SPREADSHEETS – A HANDY GUIDE FOR JOURNALISTS

If you're getting started in Datajournalism, you will need to get to grips with spreadsheets as the first basic tool. As the Open Data movement grows the amount of data available on the web for download and analysis increases daily (more later on where to find them). Most datasets are in a format which makes them readable by Excel, and by Google Spreadsheets, to name the obvious ones. A lot of the following assumes that you will be using Excel, but almost everything you need to get started works in Google Spreadsheets. (One thing about Excel is the way that each new version has a new set of menus, so the following is written with Excel 2010 in mind. There are some separate screenshots for the relevant instructions in Excel 2003 (still in use in many parts of the BBC, for example) at the end, but it's worth saying, that if you're going to get to know Excel in any depth, then 2010 is a huge step forward and it's worth you working with that from the beinning.

Many journalists who have never opened a spreadsheet before are confused by the array of rows and columns they see in front of them. It may help to think of a dataset as an interviewee with total recall of a set of facts and figures; all you need to do is to work out what to ask. The obvious questions – "what's the oldest/newest/biggest/smallest/most expensive?" can be answered by knowing about three main functions – sorting, filtering, and pivoting. So that's what the rest of this chapter (and the morning of the BBC College of Journalism course "Sources, Scoops &Spreadsheets") aims to help you to do.

If you're new to datajournalism and spreadsheets, it may help at this point to download and open one of the thousands of datasets you can find on the web now. A good place to start is with council spending. All councils and central government departments now upload details of their spend on items over £500 (in the case of bigger bodies and agencies, £25k is the minimum).

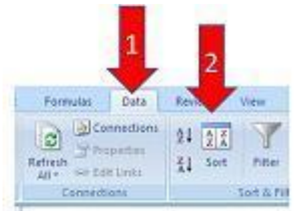
Finding a dataset couldn't be easier – and since it would be invidious to single out one council or agency – so all you need to do in order to find a sample dataset is to google the following (NB no spaces or full stops after the colon)

Filetype:xls site:gov.uk

You won't need any other search term unless there's a particular subject body or agency which interests you. If you want a bit of structure to your search, you could go via a portal like the Guardian's Data blog or the government's data.gov.uk instead.

Sorting

When you open the dataset, have a look across the rows to see what you've got – dates, spends, departments, budget codes, etc. A good first step is to sort one of the columns – largest to smallest, a-z, etc. So you're asking and quickly getting answers to questions like "what's the biggest single payment?"



Finding the SORT menu

To sort – click the “Data” Tab, highlight one cell (preferably in the column you’re sorting) and then click sort. In the dialogue box choose which column you want to sort, what values, and what order (eg largest to smallest) you want them in.

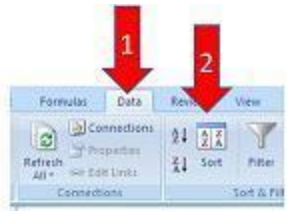
If your columns have a title in the first row and Excel hasn’t done it already, tick the “my table has headers” box and the column selector will display the relevant names instead of Col A, B etc.

Have a play by sorting the various columns – see if anything jumps out at you. One of the nice little features of Excel 2010 is that if you highlight a whole column of figures, then automatically in the bottom right hand corner of the Excel screen you will see the total number of entries in that column, the sum of the entries, and the average. You can’t copy that display except using pen and paper, but without any other keystrokes you will be able to grab, say, the total spend, or the average spend, or the number of items in a column for your article. (Later, when you filter a column, bear in mind that the totals at the bottom of the screen will be for the items currently displayed rather than everything in that column.)

Entering data - The cursor’s changing appearance

The big fat white cross – is just excel’s main cursor, it just helps you move to where you want to be Move it to the bottom right corner of a cell and it becomes black cross. Drag it down a column, or across a row and it will copy the values, or formula into each row you highlight. It will also fill in sequences like “1,2,3”, “2,4,6” or “Monday, Tuesday, Wednesday” – just highlight the first two cells in the pattern and it will continue the same pattern. Right click when you have the black cross, and you get a helpful menu . If the black cross develops into an arrow cross – beware – you are about to drag one cell to somewhere else. If you didn’t want to do that, click away or press “escape” to get back to the general purpose white cross.

Sorting (made easier if you give your columns a title in the first row “a header row”)



Finding the SORT menu

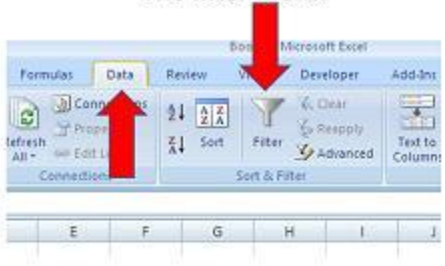
To sort – click the “Data” Tab, highlight one cell (preferably in the column you’re sorting) and then click sort. In the dialogue box choose which column you want to sort, what values, and what order (eg largest to smallest) you want them in.

If your columns have a title in the first row, tick the “my table has headers” box and the column selector will display the relevant names instead of Col A, B etc.

Filtering

Sorting a column will only get you so far in your analysis. You will soon find that your interrogation of a dataset leads you into filtering columns so that you only see certain entries, and then work out their totals, averages, or just see how many there are.

The filter menu



In the data tab you “switch on” the filter. This will be marked by a little dropdown symbol at the top of each column. You can then ask excel to display only the rows with the value or values you have selected in that column. When you click on a filter down arrow at the top of a column you will see a list of all the items in that column in alphabetical and/or numerical order. You can unclick “select all” and then just click the ones you want to see – just one department’s spend, or just one company’s invoices, for example. OK the selection, and then the column will show only the items you have selected. You can then sort or sum just those entries.

Excel 2010 has some powerful tools hidden in the filter menu as “text filters” – you can ask it to filter using “contains” – so for example if you had a list of companies with “British” somewhere in their name, but not necessarily the first word, you could filter by “contains”= British.

NB Remember to “turn off” a filter if you don’t need it when doing the next calculation.

PIVOT TABLES

You should soon see that you can do quite a lot of useful analysis just by sorting and filtering, but at some point you will realise that you could do some things much quicker if you had a few more tools at your fingertips. Pivoting tables is a key tool that, once you get the hang of it, will save you hours of repeating the same operation to get vital figures.

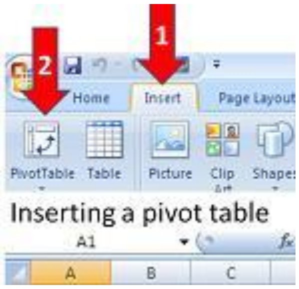
Pivoting table – as the name suggests – is a way of turning a table around – making one column into a row or vice versa.

A pivot table can help transform endless rows and columns of numbers into a meaningful presentation of the data. For example, a pivot table can count how many times a the same value appears in a given column or row. You can also display subtotals and any level of detail that you want. After you create a pivot table, you can rearrange the information in almost any way imaginable and even insert special formulas that perform new calculations.

It may not sound too exciting, but if you turn a list of members of a group into the title column of a table, you can then re-arrange the data associated with them to appear in ways which makes it easier for you to spot new patterns, and from them, stories.

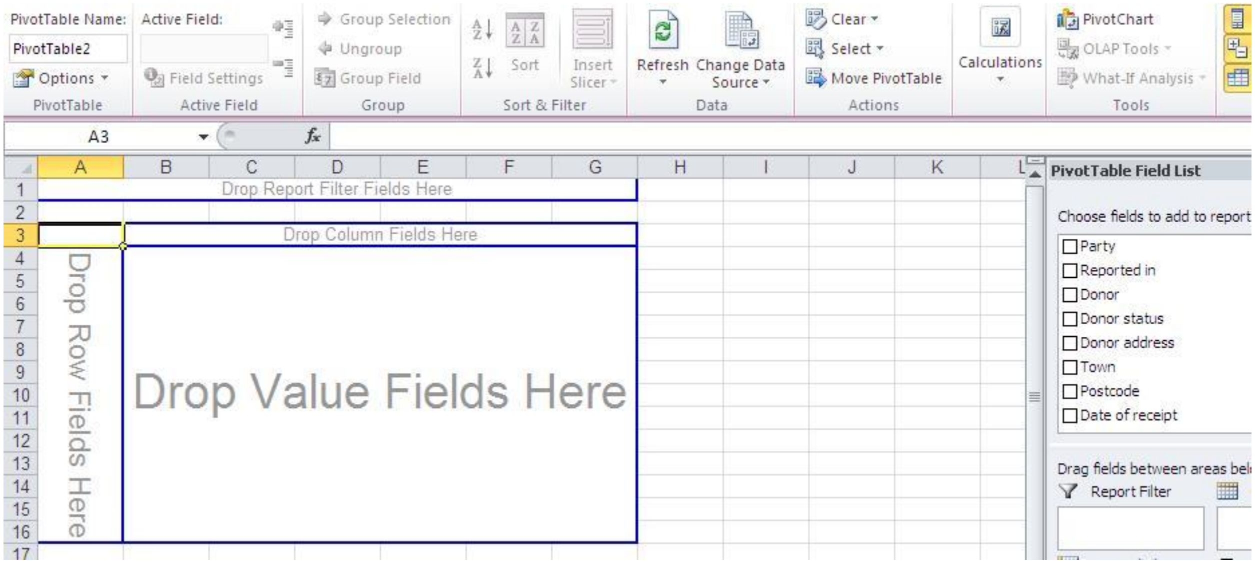
It's easier if you play with a real life example. Donations to political parties, for example. You can find that kind of dataset at the Electoral Commission website – (it's easier to use the page http://search.electoralcommission.org.uk than the main landing page). Then follow the links to get the data you want. Or follow this link

Download the data, open the spreadsheet, and go to the worksheet tab called "donations". Highlight any cell in the body of the spreadsheet, then use the insert menu to create a pivot table as shown;



Make sure you tick the box saying you want the pivot table in a new worksheet, not the existing one, and make sure you select pivot **table**, not **chart**.
Excel can be a bit picky about what it will accept for pivoting, so you may have to work around various error messages about cell sizes being the same. If you get stuck, start again.

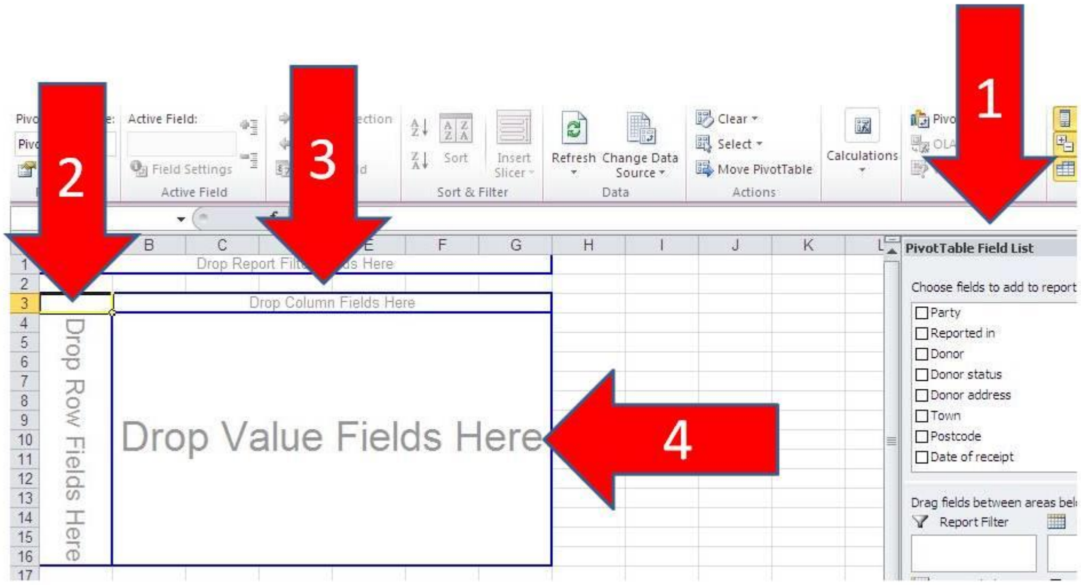
If it all works you should get a menu like this –



Which can be a bit daunting. Don’t panic. If you try these four steps you will soon see the power of pivot!

You need to select one of the data fields from the menu in 1, and drag it (don’t use the tick box next to it) where you want it. A good start would be to drag “party” into the area marked 2 (Drop row fields here”). Then take “donor status” and drag it into area 3 (Drop column fields here). Finally, drag “amount” into 4 (Drop value fields here). If it has worked, you should instantly be able to see how much each party received from each category of donor – individual, trade union, etc. Totals for the parties will be at the extreme right of the row, and the totals for each category are at the bottom of each column.

Now try dragging the grey boxes with the labels “party”, “donor status” and “amount” back to the field store at 1, and try again – it usually makes sense to decide in advance what row and column headers you want, then drop the values into the big central area – 4 at the end. Sometimes there will be too many columns and you won’t be able to see much on the screen without scrolling. Fine – start again and make those fields the rows, so they read downwards. Pivot tables are very flexible and save hours (for example if you wanted to add up what each party received from individual donors, a pivot will be inifinitely quicker than filtering for one party, adding the total, filtering for the next party, adding that total, and so on!)



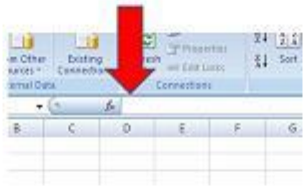
If you only master sorting, filtering, and pivoting, you will find datasets a whole lot easier to handle. There are a couple of other things you ought to know.

REMINDER Entering data - The cursor's changing appearance
The big fat white cross – is just excel's main cursor, it just helps you move to where you want to be Move it to the bottom right corner of a cell and it becomes black cross. Drag it down a column, or across a row and it will copy the values, or formula into each row you highlight. It will also fill in sequences like "1,2,3", "2,4,6" or "Monday, Tuesday, Wednesday" – just highlight the first two cells in the pattern and it will continue the same pattern. Right click when you have the black cross, and you get a helpful menu . If the black cross develops into an arrow cross – beware – you are about to drag one cell to somewhere else. If you didn't want to do that, click away or press "escape" to get back to the general purpose white cross.

Formulae
Not as scary as they sound. Just adding cells taken from two columns (eg January's income and February's need a formula. You need to decide where you want the results of your sum to appear, let's say cell A2. Start with a = in the f(x) field to tell Excel you're entering a formula. Then you tell it which cells you want it to take the data from – for example B2+C2. So looking in the formula bar the entry in cell A2 will now read =B2+C2, while down in the cell itself (once you press "enter" to move the cursor out and perform the operation) the cell will contain the result of the sum of cells B2 and C2. If you copy

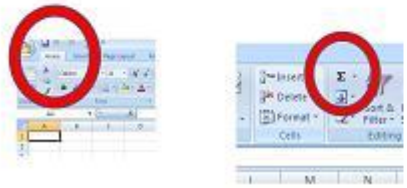
the formula down the column using the little black cross at the bottom right of the cell, then Excel will assume you want the formula to adapt to take values in each row – so it will become B3+C3, B4+C4, etc as you move down the row.

The formula bar



+ and – are themselves. /=divide, *=multiply. Remember to use brackets if you want excel to perform one calculation ahead of another – otherwise it goes in strict left to right. So E2+C2/D2 is not the same is E2+(C2/D2)

Finding the "sum/average/median"



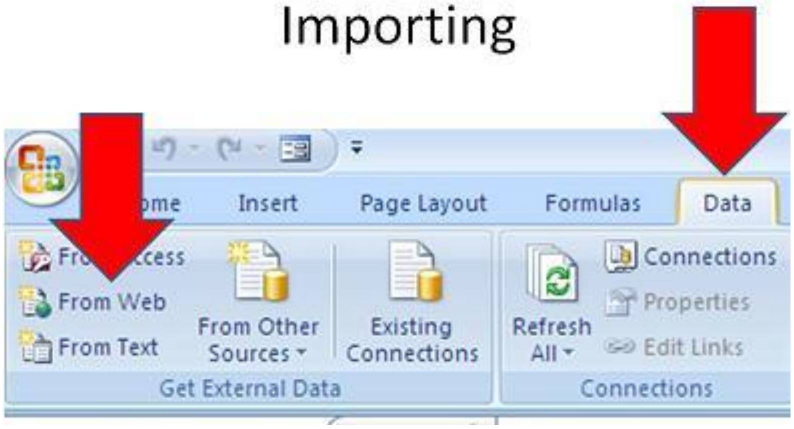
Excel will add up whole rows or columns – if you use the Σ button at top right of the main menu. It has a dropdown which also allows you to select average, means, etc.

If you copy formulae in each cell down a column, excel assumes you want to add the relevant cells in each row. So B2+C2 will become B3+C3 and then B4+C4 and so on.

If you want the formula to make use of the same cell all the way down a column, you need to put $ in front of the column letter and/or the row number.

Excel can also work with dates in a formula. So if you select a column with 23/5/11 and subtract it from 28/5/11, excel will give the result as 5 days.

Importing

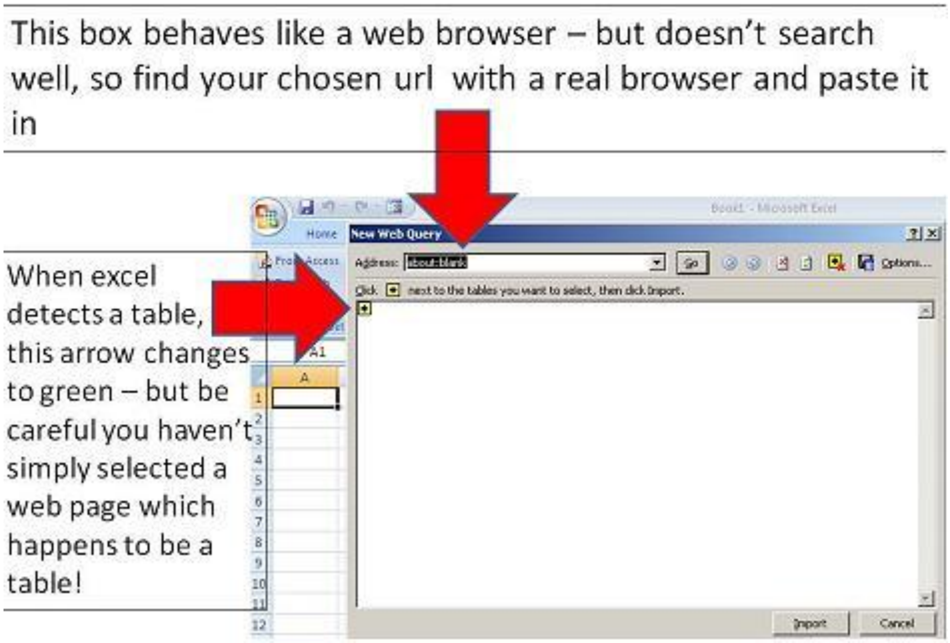


Importing – Excel* will save you hours of typing by pulling whole tables of data into a spreadsheet so you can analyse it. It takes a bit of trial and error to find tables on the web which are suitable for this, but when you find them, Excel is invaluable. (Google Spreadsheets –has a function called “Googlelookup” which will do the searching for you – but can also be a bit hit and miss)

**These screenshots are for versions of Excel between 2010 and 2013. For later versions, see Appendix below. The appendix also includes a short guide to scraping/importing data using GoogleSheets*

A good site to start experimenting on (but by no means the only one, or the best) is the register of parliamentary interests, the Lords for example - www.publications.parliament.uk/pa/ld/ldsecret/memi01.htm

Otherwise, you could just navigate to a table in Wikipedia in order to practice importing it.



This box behaves like a web browser – but doesn’t search well, so find your chosen url with a real browser and paste it in

When excel detects a table, this arrow changes to green – but be careful you haven’t simply selected a web page which happens to be a table!

New Web Query

Address: about:blank

Click ▸ next to the tables you want to select, then click Import.

Keyboard shortcuts are very handy – sometimes the mouse lets you do things you didn't mean to.

Ctl + C= copy	Ctl+V= paste
Alt I C =inserts column	Alt I R =insert row
Alt ED = deletes the highlighted cells	

Sources of datasets to get you started;

Guardian data log http://www.guardian.co.uk/data

Google http://code.google.com/p/google-refine/wiki/DataSources

UK government http://data.gov.uk/data

http://wheredoesmymoneygo.org/ visualizes UK government spending per person

http://www.whatdotheyknow.com/ FOI request aggregator

http://www.theyworkforyou.com/ - MP aggregator

Get the data http://getthedata.org/

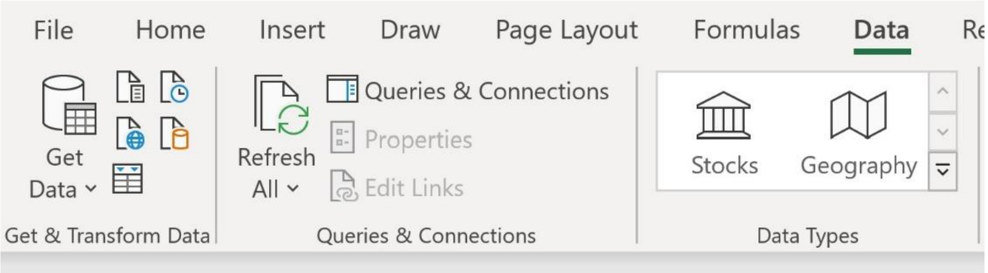
Jonathan Stoneman (jonathan.stoneman@gmail.com)

Appendix

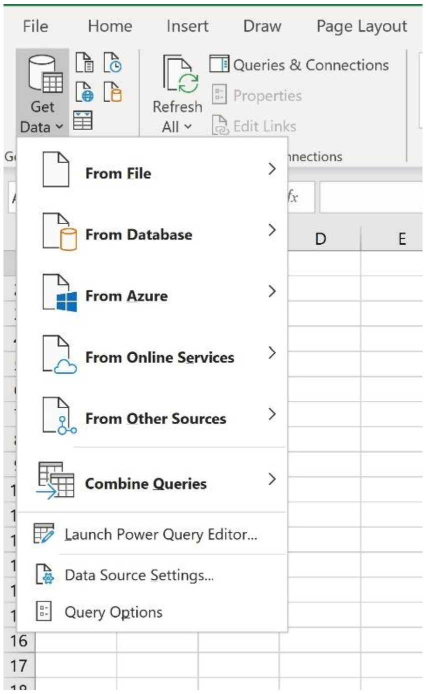
Current Excel steps for importing data from websites

Since 2013 Excel has changed it's import menu several times. Since 2016 the steps needed are as follows:

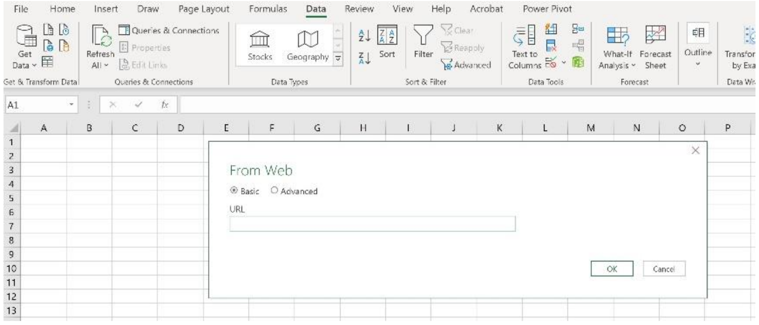
Click the "data" ribbon – at far left of the menu this appears



Click on “Get Data” to get this menu



Select – “from other sources” and select “from web” to get this dialogue box:

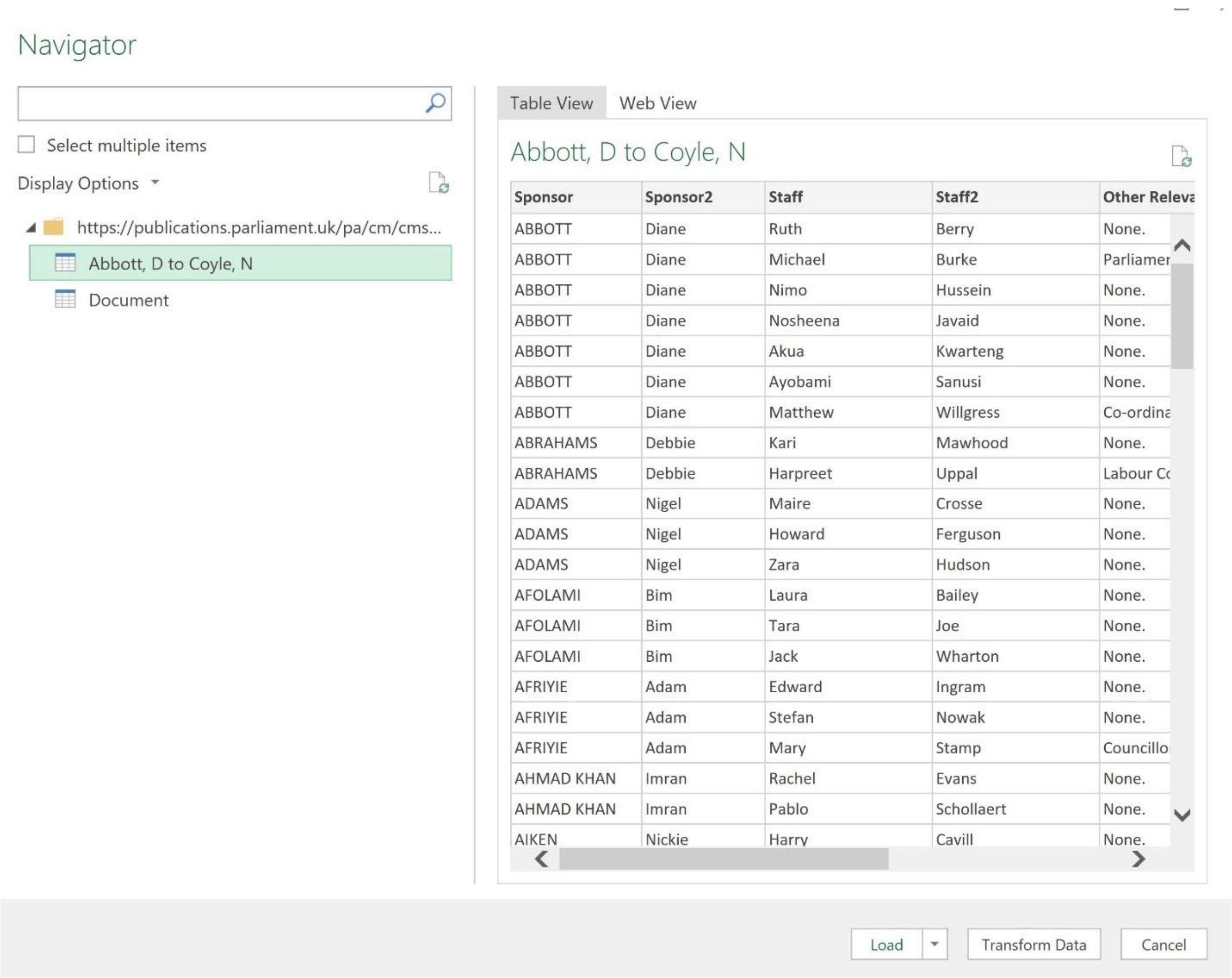


Paste your chosen web url into the relevant box, click OK and follow the steps until your chosen dataset appears in the preview (this may take a few clicks using trial and error).

For example copy and paste this url:

https://publications.parliament.uk/pa/cm/cmsecret/sponsor-01.htm

After a few seconds you get this dialogue box –



Abbott, D to Coyle, N

Sponsor	Sponsor2	Staff	Staff2	Other Releva
ABBOTT	Diane	Ruth	Berry	None.
ABBOTT	Diane	Michael	Burke	Parliamer
ABBOTT	Diane	Nimo	Hussein	None.
ABBOTT	Diane	Nosheena	Javaid	None.
ABBOTT	Diane	Akua	Kwarteng	None.
ABBOTT	Diane	Ayobami	Sanusi	None.
ABBOTT	Diane	Matthew	Willgress	Co-ordina
ABRAHAMS	Debbie	Kari	Mawhood	None.
ABRAHAMS	Debbie	Harpreet	Uppal	Labour Cc
ADAMS	Nigel	Maire	Crosse	None.
ADAMS	Nigel	Howard	Ferguson	None.
ADAMS	Nigel	Zara	Hudson	None.
AFOLAMI	Bim	Laura	Bailey	None.
AFOLAMI	Bim	Tara	Joe	None.
AFOLAMI	Bim	Jack	Wharton	None.
AFRIYIE	Adam	Edward	Ingram	None.
AFRIYIE	Adam	Stefan	Nowak	None.
AFRIYIE	Adam	Mary	Stamp	Councillo
AHMAD KHAN	Imran	Rachel	Evans	None.
AHMAD KHAN	Imran	Pablo	Schollaert	None.
AIKEN	Nickie	Harry	Cavill	None.

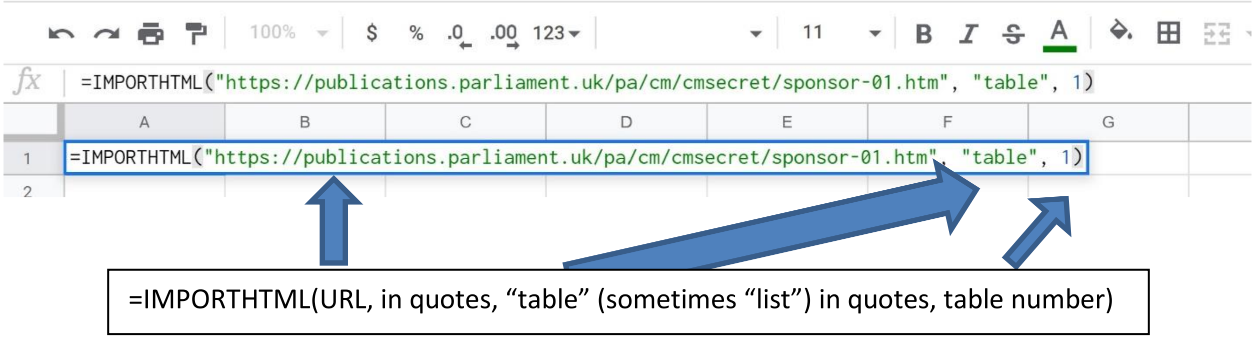
To get this we selected “Abbott, D to Coyle, N” in the left window. To insert the whole page as currently shown into a spreadsheet, click “Load” – bottom right of this screen. If you want to edit the data before importing it (eg if you don’t need all the columns) click “Transform Data”. However, it’s often easier to import everything and edit it later.

Importing tables using Google’s importHTML function

Googlesheets offer Another way of scraping a table into a spreadsheet you can keep and analyse. This requires the use of a built-in function – importHTML().

This function or formula is a little piece of code, which requires you to enter three elements, in a precise order – the URL of the site you want to import from, the element you want to import from that site (either a table or a list – there are no other options), and the number of the table you want (many websites contain more than one table, and there's no easy way of predicting the number of the element you want, so you just use trial and error, start with 1, and see if that works. If not, try 2, and so on)

The URL and the word "table" are entered between double quote marks. The table number is entered as a plain digit. So the complete formula looks like this:



```
=IMPORTHTML("https://publications.parliament.uk/pa/cm/cmsecret/sponsor-01.htm", "table", 1)
```

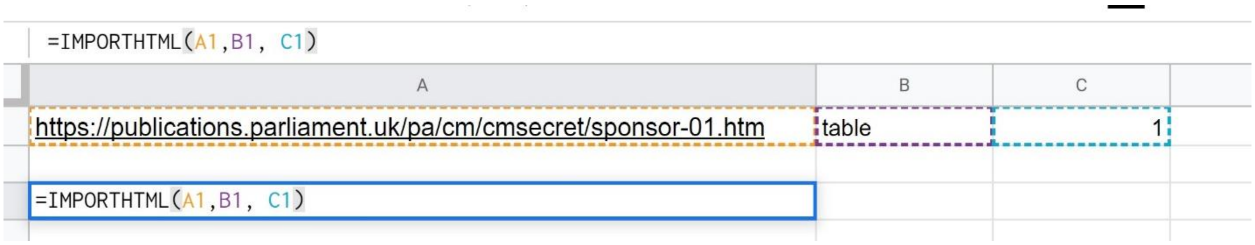
=IMPORTHTML(URL, in quotes, "table" (sometimes "list") in quotes, table number)

In this example, which is a fairly typical one, we will probably want all sub-tables containing the data – the format is the same for each table

https://publications.parliament.uk/pa/cm/cmsecret/sponsor-01.htm - the page URL changes from sponsor-01.htm to sponsor-02.htm and so on, up to sponsor-05.htm. Each page contains about 400 records so you would need to find the end of the range you just imported, then rewrite the formula one line below it, and repeat the process. It's a little tedious, and there is an easier way.

Instead of writing the whole formula in one go, you could split it into its three parts and put each part in a different cell – A1, B1, C1 for example.

Like so –



```
=IMPORTHTML(A1,B1, C1)
```

A	B	C
https://publications.parliament.uk/pa/cm/cmsecret/sponsor-01.htm	table	1
=IMPORTHTML(A1,B1, C1)		

When you close the final parenthesis and hit "enter" the imported table will appear, starting in A3. You can then copy and paste that page of data into a new worksheet, and then go back into A1 and change the -01.htm to -02.htm. Hit enter to leave the cell ,and the table starting in A3 will update to import the page called – sponsor-02.htm. You can then copy and paste the new page of data at the end of the worksheet containing page 1, return to edit 02.htm to 03.htm and repeat.

BUT – there's one vital detail. You must not use the standard paste function – this will copy the formula which imports the data – so every time you reopen the file and try to work on the data it will reimport it, and if you were to delete the cell which includes the formula, all the data will disappear (try it – delete the word sponsor in A3 and see what happens! Use CTRL+Z to go back a step).

To get the data without the formula, highlight it all type CTRL+C, then go to a new worksheet, and click Edit -> Paste Special ->->Paste Values Only [CTRL+SHIFT+V if you prefer keystrokes only]. This pastes only the data. The cell containing the word "Sponsor" contains only that word, and you could delete it without the rest of the data disappearing.

If you don't like remembering formulas and you get the importhtml function working in one spreadsheet you may prefer to keep an "import sheet" in which the URL, table, table number are all in, say, cells A1, B1, C1 and your formula is in cell A3. You would normally need to change the URL in A1, and perhaps the table number in C1, and then copy the results of your import – beginning in A3 – to wherever you need them, remembering to use the Paste Special – Paste Values Only command.