Enriching Data (Singapore edition)

In this exercise we are going to use data from OpenCorporates to enrich a dataset containing company data.

The dataset for this exercise was located by using an advanced google search containing the following: “site:sg filetype:xls contracts”. This located the following dataset:

www.citygas.com.sg/pdf/AGENT_LISTING.xls

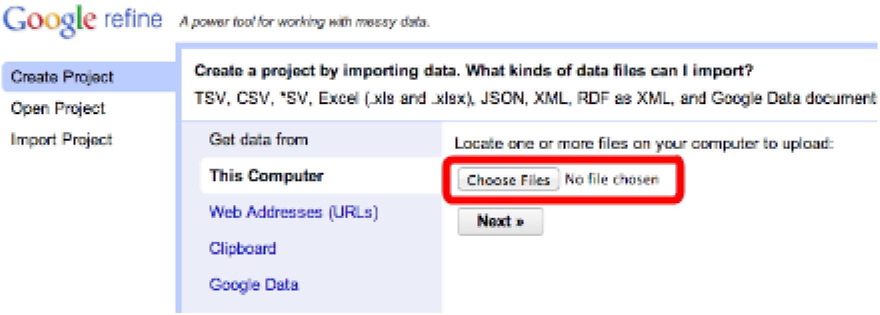
This appears to be a listing of agents of the company that was created in November 2006 (according to the file properties).

The question we are going to try and answer is:

“How many of the agents are still active, how many are dissolved and how many have been liquidated?”

Step 1 - Import into Refine

Simply import this data into refine as a new project. Ensure that the formatting looks correct before clicking create project.



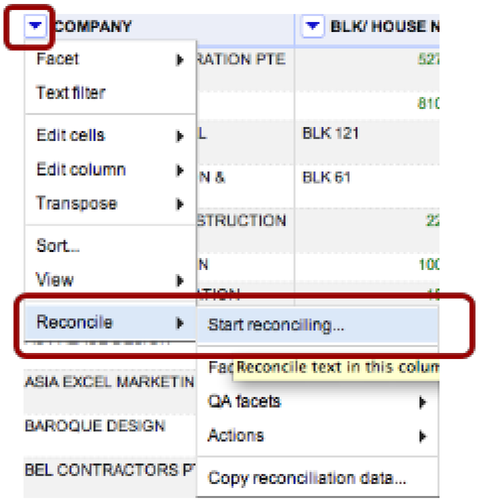
Create a project by importing data. What kinds of data files can I import?
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data document

Locate one or more files on your computer to upload:

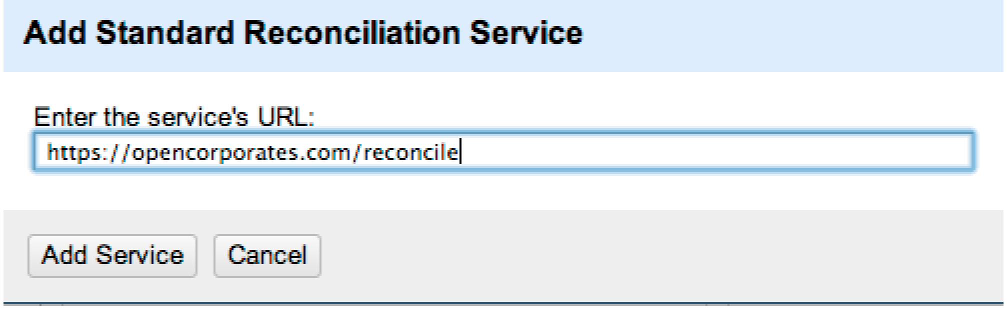
Step 2 - Reconcile and link

With the data now imported we have a column listing company names in plain text.

In order to link the data to publicly available data we are going to use the OpenCorporates reconciliation API for Singapore. From your ‘company’ column drop down, select the **Start Reconcile** option from the **Reconcile** menu. You will note that OpenCorporates is not yet available as an option so a new **standard reconciliation service** will need to be added. The service url is as follows, please ensure you copy this exactly (including the https):



`https://opencorporates.com/reconcile/sg`

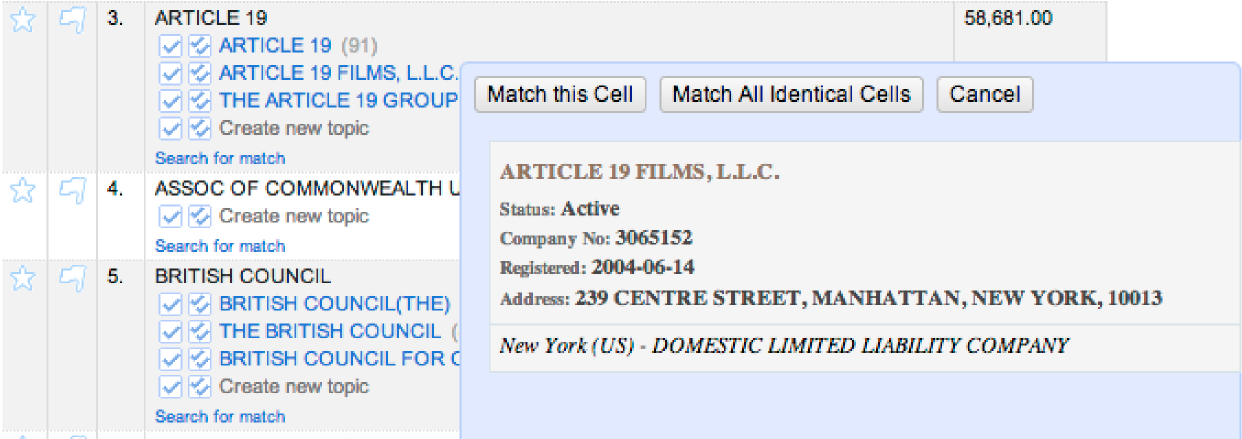


Add Standard Reconciliation Service

Enter the service's URL:
https://opencorporates.com/reconcile

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

When complete you will notice that each company may have several matches in OpenCorporates, 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.



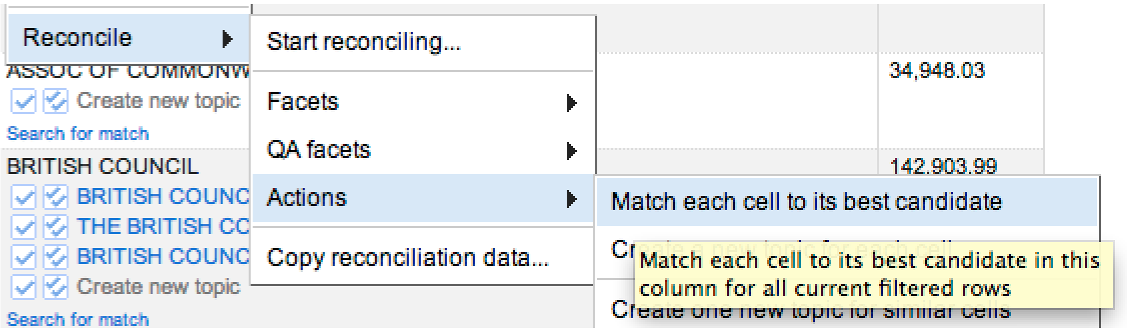
3. ARTICLE 19 — 58,681.00
ARTICLE 19 (91)
ARTICLE 19 FILMS, L.L.C.
THE ARTICLE 19 GROUP
Create new topic
Search for match

4. ASSOC OF COMMONWEALTH U
Create new topic
Search for match

5. BRITISH COUNCIL
BRITISH COUNCIL(THE)
THE BRITISH COUNCIL (
BRITISH COUNCIL FOR (
Create new topic
Search for match

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 process 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 by selecting the **reconcile** menu, followed by **actions** and



ASSOC OF COMMONW — 34,948.03
Create new topic
Search for match
BRITISH COUNCIL — 142,903.99
BRITISH COUNC
THE BRITISH CC
BRITISH COUNC
Create new topic
Search for match

Match each cell to its best candidate in this column for all current filtered rows

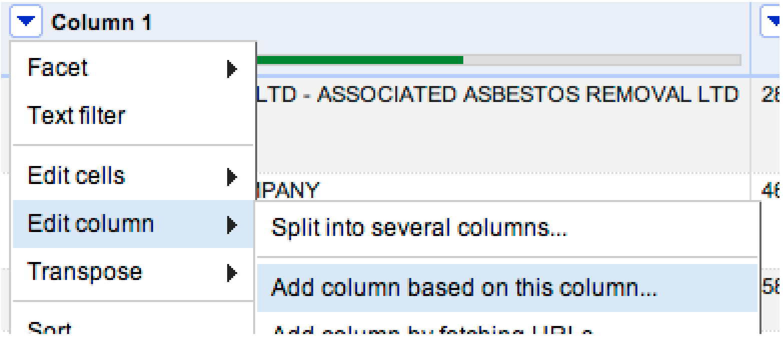
Reconcile
Start reconciling...
Facets
QA facets
Actions
Copy reconciliation data...
Match each cell to its best candidate
Create one new topic for similar cells

then **match.** .

Step 4 - Reveal the URI

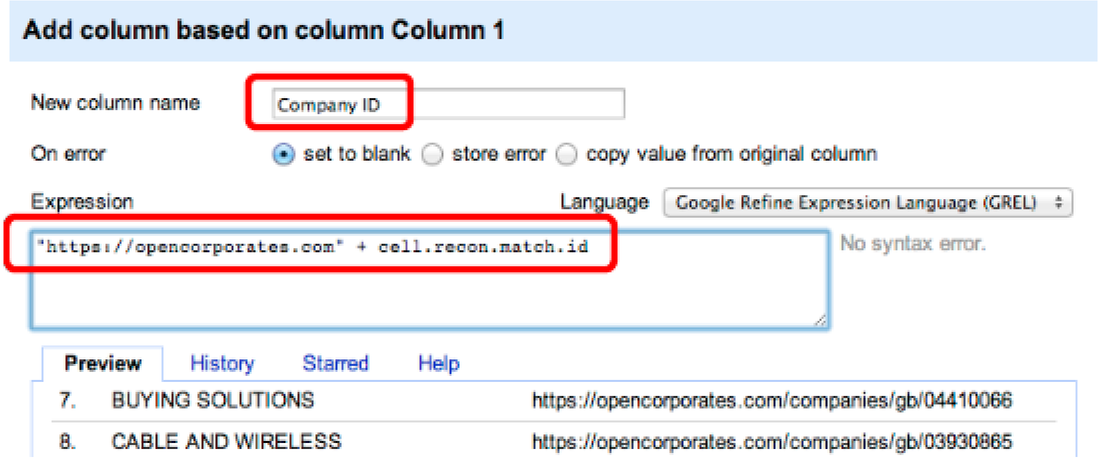
Once you have matched as many companies as possible you will notice that clicking on the company will now take you to that company’s page in OpenCorporates. This URL is also the company URI (identifier) and we can also reveal this in the dataset as follows.

From the ‘company’ column drop down select **Edit column** then **Add column based upon this column**.


Column 1
Facet
Text filter
Edit cells
Edit column
Transpose
Sort
LTD - ASSOCIATED ASBESTOS REMOVAL LTD
IPANY
Split into several columns...
Add column based on this column...
Add column by fetching URLs...

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
```



Add column based on column Column 1

New column name: Company ID

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

Expression — Language: Google Refine Expression Language (GREL)

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

No syntax error.

Preview | History | Starred | Help

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

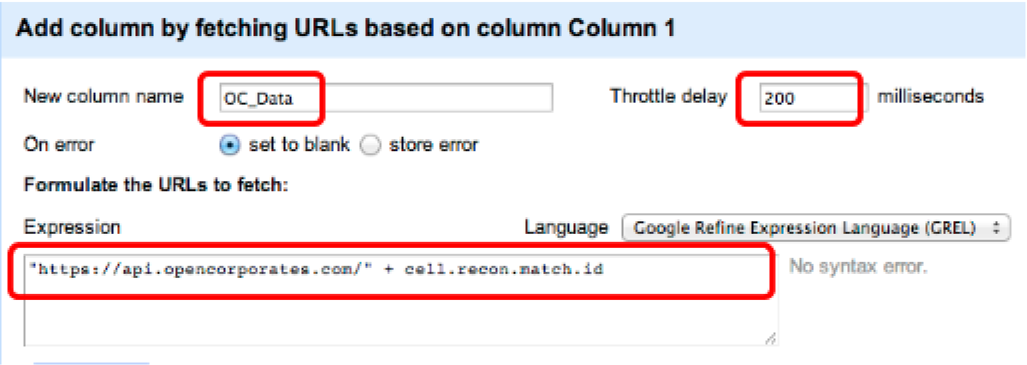
When done click OK. We could now export this dataset as CSV, XML, RDF etc and it would contain a URI from which further data could be obtained. This is a simple but effective example of linked data and linked csv.

Step 5 - Get the data

We now have a link to the OpenCorporates' page for each company, listing data about that company. It is the current status that we need in order to answer the original question. We will use the OpenCorporates API to download this 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, so 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.

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



Add column by fetching URLs based on column Column 1

New column name: OC_Data

Throttle delay: 200 milliseconds

On error: ● set to blank ○ store error

Formulate the URLs to fetch:

Language: Google Refine Expression Language (GREL)

Expression:

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

No syntax error.

Name the column 'OC_Data' and turn down the **throttle delay** to 200 milliseconds 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, take one of the returned values and paste it into the JSON validator at **jsonlint.com**.

Step 6 - Extract the company status

Add one last column, based upon the 'OC_Data' column this time with the following expression. This will extract just the company current status from the data.

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

Once complete you might want to **collapse** the ‘OC_Data’ column so that you can see more rows on the screen. It is then possible to **apply a text facet on the new column** (available from the column drop down) to find out how many companies are still active, liquidated, dissolved or other.

Extension Exercises

Why not try extracting different data from the OpenCorporates data, other than company status?

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

Can you produce a visualisation of the data?