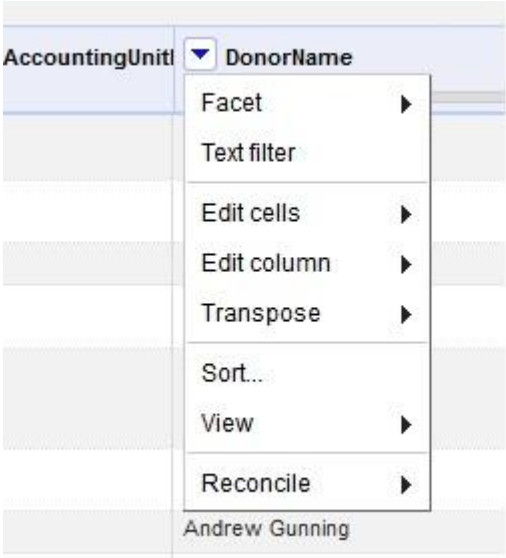
Reconciling in OpenRefine

A powerful feature built into OpenRefine is the reconciliation function. This is a way of automating the searching and matching data which are in one column of your dataset with a much bigger database online – the example given in one tutorial is matching a list of films which have won Oscars, to the details about those films on the Internet Movie Database, IMDb. A more common use within investigative journalism is to match a list of companies from, say, a political donation register, to the company details on OpenCorporates.com.

To reconcile any dataset you have opened (otherwise known as a "project") in OpenRefine with an online database you need to set up, one time, an endpoint. Some are already built into OpenRefine, but most of these are deprecated and need replacing! (I know this from many hours of spinning wheels on my laptop – reconciliation not working because the endpoint no longer existed!).

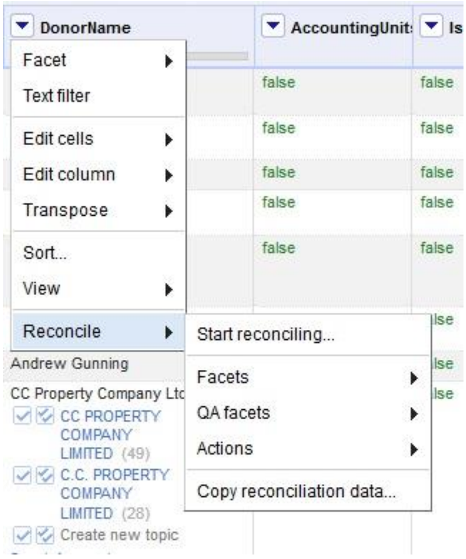
To set up the endpoint for OpenCorporates you need to copy this string https://opencorporates.com/reconcile

Then, with OpenRefine running, and with any dataset open on the screen (ie you don't have to do this with the data you're really going to analyse, but can set it up ahead of time), click the dropdown arrow at the head of any column –

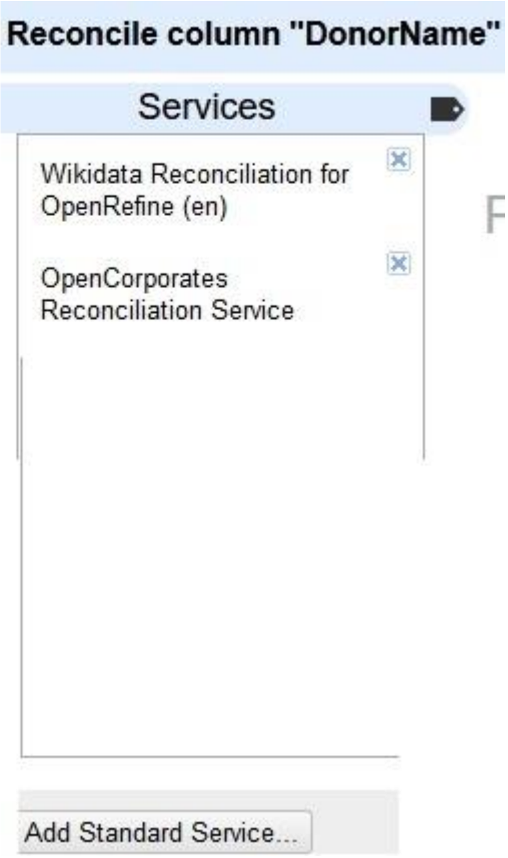


Choose "reconcile at the bottom of the menu.

Then “start reconciling”



This makes Refine do some preliminary processing, then it asks you to choose the service you’re going to use


Reconcile column "DonorName"
Services
Wikidata Reconciliation for OpenRefine (en)
OpenCorporates Reconciliation Service
Add Standard Service...

If OpenCorporates is not listed - which it won't be if you're using Refine for the first time – click "Add Standard Service" at the bottom of the page. You will be asked for a url – and you paste the opencorporates one you copied earlier. Close the dialogue, and OpenCorporates will be there next time you use the program, forevermore.

More detailed documentation is available via the "help" link on the OpenRefine main screen

This takes you to

https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users

and a quick word search for "reconcile" will take you here:

https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API

Other documentation is around on the web – I google "openrefine reconcile filetype:pdf" to get some of it. The endpoints are changing all the time – some disappearing, and new ones being created. You can of course build your own!

To take a good example of a dataset you can reconcile with opencorporates go to

Search.electoralcommission.org.uk and select donation data for a year or more (if you're interested in clustering dirty data, less if you only want to reconcile…..

This will give you a dataset containing, among other things, names of companies which have donated to political parties and MP's. Using the instructions above, you should be able to reconcile the "donor name" column with opencorporates.