L.C. <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> THE ARTICLE 19 GROUP <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match	58,681.00
☆	🗨	4.	ASSOC OF COMMONWEALTH L <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match	
☆	🗨	5.	BRITISH COUNCIL <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> BRITISH COUNCIL(THE) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> THE BRITISH COUNCIL (C <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> BRITISH COUNCIL FOR C <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match	

Match this Cell
Match All Identical Cells
Cancel

ARTICLE 19 FILMS, L.L.C.

Status: **Active**

Company No: **3065152**

Registered: **2004-06-14**

Address: **239 CENTRE STREET, MANHATTAN, NEW YORK, 10013**

New York (US) - DOMESTIC LIMITED LIABILITY COMPANY

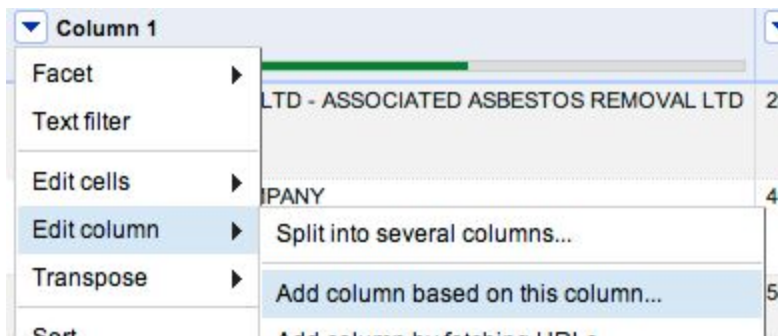
Rather than manually processing the entire dataset you can also filter to high or low quality matches using the facet on the left. You can also just choose to match all companies against their best candidate using the **match** option in the **reconcile** menus **actions**.

<div style="border: 1px solid #ccc; padding: 5px;"> Reconcile ▶ </div> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match </div> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> BRITISH COUNCIL <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> BRITISH COUNCIL <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> THE BRITISH CC <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> BRITISH COUNCIL <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic Search for match </div>	<div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> Start reconciling... </div> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> Facets ▶ </div> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> QA facets ▶ </div> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> Actions ▶ </div> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> Copy reconciliation data... </div>	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 70%;"></td> <td style="width: 30%; text-align: right;">34,948.03</td> </tr> <tr> <td style="width: 70%;"></td> <td style="width: 30%; text-align: right;">142,903.99</td> </tr> </table> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px; background-color: #fff9c4;"> Match each cell to its best candidate Match each cell to its best candidate in this column for all current filtered rows Create one new topic for similar cells </div>		34,948.03		142,903.99
	34,948.03					
	142,903.99					

Step 4 - Reveal the URI

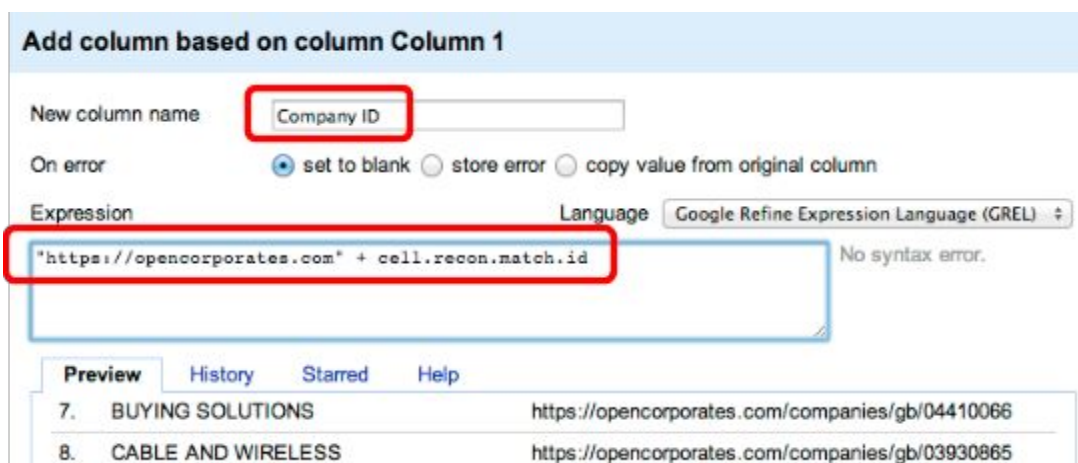
Once you have matched as many companies as possible you will notice that clicking on the company will now take you to that company's page in Open Corporates. This URL is also the company URI (identifier) and we can also reveal this in the dataset as follows.

From the company column drop down select **Add column based upon this column**.



In the box that appears, give the column a name of **Company ID** and type the following in the expression box:

```
"https://opencorporates.com" + cell.recon.match.id
```



Preview	History	Starred	Help
7.	BUYING SOLUTIONS	https://opencorporates.com/companies/gb/04410066	
8.	CABLE AND WIRELESS	https://opencorporates.com/companies/gb/03930865	

When done click OK. We could now export this dataset as CSV, XML, RDF etc, and it would contain a URI from which further data could be obtained. This is a simple but effective example of linked data and linked csv.

Step 5 - Get the data

We now have a link to the Open Corporates page for each company that lists data about that company. It is the current status that we need in order to answer the original question. In order to fetch it for the companies we shall use the Open Corporates API in order to download the data for each company.

This stage involves adding a **column by fetching URLs**. Although we could fetch the URI from step 4 using content negotiation, Refine does not understand HTTP redirection thus we would not get the data back from the URI. For this reason it is necessary to cut a corner and download the data directly from where the redirect would have sent Refine.

Enter the following in the **Expression** box:

```
"https://api.opencorporates.com" + cell.recon.match.id + ".json"
```

When required to use an **API key**, enter the following expression replacing the *APIKEY* with the one made available by the course leader:

```
"https://api.opencorporates.com" + cell.recon.match.id +  
".json?api_token=APIKEY"
```

Add column by fetching URLs based on column Column 1

New column name Throttle delay milliseconds

On error set to blank store error

Formulate the URLs to fetch:

Expression Language No syntax error.

Make sure you name the column and turn down the **throttle delay** so that this process completes in reasonable time. The throttle delay is there so we don't overwhelm an API with requests.


Once done click OK and you should end up with a column full of JSON data from which we need to pick out the **company status**.

In order to see what the JSON data looks like why not take one of the returned values and paste it into the JSON validator at jsonlint.com.

Step 6 - Extract the company status

One last column to add based upon the data column this time with the following expression to parse the JSON data:

```
value.parseJson()["results"]["company"]["current_status"]
```



Add column based on column OC_Data

New column name:

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

Expression: Language: Google Refine Expression Language (GREL) ⌵

No syntax error.

Preview History Starred Help

3.	...	Active
	{"api_version": "0.3.1", "results": {"company": {"name": "ARTICLE 19", "company_number": "02097222", "jurisdic": "02-05", "dissolution_date": null, "company_type": "(Private, Limited by guarantee, no share capital, use of 'limited exemption')", "registry_url": "http://data.comps	

Once complete you might want to **collapse** the data column so that you can see more rows on the screen. It is then possible to apply facets to find out how many companies are still active, liquidated, dissolved or other.

Extension Exercises

Why not try extracting data other than current status from the Open Corporates data?

Can you link to any other data available from Open Corporates or suppliers of other reconciliation endpoints?

How about a visualisation of the data?