Preparation for Module 3

Cleaning data

To allow us more time to talk and try out some of the techniques you may need to use to prepare data for analysis, we ask you to watch these videos as preparation for session 3.

Rather than spending a whole live session being bombarded with new information, you can come prepared, possibly having had a chance to try some of the techniques for yourself, or ready to ask questions and solve them in the session.

Links to the videos and the material used in them are as follows:

Subject	Purpose	Duration
1. ImportHTML	To convert a table from a web page to a spreadsheet, using Google Sheets	7.02
2. Clean with Open Refine	To learn how to clean names in a dataset where slight differences make the computer see them as different entries	13.37
3. Export your OpenRefine project	How to convert your data after cleaning in OpenRefine back to Excel/Google sheets. The source table is this Wikipedia page if you want to follow the same steps	3.16
4. Reconcile with OpenRefine	Matching names in a column in OpenRefine with company names on OpenCorporates.com NB – to make use of the OpenCorporates reconciliation service you need to add this address to the reconciliation menu in OpenRefine https://opencorporates.com/reconcile (see 1.18" in this video). See here for documentation	7.53
5. Tabula	Convert pdf file to csv spreadsheet with Tabula	5.09
6. Merging data using VLOOKUP formula	Enhance one data by matching values from another to add extra information	8.43
7. Cleaning names with OpenRefine	Removing extras such as "Mr", "Mrs", "Dr, and creating a reproducible script to save time doing the same job on future occasions	13.37
8. Power Query	A relatively new Excel feature – which allows you to merge data without learning VLOOKUP. It also keeps a record of what you have done, so you can check or reproduce your work – eg when you use a later edition of the same data.	6.39

Practice material

- To practice "6. VLookup" you need to download a copy of this worksheet containing the two datasets.
- If you want to practise cleaning the MPs names (removing titles etc) using OpenRefine as in the video "7. Cleaning names with OpenRefine, you will need to download a copy of this dataset
- To get a copy of the company donation data demonstrated in nos 2 and 4, follow this link and download a copy

- To practice the Power Query demo with the same data you will need to download copies of these two datasets – the donations to MPs and the list of MPs elected in 2019