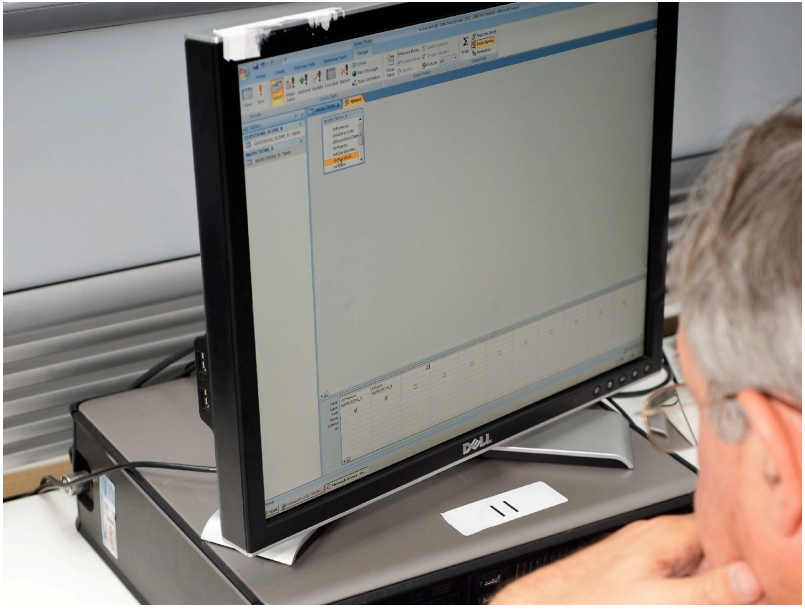
Data Journalism

THE CENTRE FOR INVESTIGATIVE JOURNALISM

This handbook is for all journalists who want to master the art of interrogating and questioning numbers competently. Being able to work with figures and investigate numbers is not a new form of journalism but a skill that all journalists can acquire.

Data journalism is the ability to analyse and examine numbers and to know how to manage large datasets and read them correctly. This handbook will guide you through the basics and allow you to use numbers to find and support your stories.



by **Elena Egawhary** and **Cynthia O'Murchu**

about the authors

Elena Egawhary studied to be a human rights lawyer and ended up as an investigative journalist. Whilst working as a legal reporter, she won a bursary from the Marjorie Deane Financial Journalism Foundation to study for an MSc in Financial Regulation and Corporate and Financial Crime at the London School of Economics.

Elena's writing has been published in The Guardian, The New Statesman and The Independent.

She can be contacted at egawhary.elena@gmail.com

Cynthia O'Murchu is a multimedia specialist and investigative reporter at the Financial Times. She was previously the London business daily's Deputy Interactive Editor where she researched and produced multimedia features and data visualisations.

about the cij

The centre for investigative journalism (cij) came into being in 2003 to address a deepening crisis in investigative reporting.

The cij provides high-level training, resources and research to journalists, researchers, non-governmental organisations, academics, graduate students and others interested in public integrity and the defence of the public interest.

The cij is a non-profit organisation and runs international summer schools, training programmes in basic and advanced investigative techniques and organises public meetings – all designed to raise and sustain the standards of investigative reporting. Our handbooks, archive material, web and audio resources have helped bring additional investigative tools to journalists and the community unable to attend cij workshops and training programmes.

The cij offers particular assistance to those working in difficult environments where freedom of the press in under threat and where reporting can be a dangerous occupation.

contents

what is data journalism?

At its most basic, data journalism begins by asking questions of numbers or proving something that you know is happening and is probably widespread, through numbers. Two classic examples are below.

Example 1: questioning figures
A government says that X percent of children are living in poverty and that the numbers have fallen over the last five years. How do they measure this? What do they define as poverty? What age range does their definition of a child fall within? Have the measurements of poverty changed over the years, meaning that the claim that it has fallen over the past five years is a distortion?

Example 2: proving a problem with numbers
You know that your local police force settles lawsuits and would prefer to avoid the shame of going to court or risking high legal fees by settling cases of alleged police brutality outside of court. You wonder if other forces might do this too. You send an FOI to all police forces to establish how much money they spent on out of court settlements and how many cases they settle, how many go to court and how many are withdrawn in the last year.

Where Do I Get the Data?
There are a number of ways for getting hold of datasets:

- A source comes to you and tells you about it
- You are examining an area and realise that data has been collected
- Regulatory bodies; if an organisation investigates or regulates anything, it will hold data to measure the performance of whatever it is regulating.
- Whenever you fill in a form or tick a box ask yourself where the information goes
- HR departments will often hold a lot of statistical information about their employees.
- Every public body will have some sort of reporting requirements – find out what they are.
- Information that is not in the format of statistics but you can pull statistics out of it
- Government statistics that, when combined with other statistics produce a new picture.

Once you have started asking questions you are on your way, but you need a computer programme to collect your figures and to let you analyse them.

freedom of information

A lot of the time you may need a freedom of information request (FOI) for data. There are three main types of FOI requests:

- For an entire database
- For an unpublished dataset that you know is collected by a public body
- To various authorities to collect the data yourself

software to manage your data

When you've got your data you'll need some software to help you manage and make sense of the large lists of numbers or statistics. Microsoft Excel (Open Office works but it can be slow) is the spreadsheet software most people use for data journalism. But while Excel is extremely useful for analysing numbers, it is not always the most efficient tool for dealing with large volumes of data. If your data entries run to tens or hundreds or thousands, or you have two or more datasets that you want to cross-reference you will need use a database manager.

Database managers, MySQL being the most popular, can handle many more records than a spreadsheet. So when you obtain a huge database, for example the Combined Online Information System (COINS) database of government spending, the Iraq and Afghanistan files or indeed any government database, you'll need to use a database manager.

best practice – to avoid getting fired

It is imperative when doing data journalism that you question yourself and are the most sceptical person about your story. Data-based stories can earn you a front page, but get it wrong and force your company to print a front page retraction and you may well find yourself looking for a new job.

Before you do anything, save your spreadsheet and keep it in its original state as a master copy, this way you will always have the original to go back to.

Think of the spreadsheet as a notebook and keep a record of each of each calculation. Every time you do a new calculation save a version. If anyone questions your story you need to be able to retrace your steps and show how you reached your conclusion.

Make sure you know the data and understand what it is that you are looking at fully:
Who does your data include and exclude?
What definitions have been used for the terms that the data is said to represent?

Ahead of publishing a story give whoever you are targeting the right to reply; talk them through how you found your story in the data. It is better to receive questions or criticisms before you publish rather than after - if you are right there is nothing they can do to stop you running the story.

Excel opens with a spreadsheet - it has columns along the top, labelled with letters and rows running down the side labelled with numbers. If you click on a cell you will see it's location in the address bar and the cell will have a thick black border to show it's active.

Basic Calculations

The best way to learn how to analyse data is by doing it. We are going to create a fictional spreadsheet of local authority spending over two years to give you an example of some of the basic calculations that can be done and how you find tips for stories in Excel.

Copy the data so that your table looks like this:

	A	B	C	D
1	**Category of Spend**	**First Year**	**Second Year**	
2	Administrative Costs:	25000000	50000000	
3	Schools:	200000000	250000000	
4	Housing:	20000000	25000000	
5	Recreation and Tourism:	15000000	10000000	
6	Salaries of Council Staff:	15000000	40000000	
7	Refuse Collection:	25000000	15000000	
8	Highways:	30000000	25000000	
9	Child Protection Social Services:	50000000	55000000	
10	Adult Social Services:	25000000	15000000	
11				

As a journalist the first question might be what did the local authority increase/decrease spending on? This would signify the area they are putting the most/least resources into. We could then start to examine why they might be moving resources around in this manner.

Calculating increase and decrease in terms of amount

Whenever you do a calculation in Excel you must let it know this is what you are doing by typing an = sign before the calculation. Excel works by referring to the cell addresses. The formula for calculating the increase or decrease from the previous year needs to be done using the cell addresses.

If we were to calculate the administrative costs on a calculator we would subtract 25,000,000 from 50,000,000. In Excel the same calculation would be a formula and refer to the cells, not the numbers, so it would be =C2-B2

The beauty of Excel is that this can be replicated very quickly - hover over the bottom right hand corner of the cell until you see a thin black cross and then double click - you should see that your calculation has been repeated for all of the items.

At this point we might want to sort our numbers to find out which item had the biggest increase and which had the biggest decrease. To do this, click any cell in the column that you want to sort. When you sort in Excel there are three important steps you must follow:

- Highlight everything apart from the column titles.
- Select the column title you wish to sort by, in this case increase/decrease or values, which is the default setting.

So at first glance we can clearly see that expenditure on schools has had the largest injection of money and both adult social services and refuse collection have had the highest decrease in spending.

However this is not the whole picture. Yes the schools category has had the largest increase of money but that may well be expected as it is the most expensive item.

Which item has had the largest percentage increase in expenditure?

- Highlight everything apart from the column titles.
- Select the column title you wish to sort by, in this case increase/decrease or values, which is the default setting.
- Then finally select largest to smallest so the biggest increase is at the top and click ok.

You should find that your dataset now looks like this:

	A	B	C	D
1	**Category of Spend**	**First Year**	**Second Year**	**Increase/decrease**
2	Schools:	£ 200,000,000.00	£ 250,000,000.00	£ 50,000,000
3	Administrative Costs:	£ 25,000,000.00	£ 50,000,000.00	£ 25,000,000
4	Salaries of Council Staff:	£ 15,000,000.00	£ 40,000,000.00	£ 25,000,000
5	Housing:	£ 20,000,000.00	£ 25,000,000.00	£ 5,000,000
6	Child Protection Social Services:	£ 50,000,000.00	£ 55,000,000.00	£ 5,000,000
7	Recreation and Tourism:	£ 15,000,000.00	£ 10,000,000.00	-£ 5,000,000
8	Highways:	£ 30,000,000.00	£ 25,000,000.00	-£ 5,000,000
9	Refuse Collection:	£ 25,000,000.00	£ 15,000,000.00	-£ 10,000,000
10	Adult Social Services:	£ 25,000,000.00	£ 15,000,000.00	-£ 10,000,000
11				

Calculating the percentage change

There is a very simple way to remember how to calculate percentage change: NOO

NOO = (New number – Old number)/Old Number

The formula for the schools would be: =(C2-B2)/B2

The reason we use brackets is that we are telling Excel to first do the calculation in the bracket and then divide the result of that calculation by the old number. To format the cell, go to the formatting section of the home ribbon and select percentage.

Once again we can take our copy tool and copy down the results. And then sort yet again in the same way as before. Suddenly a different picture begins to emerge. The salaries of the local authority staff have increased by 167% and the administrative costs have risen by 100% since the previous year.

This may well lead a journalist to ask why, and the answers could be many ranging from innocent to an abuse of power. For example it could be that staff numbers have increased, leading to an increase in salaries and HR costs. Or, it could be there has been no increase in staff numbers, but that councillors have allocated large salary increases and that the department now spends more on salaries and administrative costs as a result. The answer to what is in essence a tip-off lies in old-fashioned journalistic inquiry.

Calculating the total

Excel can calculate totals in seconds; enter a new title 'total' in row 11.
The formula is:

**=SUM(*first cell in a range*
:last cell in the range)**

in this example it is =SUM(B2:B10)

The result should and would look like this:

	A	B	C	D	E
1	**Category of Spend**	**First Year**	**Second Year**	**Increase/decrease**	**% Increase/Decrease**
2	Schools:	200000000	250000000	=C2-B2	=(C2-B2)/B2
3	Administrative Costs:	25000000	50000000	=C3-B3	=(C3-B3)/B3
4	Salaries of Council Staff:	15000000	40000000	=C4-B4	=(C4-B4)/B4
5	Housing:	20000000	25000000	=C5-B5	=(C5-B5)/B5
6	Child Protection Social Services:	50000000	55000000	=C6-B6	=(C6-B6)/B6
7	Recreation and Tourism:	15000000	10000000	=C7-B7	=(C7-B7)/B7
8	Highways:	30000000	25000000	=C8-B8	=(C8-B8)/B8
9	Refuse Collection:	25000000	15000000	=C9-B9	=(C9-B9)/B9
10	Adult Social Services:	25000000	15000000	=C10-B10	=(C10-B10)/B10
11	**Total**	=SUM(B2:B10)			
12					
13					

Use the copy tool to drag the formula to the right and into the adjacent two cells to work out what the total for the second year was and what the total increase or decrease was.

Finally we can work out what the overall percentage change is by going to the last cell of the percentage change column, cell E10 and using the copy tool to drag down the formula. You should find that the overall expenditure increased by 20%.

Calculating the average and the median

The **average** is the sum of all of the items divided by the total number of items. To calculate the average amount spent in each category, divide the items by the total number of items using this formula:

=AVERAGE(first cell in a range:last cell in the range)

in this example it is =AVERAGE (B2:B10)

However, beware of averages as the total can be distorted by a one-off large spend. It is always worth calculating the median as well.

The median is the mid-point, where half of the numbers of your selection fall below and half of the numbers in your selection are above. Calculating the **median** is very similar to the average, the formula is:

=MEDIAN(first cell in a range:last cell in the range)

in this example it is =MEDIAN (B2:B10)

By copying the formulas for both the average and the median into the cells on the right by using the copy tool, you should end up with the following:

	A	B	C	D	E	F
1	**Category of Spend**	**First Year**	**Second Year**	**Increase/decrease**	**% Increase/Decrease**	
2	Schools:	£ 200,000,000.00	£ 250,000,000.00	£ 50,000,000	25%	
3	Administrative Costs:	£ 25,000,000.00	£ 50,000,000.00	£ 25,000,000	100%	
4	Salaries of Council Staff:	£ 15,000,000.00	£ 40,000,000.00	£ 25,000,000	167%	
5	Housing:	£ 20,000,000.00	£ 25,000,000.00	£ 5,000,000	25%	
6	Child Protection Social Services:	£ 50,000,000.00	£ 55,000,000.00	£ 5,000,000	10%	
7	Recreation and Tourism:	£ 15,000,000.00	£ 10,000,000.00	-£ 5,000,000	-33%	
8	Highways:	£ 30,000,000.00	£ 25,000,000.00	-£ 5,000,000	-17%	
9	Refuse Collection:	£ 25,000,000.00	£ 15,000,000.00	-£ 10,000,000	-40%	
10	Adult Social Services:	£ 25,000,000.00	£ 15,000,000.00	-£ 10,000,000	-40%	
11	**Total**	£ 405,000,000.00	£ 485,000,000.00	£ 80,000,000.00	20%	
12	**Average**	£ 45,000,000.00	£ 53,888,888.89	£ 8,888,888.89		
13	**Median**	£ 25,000,000.00	£ 25,000,000.00	£ 5,000,000.00		

Clearly there is a big difference between the average and median amount spent per category. When deciding which figure to use you need to decide which will give as accurate a picture of the data as possible – when the difference between the highest and lowest numbers is very great the median will usually paint a better picture.

Calculating both the median and the average will enable you to sop outliers. These are numbers that fall far above or below the average. They can be tip offs or they can be errors in the data, it is up to you as a journalist to find out before you even begin to think about printing a story.

	A	B	C	D	E	F	G
1	**Category of Spend**	**First Year**	**Second Year**	**Increase/decrease**	**% Increase/Decrease**	**% of whole in second year**	
2	Schools:	£ 200,000,000.00	£ 250,000,000.00	£ 50,000,000	25%	52%	
3	Administrative Costs:	£ 25,000,000.00	£ 50,000,000.00	£ 25,000,000	100%	93%	
4	Salaries of Council Staff:	£ 15,000,000.00	£ 40,000,000.00	£ 25,000,000	167%	160%	
5	Housing:	£ 20,000,000.00	£ 25,000,000.00	£ 5,000,000	25%	#DIV/0!	
6	Child Protection Social Services:	£ 50,000,000.00	£ 55,000,000.00	£ 5,000,000	10%	#DIV/0!	
7	Recreation and Tourism:	£ 15,000,000.00	£ 10,000,000.00	-£ 5,000,000	-33%	#DIV/0!	
8	Highways:	£ 30,000,000.00	£ 25,000,000.00	-£ 5,000,000	-17%	#DIV/0!	
9	Refuse Collection:	£ 25,000,000.00	£ 15,000,000.00	-£ 10,000,000	-40%	#DIV/0!	
10	Adult Social Services:	£ 25,000,000.00	£ 15,000,000.00	-£ 10,000,000	-40%	#DIV/0!	
11	**Total**	£ 405,000,000.00	£ 485,000,000.00	£ 80,000,000.00	20%	#DIV/0!	
12	**Average**	£ 45,000,000.00	£ 53,888,888.89	£ 8,888,888.89			
13	**Median**	£ 25,000,000.00	£ 25,000,000.00	£ 5,000,000.00			

To find out which category gets the biggest piece of the expenditure pie use the following formula, in this instance using schools as an example:

=C2/C11

Format the cell as a percentage. You may think that you can just copy this formula as you did before, by double clicking the thin black cross. But if you do, you will get the above result.

When you get results like this, examine the formula in each cell. If we look at the formulas we should see that this is what is happening:

	A	B	C	D	E	F
1	**Category of Spend**	**First Year**	**Second Year**	**Increase/decrease**	**% Increase/Decrease**	**% of whole in second year**
2	Schools:	200000000	250000000	=C2-B2	=(C2-B2)/B2	=C2/C11
3	Administrative Costs:	25000000	50000000	=C3-B3	=(C3-B3)/B3	=C3/C12
4	Salaries of Council Staff:	15000000	40000000	=C4-B4	=(C4-B4)/B4	=C4/C13
5	Housing:	20000000	25000000	=C5-B5	=(C5-B5)/B5	=C5/C14
6	Child Protection Social Services:	50000000	55000000	=C6-B6	=(C6-B6)/B6	=C6/C15
7	Recreation and Tourism:	15000000	10000000	=C7-B7	=(C7-B7)/B7	=C7/C16
8	Highways:	30000000	25000000	=C8-B8	=(C8-B8)/B8	=C8/C17
9	Refuse Collection:	25000000	15000000	=C9-B9	=(C9-B9)/B9	=C9/C18
10	Adult Social Services:	25000000	15000000	=C10-B10	=(C10-B10)/B10	=C10/C19
11	**Total**	=SUM(B2:B10)	=SUM(C2:C10)	=SUM(D2:D10)	=(C11-B11)/B11	=C11/C20
12	**Average**	=AVERAGE(B2:B10)	=AVERAGE(C2:C10)	=AVERAGE(D2:D10)		
13	**Median**	=MEDIAN(B2:B10)	=MEDIAN(C2:C10)	=MEDIAN(D2:D10)		
14						

What is happening is that Excel has continued with the formula and instead of dividing by the total has divided by the average and then by the median and then by the blank cells beneath causing it to say “I can’t divide by zero!” with the **DIV/0!** symbol.

You can fix this by anchoring the total cell down in the formula so it can be copied over easily. You anchor a cell by using a **$** sign. To anchor the cell for the salaries of council staff the formula would be:

=C2/$C$11

To display the formula, click on control.

And then format it as percentage. Now you can copy this formula and your results should look like this:

	A	B	C	D	E	F
1	**Category of Spend**	**First Year**	**Second Year**	**Increase/decrease**	**% Increase/Decrease**	**% of whole in second year**
2	Schools:	200000000	250000000	=C2-B2	=(C2-B2)/B2	=C2/$C$11
3	Administrative Costs:	25000000	50000000	=C3-B3	=(C3-B3)/B3	=C3/$C$11
4	Salaries of Council Staff:	15000000	40000000	=C4-B4	=(C4-B4)/B4	=C4/$C$11
5	Housing:	20000000	25000000	=C5-B5	=(C5-B5)/B5	=C5/$C$11
6	Child Protection Social Services:	50000000	55000000	=C6-B6	=(C6-B6)/B6	=C6/$C$11
7	Recreation and Tourism:	15000000	10000000	=C7-B7	=(C7-B7)/B7	=C7/$C$11
8	Highways:	30000000	25000000	=C8-B8	=(C8-B8)/B8	=C8/$C$11
9	Refuse Collection:	25000000	15000000	=C9-B9	=(C9-B9)/B9	=C9/$C$11
10	Adult Social Services:	25000000	15000000	=C10-B10	=(C10-B10)/B10	=C10/$C$11
11	**Total**	=SUM(B2:B10)	=SUM(C2:C10)	=SUM(D2:D10)	=(C11-B11)/B11	=C11/$C$11
12	**Average**	=AVERAGE(B2:B10)	=AVERAGE(C2:C10)	=AVERAGE(D2:D10)		
13	**Median**	=MEDIAN(B2:B10)	=MEDIAN(C2:C10)	=MEDIAN(D2:D10)		
14						

When you look at the figures you can see very quickly that the local authority is spending more than three times the amount it spends on adult social services on administrative costs, in the second year which should raise some questions.

spreadsheets – rates, ratios and filtering

Ratio and Rates
Ratios and rates will let you compare figures fairly. For example just comparing the number of assaults by area would be very unfair as the population size varies depending on which town you are looking at. Calculating the number of assaults per 100,000 people by area would be a much fairer comparison to make.

Ratios enable you to compare the rate of incidence of one event happening to compared to another.

Filtering is a way of sorting through data; it can help you quickly find what you're looking for.

Whilst you can sort your data, it would be even more useful to filter it. The following spreadsheet has Conservative, Labour, Liberal Democrats all mixed together. Filtering will let you sort through and filter out information you are not interested in.

	A	B	C	D	E	F	G
1	EC reference	Name of MP	Party	Donor name	Type of donation	Value	Accepted date
2	C0003092	Dr Andrew Murrison MP	Conservative	na J Townsend	Cash	£3,000.00	22/12/2009
3	C0003093	Dr Andrew Murrison MP	Conservative	Mr Andrew Scott	Cash	£3,750.00	22/12/2009
4	C0003094	Dr Andrew Murrison MP	Conservative	Mr Charles Harman	Cash	£2,500.00	22/12/2009
5	C0003095	Dr Andrew Murrison MP	Conservative	Mr Stephen Morant	Cash	£1,250.00	22/12/2009
6	C0003096	Dr Andrew Murrison MP	Conservative	Mr Michael Spencer	Cash	£6,800.00	22/12/2009
7	C0003097	Dr Andrew Murrison MP	Conservative	Mrs A Pugh	Cash	£1,750.00	22/12/2009
8	C0003098	Dr Andrew Murrison MP	Conservative	Mr Dominic Taylor	Cash	£1,750.00	22/12/2009
9	C0003099	Dr Andrew Murrison MP	Conservative	Lord na Margadale	Cash	£2,000.00	22/12/2009
10	C0003100	Dr Andrew Murrison MP	Conservative	Mr Ivan Shenkman	Cash	£2,050.00	22/12/2009
11	C0000082	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£7,500.00	7/2/2008
12	C0000083	Dr Liam Fox MP	Conservative	Mr John Moulton	Cash	£50,000.00	27/03/2008
13	C0000084	Dr Liam Fox MP	Conservative	Mr Stanley Fink	Cash	£10,000.00	24/01/2007
14	C0000085	Dr Liam Fox MP	Conservative	Mr Michael Hintze	Cash	£10,000.00	24/01/2007
15	C0000086	Dr Liam Fox MP	Conservative	Mr Alan Howard	Cash	£15,000.00	24/01/2007
16	C0000087	Dr Liam Fox MP	Conservative	Mr Jon Moulton	Cash	£50,000.00	13/06/2007
17	C0000088	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£22,500.00	31/10/2007
18	C0000089	Dr Liam Fox MP	Conservative	Mr Stanley Fink	Cash	£10,000.00	29/11/2007

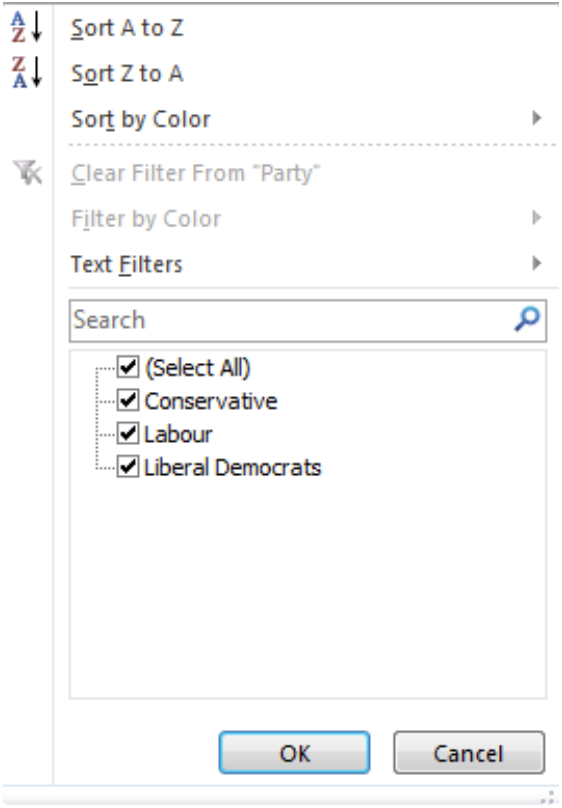
Sheet1 Sheet2 Sheet3

Click on a cell in your title row – this will usually be row 1, A1 in this example. Go to the menu ribbon and click on data, then the filter button that looks like a funnel.

An arrow will appear at the right of each box in the title row - these are your filters.

	A	B	C
1	EC referen	Name of MP	Party
2	C0003092	Dr Andrew Murrison MP	Conservative

To filter out all donations except those to the Conservative party go to the cell with the title ‘party’ and click on the arrow, a menu appear with a list of all entries in this category.



You can sort by this row by clicking on the two sort options above. If you had colour coded the cells you could sort by colour. Most important is the bottom part of the menu where all of the parties that are represented.

You can see that currently it is on (Select All). First we want to de-select (Select All) by clicking on the arrow. This should result no boxes having tick signs. We then tick the box next to Conservatives to select just this option. This will give us all the donations for Members of Parliament from the Conservative party.

When you’ve filtered results you will see that the arrow has changed to the funnel icon. This is to let you know that the list is showing a selection of, and not all the available data.

Another indication is in the bottom left-hand corner where it tells you how many records you are looking at out of the total. This is records and not people, in the case of our Conservative MPs, the same name may be entered more than once.

To remove filters and start again, click on the large filter button in the data menu. This button is a bit like the filter light switch, you can turn them on and off with it.

If you want to filter the results further, for example to find all donations to the Conservative party made in cash, you would need to use two filters - one on the ‘party’ and one on the ‘type of donation’’.

On the ‘party’ filter select ‘Conservative’.

On the ‘type of donation’ filter select ‘cash’.

We can also find out the top ten donations made to MPs within this dataset.

First, clear the filters by clicking the large funnel button in the Data Menu Ribbon twice - once to clear the filters and the second time to reinstate them.

In the 'value' column, click on the filter and select 'number filters' and then select 'top 10 AutoFilter', then ok

Now the data is unsorted - go to the 'value' column and select 'Z to A largest to smallest' again and you should see the following figures:

You can filter for other results, for example the bottom ten donors, using the same method but selecting different options in the **'top 10 AutoFilter.'**

	A	B	C	D	E	F	G
1	EC referen	Name of MP	Party	Donor name	Type of donatic	Value	Accepted da
12	C0002637	Mr David Willetts MP	Conservative	Mr Michael Hintze	Cash	£150,000.00	5/4/2007
16	C0002196	Mr Chris Huhne MP	Liberal Democrats	Lord Timonthy Clement-Jones	Cash	£120,000.00	2/24/2006
29	C0003174	The Rt Hon Edward Balls MP	Labour	na Ken Follett	Cash	£100,000.00	6/24/2010
48	C0002108	The Rt Hon David Cameron	Conservative	Lord Philip Harris	Cash	£ 90,000.00	1/8/2005
57	C0002424	Mr Boris Johnson MP	Conservative	Mr John Cluff	Cash	£ 71,000.00	3/7/2008
64	C0000723	Mr George Osborne MP	Conservative	Mr Simon Robertson	Cash	£ 55,470.03	2/26/2009
73	C0000087	Dr Liam Fox MP	Conservative	Mr Jon Moulton	Cash	£ 50,000.00	6/13/2007
87	C0000097	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£ 50,000.00	1/1/2003
156	C0000092	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£ 50,000.00	8/4/2006
178	C0002067	Mr Andrew Mitchell MP	Conservative	Mrs Helena Mary Frost	Cash	£ 50,000.00	5/13/2008
202	C0002060	Mr Andrew Mitchell MP	Conservative	Mr Roderick Fleming	Cash	£ 50,000.00	11/29/2007
235	C0002084	Mr Andrew Mitchell MP	Conservative	Mr Lennart Perlhagen	Cash	£ 50,000.00	10/19/2009
378	C0002050	Mr Andrew Mitchell MP	Conservative	Mr Michael Aeon-Buckley	Cash	£ 50,000.00	5/15/2007
421	C0003026	Mr Andrew Palmer Kerr MSP	Labour	Mr Paul McKenzie	Cash	£ 50,000.00	8/26/2008
447	C0000742	The Rt Hon David Davis MP	Conservative	Lord na Kalms	Cash	£ 50,000.00	1/7/2005
451	C0003189	The Rt Hon David Miliband MP	Labour	na David Claydon	Cash	£ 50,000.00	5/24/2010

Cleaning Data

If there is a misspelling - for example if Labour is spelt Labor - in any of your data you will have problems filtering accurately.

Before doing any work on the spreadsheet you'll need to use the filtering system to tidy up the data. When you go to the drop down filter menu for any column you will see a list of all categories, this is where you can find spelling errors and inconsistencies. Unfortunately you'll have to correct them manually, but you can use the filter to locate them.

For a beginner a pivot table is a fairly complex tool and is used for aggregating data. A pivot table lets you summarise the raw data, putting all the single records into groups and finding new ways of looking at the data.

Using the above data as an example, what would it be interesting to find out? For example:

- Which party received the most money?
- Which party received the most donations?
- Which party received the most money and from which type of donor?

We know that our data conforms to this having worked with it before but it is always worth checking.

As always, before you do anything, save your data so you don't have to repeat the work done so far should you make a mistake.

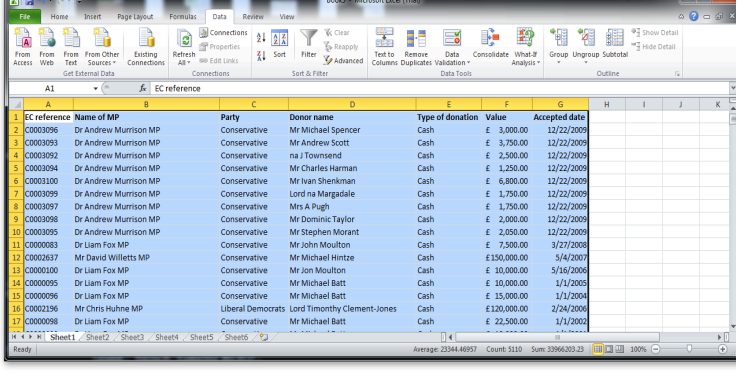
There are three conditions for data that you want to pivot:

- There must be a heading for each column in the first row.
- There can't be any empty rows or columns, the data must be in one contiguous block.
- Each column must have only one type of data, so either textual data, or numeric data or dates etc.

Step 1: Select your data

Take the white cross and click on cell **A1**. Hold down **Ctrl**, **Shift** and then press ***** (number 8 on your keyboard) this should have selected the entire block of the data.

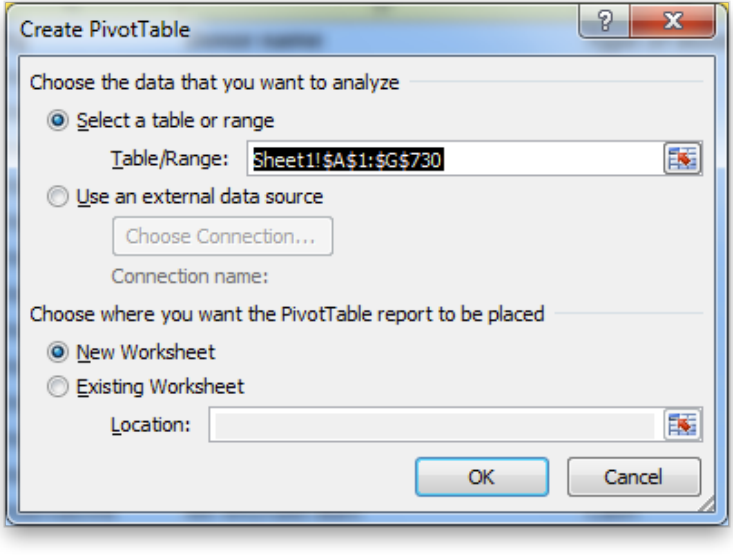
To ensure that all of the data has been selected hold down the **Ctrl** key and press **full stop** on your key board (known as period) this should move you around the outskirts of what you have selected.



EC reference	Name of MP	Party	Donor name	Type of donation	Value	Accepted date
C0003096	Dr Andrew Murrison MP	Conservative	Mr Michael Spencer	Cash	£ 3,000.00	12/22/2009
C0003093	Dr Andrew Murrison MP	Conservative	Mr Andrew Scott	Cash	£ 3,750.00	12/22/2009
C0003092	Dr Andrew Murrison MP	Conservative	na J Townsend	Cash	£ 2,500.00	12/22/2009
C0003094	Dr Andrew Murrison MP	Conservative	Mr Charles Harman	Cash	£ 1,250.00	12/22/2009
C0003100	Dr Andrew Murrison MP	Conservative	Mr Ivan Shenkman	Cash	£ 6,800.00	12/22/2009
C0003099	Dr Andrew Murrison MP	Conservative	Lord na Margadale	Cash	£ 1,750.00	12/22/2009
C0003097	Dr Andrew Murrison MP	Conservative	Mrs A Pugh	Cash	£ 1,750.00	12/22/2009
C0003098	Dr Andrew Murrison MP	Conservative	Mr Dominic Taylor	Cash	£ 2,000.00	12/22/2009
C0003095	Dr Andrew Murrison MP	Conservative	Mr Stephen Morant	Cash	£ 2,050.00	12/22/2009
C0000083	Dr Liam Fox MP	Conservative	Mr John Moulton	Cash	£ 7,500.00	3/27/2008
C0002637	Mr David Willetts MP	Conservative	Mr Michael Hintze	Cash	£150,000.00	5/4/2007
C0000100	Dr Liam Fox MP	Conservative	Mr Jon Moulton	Cash	£ 10,000.00	5/16/2006
C0000095	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£ 10,000.00	1/1/2005
C0000096	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£ 15,000.00	1/1/2004
C0002196	Mr Chris Huhne MP	Liberal Democrats	Lord Timonthy Clement-Jones	Cash	£120,000.00	2/24/2006
C0000098	Dr Liam Fox MP	Conservative	Mr Michael Batt	Cash	£ 22,500.00	1/1/2002

Step 2: Create the Pivot Table in a new worksheet

is go to the menu bar at the top of the spreadsheet and **'insert menu ribbon'** and click on the **'pivot table'** button; this menu will appear:



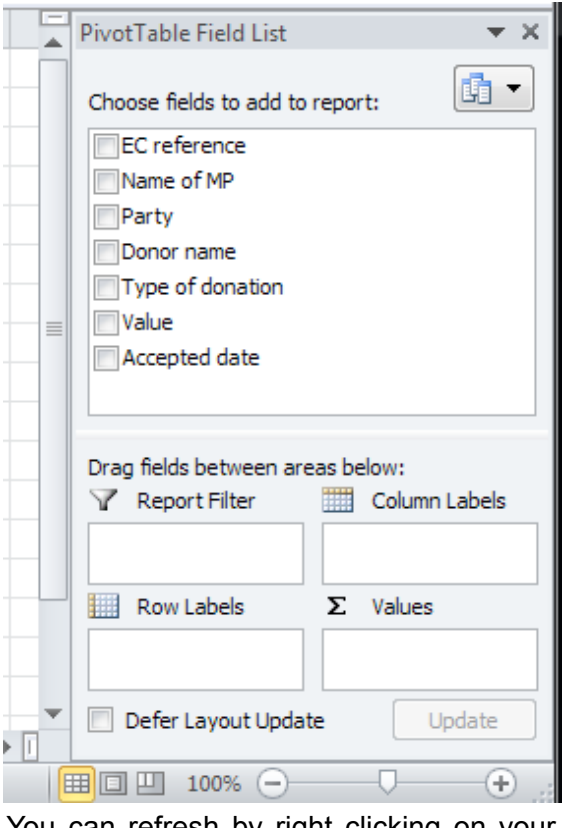
As you have already selected the data the table range will be what you selected prior to clicking the **'pivot table'** button.

On the bigger screen you should see the friendly marching ants running along the outer edges of the data you have selected.

Always put the pivot table report in a new worksheet. If you use the existing worksheet you are likely to end up getting confused and will put your original data at risk. Click **OK**.

Excel will have created a new worksheet for you and this is where you start to create your pivot table. Rename and save this sheet.

It is important to remember that your pivot table will not automatically update to reflect changes in the original data, or the addition of new records. They can however be refreshed, which will update the table to reflect changes in the source material.



You can refresh by right clicking on your table, then selecting refresh. If you have added new records to your dataset, you will need to use the Pivot Table Wizard, returning to step 2 and redefining the data range to include your new data.

The key place to manage your pivot table is in the **'pivot table field'** list.

In the **'choose fields to add to report'** you will see that the titles of the columns of your raw data appear. Excel lets you choose the fields you want to drag into boxes outlined below.

The Report Filter: this is where you drag the title of the column that would hold the data that you want to filter by. So for example if you want to only view Conservative Party donations then you would drag the **'party'** field into this box.

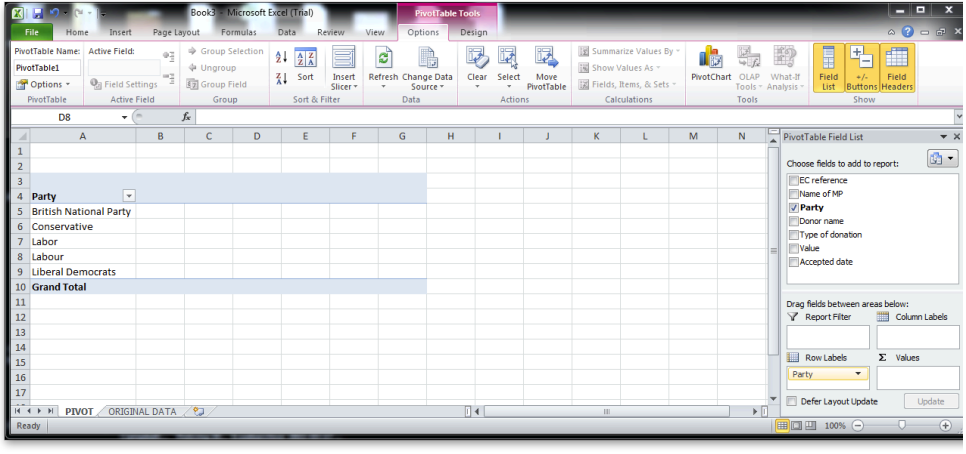
Column and Row Label Areas: this is where you drag the column with the labels you want to see, if for example you want to see which MP got the most money you would drag the **'name of MP'** field into the Row Label Area.

Values: this is where you drag the data that you want to sum-up. Usually this is the place where you drag any data that is in numeric form. So for example, if we wanted to work out how much money each MP had made in total we would drag the **'value'** field here.

So, to return to the questions:

Calculating which party received the most money

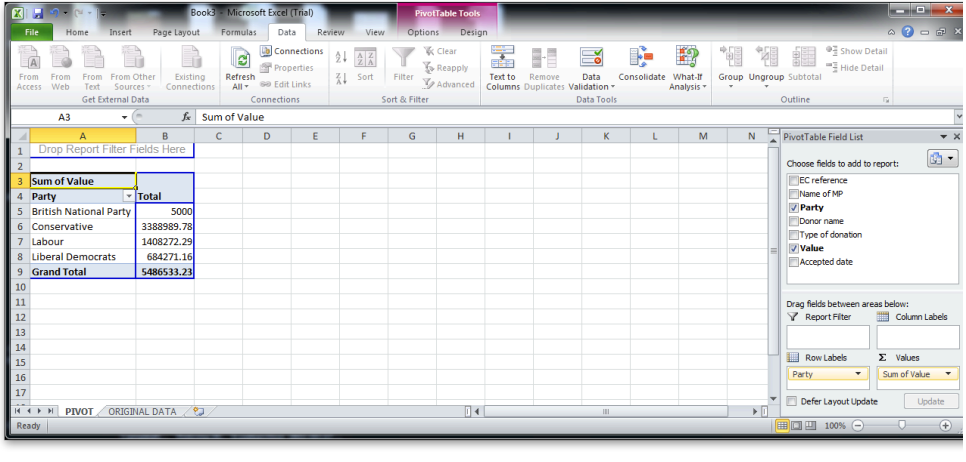
You would need to know the party name and the total value of money received. Because the data is going to be broken down by party; hover over the party field until the arrow appears then drag it into the **'row label area'**. You should see the following:

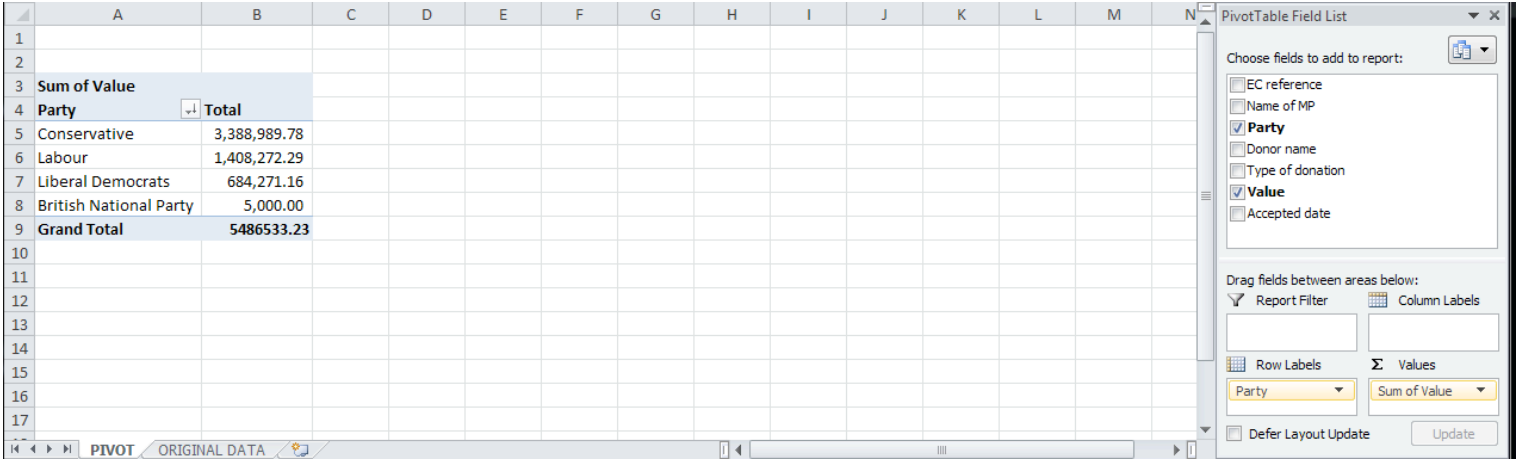


All the parties are listed on the left-hand side but the table is missing the data.

To view the amount by party you need to put the value data into the table so that Excel can sum up all of the values listed for each party.

To do this, hover over the value field until you see the arrow then drag the value field into the **'values'** box. Your spreadsheet should look like this:





Sum of Value	
Party	Total
Conservative	3,388,989.78
Labour	1,408,272.29
Liberal Democrats	684,271.16
British National Party	5,000.00
Grand Total	5486533.23

Excel is telling you what it is doing in the top left hand corner. It is summing the data in the **Value** field.

The information is there but the data isn't particularly easy to read, you can add commas by selecting the cells with the numbers and clicking on the comma button under the **'home'** menu ribbon.

You might also want to sort this data. This is easy in a pivot table; click on one cell which contains a number, say **B7** in this example, then click the small **Z** to **A** button in the data ribbon menu. You should now be looking at this.

In the space of two seconds you have now added up the entire data set to show how much money each of the parties made.

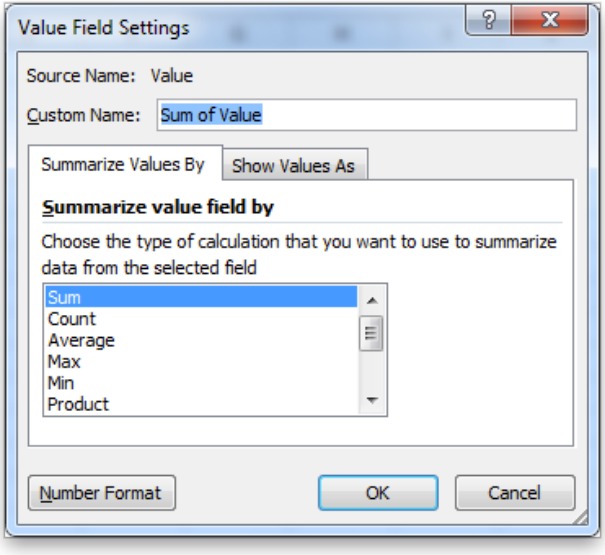
But there are still some questions.

How to calculate which party received the most donations

For this we want to know the number of the donations per party – we want to count them.

We still need to know the party and we want to know the number of donations listed in the value field so our fields are already in the right place. The only problem is that we don't have a column with the number of donations – we just have a list which Excel has grouped together for us to see the total amount. But Excel can count them rather than sum them if we ask it.

In cell **A3** there is a box saying **SUM OF Value** (see above). Hover over cell **A3** where it says sum of value until the arrow appears then double click and a new box will appear:



	A	B	C
1			
2			
3	**Count of Value**		
4	**Party**	**Total**	
5	Conservative	419.00	
6	Labour	207.00	
7	Liberal Democrats	102.00	
8	British National Party	1.00	
9	**Grand Total**	**729**	
10			
11			
12			

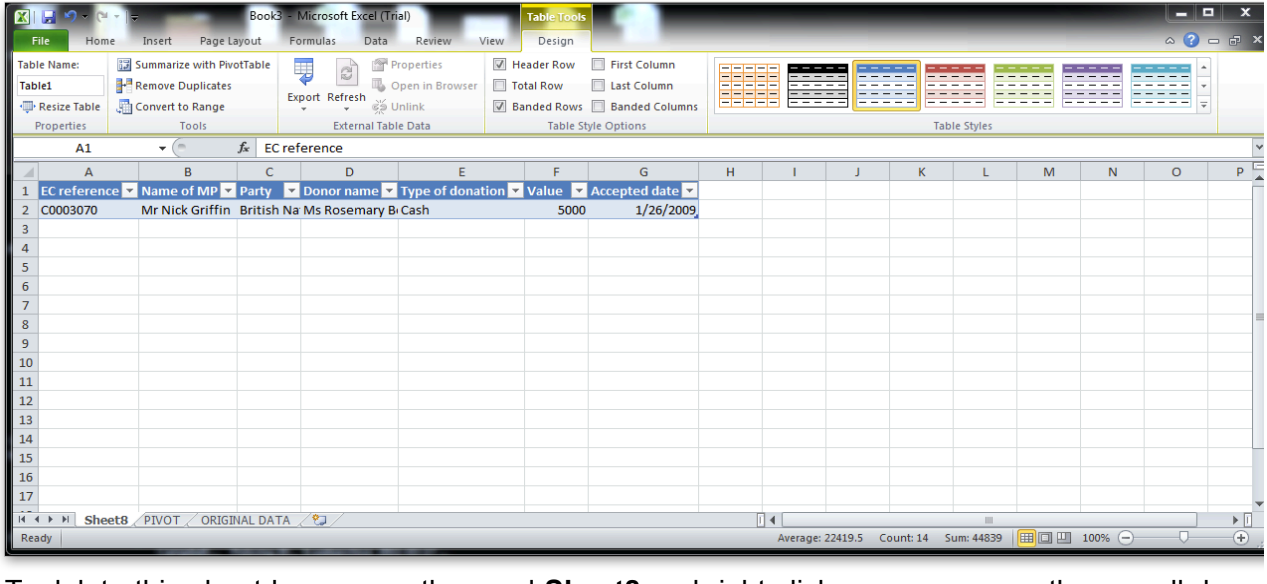
Here Excel is summing the values in the value column. If you select count by clicking on the word **'count'** then ok, you will see the result on the left.

To tidy up the formatting and remove some decimal points select a single cell in the correct column and click on the decrease decimal point button in the home menu ribbon:

In this example it becomes clear that not only did the Conservative Party receive the highest amount of money from individual donors, it also received more than double the number of donations from individual people than Labour.

Also interesting is that the British National Party, a controversial far right political party received one single donation by an individual. If you want to see more information behind the numbers displayed all you need to do is to hover over the number (in this case the number 1 which is in cell **B8**) and double click with the white cross.

What you should see is that the pivot table has created a new sheet and has placed the background information for this figure in this sheet so we can see who the donor is. Now Nick Griffin, the person under the title **Name of MP**, is not actually an MP, so this is often how you might find errors within your data. You must always proceed with caution as all data is almost always what is known as 'dirty data'. Where there are lots of records, it is likely that there will be some errors.



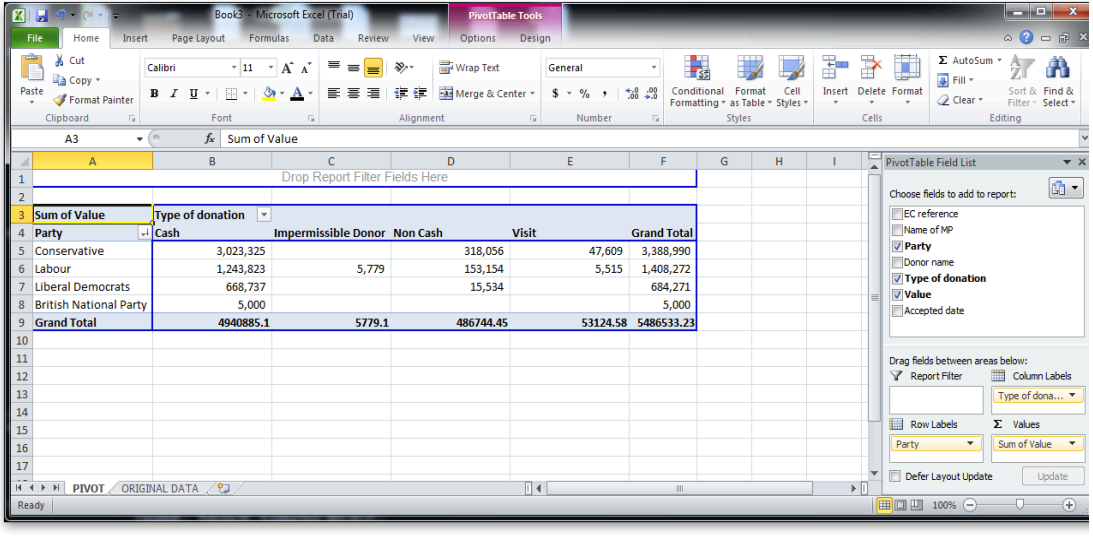
To delete this sheet hover over the word **Sheet8** and right click on your mouse then scroll down to delete and click ok on the warning screen that appears.

How to calculate which party receives the most money by the type of donation

First we want Excel to sum again so double click on the count of value cell and you should get up the same menu as before in the **'summarise value'** field box. Select SUM and click ok and you will have the amounts again.

To break it down further by donation type you need to drag another field into your column **'labels'**.

With pivot tables you are creating your own table so you need to think where would it be best placed and what information you need.



Sum of Value	Type of donation				
Party	Cash	Impermissible Donor	Non Cash	Visit	Grand Total
Conservative	3,023,325		318,056	47,609	3,388,990
Labour	1,243,823	5,779	153,154	5,515	1,408,272
Liberal Democrats	668,737		15,534		684,271
British National Party	5,000				5,000
Grand Total	4940885.1	5779.1	486744.45	53124.58	5486533.23

In this instance, drag the 'type of donation' field into the 'column labels' area and you should see something like this:

Now you can see how much each party received and the type of donor, you can always sort this if you want to. You can also see that the Labour Party received £5,779 from an impermissible donor. To find out more about this donation, double click on the number.

Finally how do we find out whom, during the Conservative leadership election campaign, received the most money?

This involves adding some report filters. Previous filtering focused on the **'party'** field, selecting only the Conservative Party, the **'type of donation'** field, selecting only cash donations and finally filtering out all dates that did not fall within 05/06/2005 and 12/06/2005.

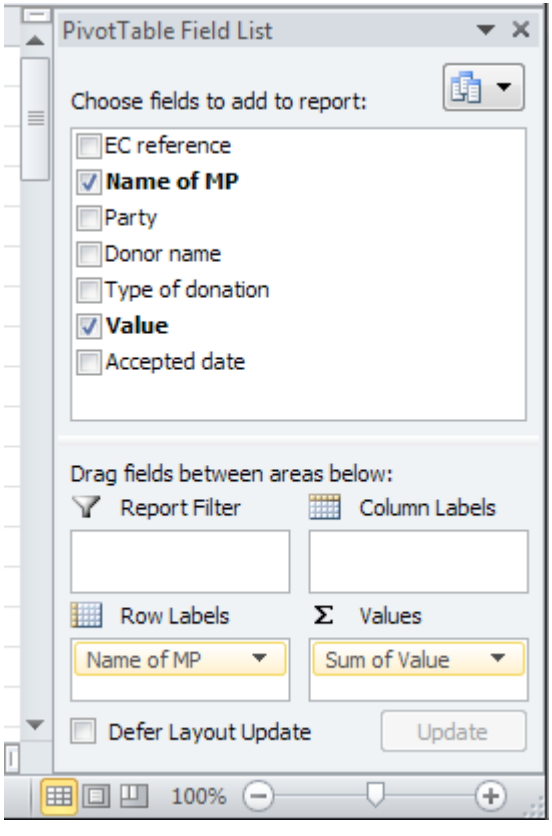
To put this information in a pivot table, first create a new one, as before, or clear out the old one by click on the **'PivotTable Tools'** menu then **'clear'** and **'clear all**'.

A useful way to think about building your pivot table is to think about what you expect to see. To find the MP with the highest amount in donations during the Conservative leadership election campaign you would expect to see MPs names on the left hand side so drag the Name of MP field into **'row labels.'**

You would expect to see the amount of money amassed by the MP, so you can add up the total value of all of the donations received by each MP. Do this by dragging the **'value field'** into the **'values box**'.

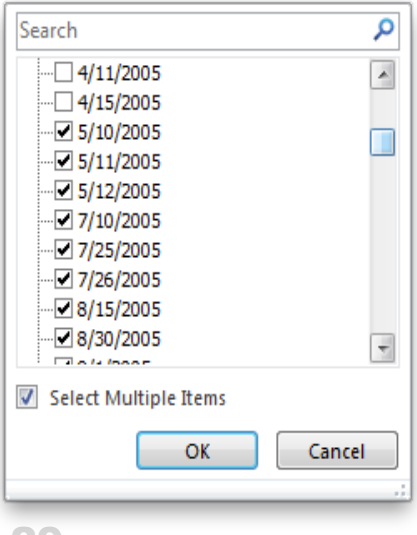
Your Pivot Table Field List should look like this:

Technically it is possible to sort at this point in the main table and see which MP received the most money overall in the dataset. But by further filtering we can get deeper into the data and look at the results for the Conservative leadership election campaign in 2005.



You do this by filtering in much the same way as before. To filter out all parties other than the Conservatives, drag the **'party'** field into the **'report filter'**. You should see at the top of the pivot table there is an arrow with drop down menu on the right of the party field:

	A	B
1	Party	(All)
2		
3	**Sum of Value**	
4	**Name of MP**	**Total**
5	Dr Andrew Murrison MP	24850
6	Dr Liam Fox MP	521200
7	Dr Vincent Cable MP	2500

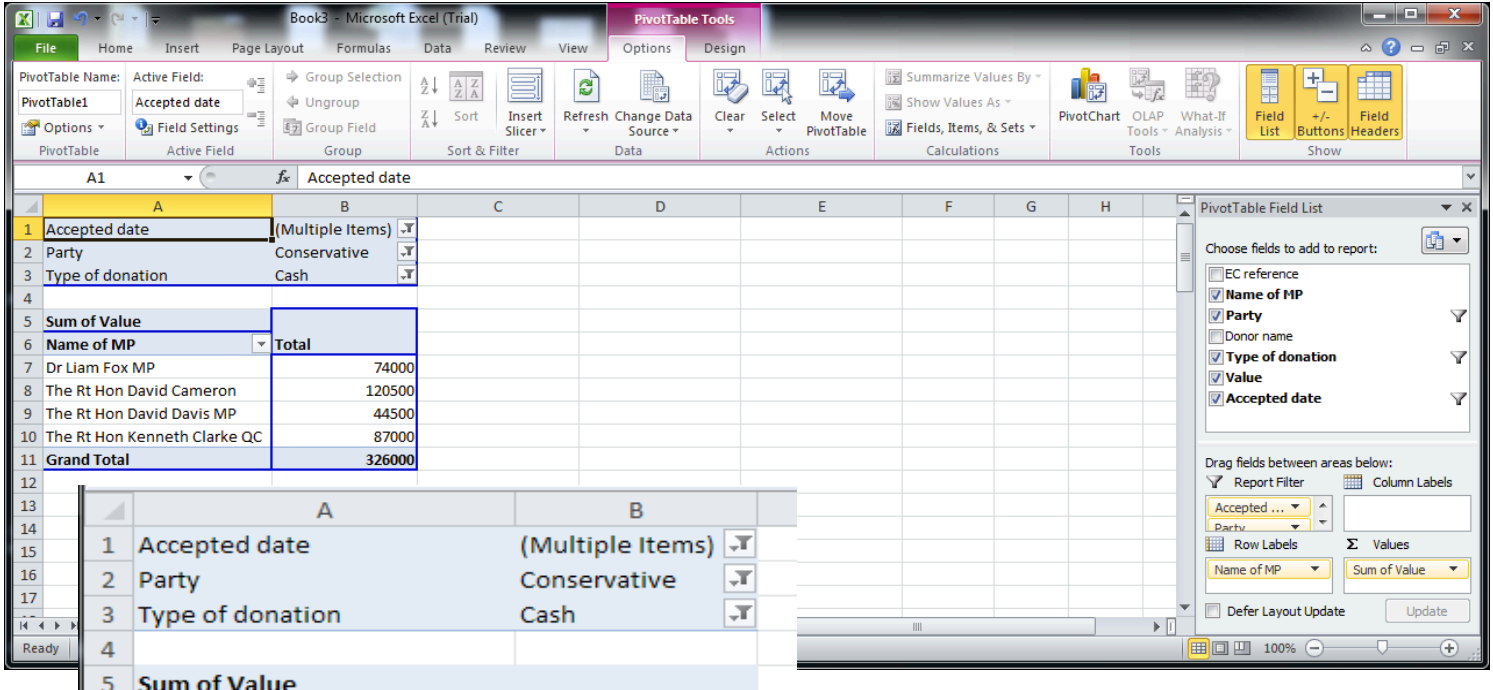


Click on the arrow next to (All) to get a menu from which you should select Conservative. Next, drag in the **'type of donation'** field and repeat the above process selecting cash.

Set the date range between 6/5/2005 and 6/12/2005 by dragging the **'accepted date'** field into the **'report filter'**, then click the arrow and untick the box that says **'select multiple items**'. Untick the box next to **'select all'** and scroll down clicking the dates within your range:

Once this is done click OK and you should see what follows in your pivot table. You can add commas and sort the figures to ensure you read them correctly using the instructions as above.

From this pivot table you can see that David Cameron had, by a huge margin, the largest amount of money to spend during the leadership election campaign and that he did indeed go on to win.



	A	B
1	Accepted date	(Multiple Items)
2	Party	Conservative
3	Type of donation	Cash
4		
5	**Sum of Value**	
6	**Name of MP**	**Total**
7	The Rt Hon David Cameron	120,500
8	The Rt Hon Kenneth Clarke QC	87000
9	Dr Liam Fox MP	74000
10	The Rt Hon David Davis MP	44500
11	**Grand Total**	**326000**
12		

In this instance only the cell referring to David Cameron has been formatted, to apply this to the entire column, use the copy tool and copy down the results. You could of course delve further by finding out who donated this money at this crucial time to Mr Cameron by double clicking on the number 120,500 and pulling up a list of all of the donors.

conclusion

By working through the above exercises you should be able to acquire a good knowledge of how you can use Excel to help manage large datasets and statistics. There are of coure, variations on all of the examples shown and as you grow in confidence you'll be able to find for yourself new ways of interrogating the data. For more links to places where you can access datasets, please visit the links directory on the CIJ website as nothing beats getting some real datasets and seeing what you can find.

Useful Links

WikiLeaks: **http://www.wikileaks.ch/**
United Nations: **http://unstats.un.org/unsd/databases.htm**
World Bank: **http://data.worldbank.org/data-catalog**
International Monetary Fund: **http://www.imf.org/external/data.htm**

finding data (online and offline)

Transparency is the building block of a democratic society. In the United States, where computer assisted reporting has been practiced for years, historically data has been much more freely available, with most states publishing data such as salaries of state employees for example. This would have been unthinkable in the UK until recently.

However, the past few years have seen increased efforts by governments and open-data campaigners to make information available to citizens. There has been a proliferation of data sets, statistics and portals to those data sets. In the UK, data.gov.uk, reflects David Cameron's ambition that UK be one of the most transparent governments in the world. Like its US counterpart data.gov, the website is run by the government and aims to be a central portal to government data.

At the time of writing, the data.gov.uk portal holds over 6,800 data sets, making it an excellent place to start when searching for any UK government data.

While many important data and information are still off limits to citizens, the starting point for any journalist should be the acknowledgement that "data is everywhere". Government agencies, private companies, non-profits and think tanks all collect data and produce statistics, and most of the information is now stored electronically.

When starting a new beat or a new project, it is important to familiarise yourself with the data available on that beat.

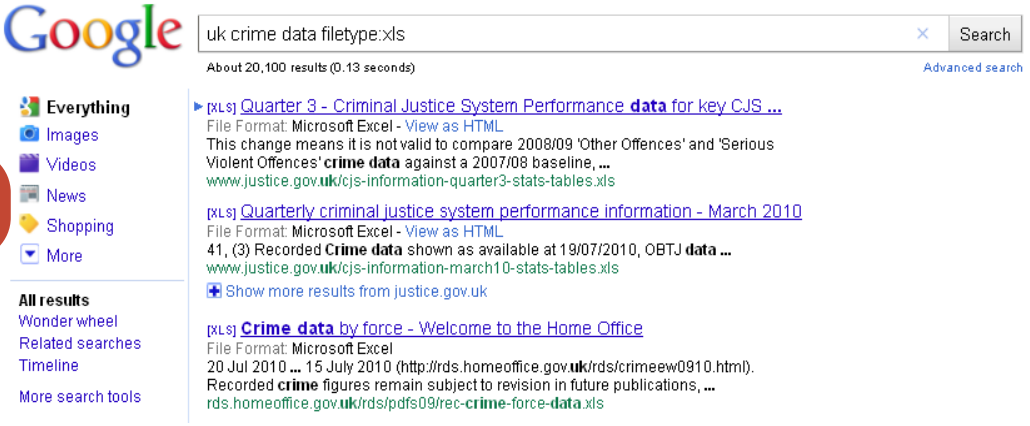
As with any source, it is crucial to understand how the data was collated, whether the source is trustworthy and whether the data set is complete as well as how often it is updated and what it contains.

The easiest and usually fastest way to obtain data is to access it via a government (or other) websites. Most government departments usually have 'publications' or 'statistics' sections on their website. Another good place to look is in the 'publication scheme' or 'information asset register' section of the website. This will usually outline the information that is already available, and mention databases that the government creates in the course of its work. While some sites will just contain downloadable files, others will post full databases online which can be queried. Bear in mind that the agency may well have data available that is not available online, but can be obtained through a Freedom of Information request.

Another method is to simply search for data via google, using the

filetype: operator, in order to force Google to only return spreadsheets for example. So a search for crime data in the UK could look as follows:

uk crime data filetype:xls or **uk crime data filetype:csv**



Remember the 'invisible web and be sure not to search just the surface of the web. Using Google or other search engines for particular types of datasets by topic will only yield surface results. The vast majority of information on the internet is hidden, in the 'deep web', and not indexed by search engines (which cannot retrieve content that is generated dynamically 'on the fly', which is hidden by paywalls or buried in databases). You can access the deep web through deep web portals/directories such as Infomine or by searching for databases of databases.

Here is a selection of portals to datasets in the UK and beyond:

http://ckan.net/, the Comprehensive Knowledge Archive Network holds several thousand data sets (not all open) collected from around the world.

http://opendatasearch.org/, a project by the Open Knowledge Foundation wosse aim is to facilitate the search for datasets internationally

ScraperWiki – a community of programmers and data users working together to scrape data sets from across the web. There will often be data sets here that are not available elsewhere as they are collected via scrapes (see below).

ONS: Office of National Statistics
http://www.neighbourhood.statistics.gov.uk/

http://data.london.gov.uk/

http://data.worldbank.org/ - economic data from the World Bank. The Data Catalog provides access to over 7,000 indicators from World Bank data sets.

If the data is not available online, it pays to spend some time with the relevant person at the agency (database manager, statistics officer) to learn about how the database is set up and what fields it contains. Press officers are often skittish about letting reporters talk to their database people and that often leads to a circuitous ping-pong of request after request and incomplete answers, when a short phone call with the right person would often resolve the query in minutes.

It is also worth collecting forms the agency holds online and offline, as they will show you fields that are contained in databases they hold. When requesting data from government agencies it's useful to ask for the 'data dictionary' or 'metadata' of the database. This information describes what fields are contained in the database and what format they are kept in, what type of data is contained within those fields or columns (eg numeric, text, date), whether they are mandatory or not etc. You will also need some explanation of what the fields/column headers mean, what they include or exclude. It is important to understand this thoroughly to ensure you don't misinterpret the data.

5.1 Performance Data

Name	*Data Type*	*Mandatory*	*Comment*
HA_AREA	Characters (2)	Y	E.g. 03 or 12. Must contain leading zero
REPORTING_PERIOD	Number (6)	Y	Must be of format YYYYMM
INCIDENT_REFERENCE NUMBER	Characters (8)	Y	Unique ID from MAC Incident Management System
EMERGENCY_FLAG	Text (1)	Y	Upper case Y or N only
PROACTIVE_FLAG	Text (1)	Y	Upper case Y or N only
INCIDENT_REPORTED_BY	Text (50)		Defined by the HA INCIDENT_REPORTED_BY spreadsheet. A list of values is shown in the Reference Data section. The default value OTHER (DETAILED DESCRIPTION REQUIRED) should be used where MAC cannot populate this field for any given Incident Response.

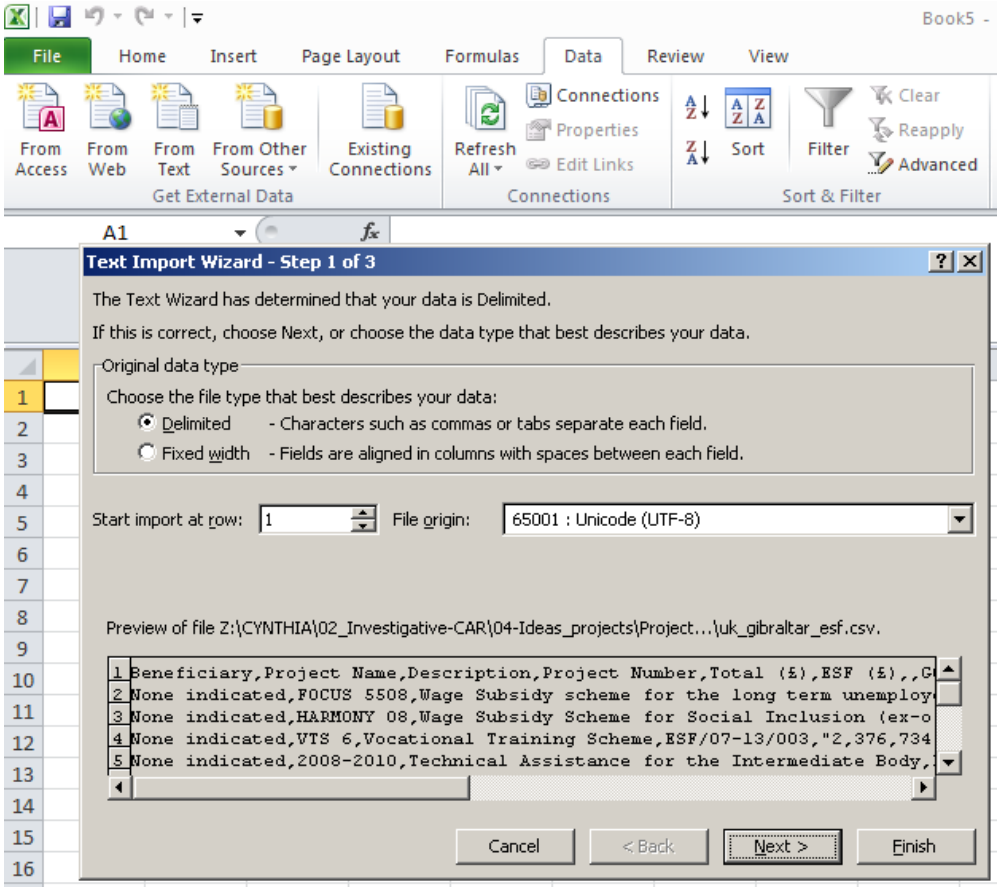
Importing text files
If you receive data as a plain txt or CSV file you may need to import it into Excel. Often, CSV files can simply be opened by double clicking on the file name and Excel will recognise how it needs to be presented. Sometimes, however, txt or CSV files need to be imported via the 'import' function in Excel.

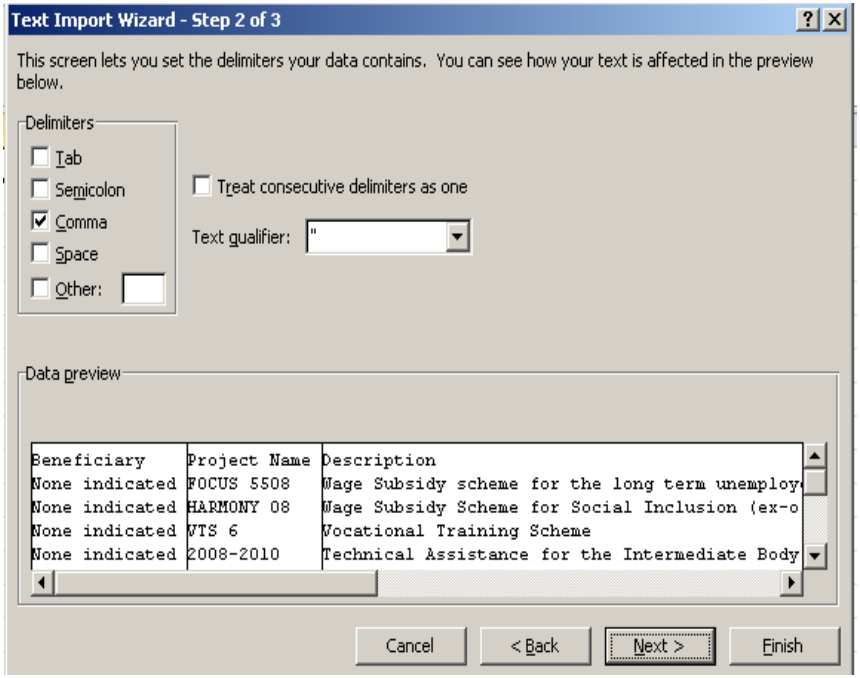
Plain files can be separated by a coma, for example, (csv – comma-separated values) or delimited (tab delimited), or they can be fixed width files.

To import them, go to file/import/ from text - this will bring up a dialog box asking you to select the text or CSV file. Choose the file and click import.

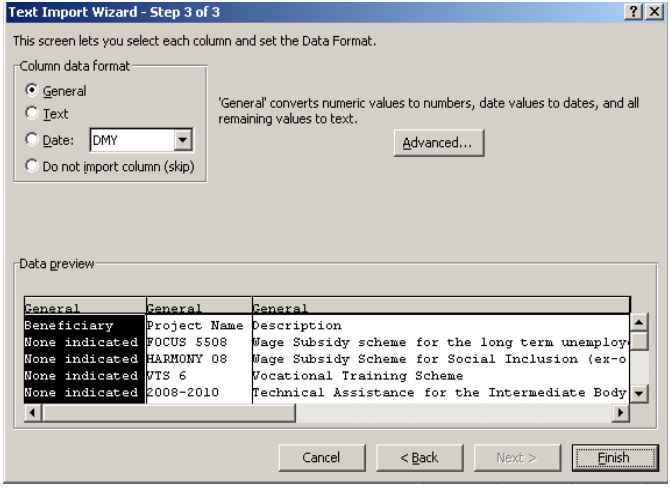
The import wizard will usually detect whether your file is a fixed width or delimited file and ask you to confirm.

In this example, the data is comma-separated (as you can see there are commas in between each of the column headers). Click 'next' and select 'comma'. This will break up the text to create columns in place of the commas.





The next menu allows you to determine whether you want to skip importing a column and whether Excel should treat it as a date or text (useful if you have ID numbers for example beginning with 0. In such cases the ID number should be imported as text, because otherwise Excel will treat 000344 as 344 rendering match between ID numbers from different datasets impossible).

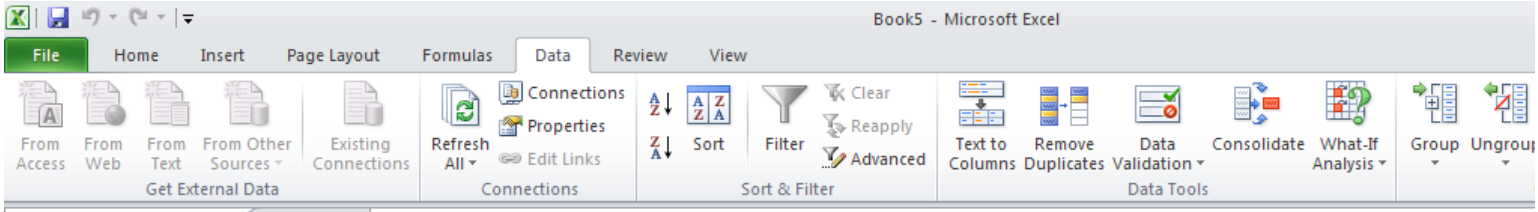


Click ‘finish’ and you have your data set in a format you can work with.

If you want to import ‘fixed width’ files you need to follow the instructions from the data dictionary, which will indicate how long each of the fields should be. Click import from text, as above, but in the import wizard select ‘fixed width’ - fields are aligned in columns with spaces between each field.

Click ‘next’. On this screen you will create the break lines in order to indicate where the columns start and end. In the ‘data preview’ window you will see a set of numbers. These represent the number of characters from the start of a row. If the data dictionary says that ‘employee name’ is 50 characters long, click on the tick underneath the number 50. This will create the break line. To delete it, double click on the line. To move the break line you can click and drag it to the desired position.

Continue like this until you reach the end of the row then click ‘next’. Choose ‘date’ (and specify the format) if needed for dates then click ‘finish’ and the data set will import.



	A	B	C	D	E	F	G
1	Beneficiary	Project Name	Description	Project Number	Total (£)	ESF (£)	
2	None indicated	FOCUS 5508	Wage Subsidy scheme for the long term unemployed	ESF/07-13/001	18,565.00	9,282.00	
3	None indicated	HARMONY 08	Wage Subsidy Scheme for Social Inclusion (ex-offenders; ex-addicts; etc)	ESF/07-13/002	78,734.00	39,368.00	
4	None indicated	VTS 6	Vocational Training Scheme	ESF/07-13/003	2,376,734.00	1,188,367.00	
5	None indicated	2008-2010	Technical Assistance for the Intermediate Body	ESF/07-13/004	90,387.00	45,193.00	
6	None indicated	CTC INTAKE 13	Apprenticeship Scheme in the Construction Trades	ESF/07-13/005	83,154.00	41,577.00	
7	None indicated	CTC INTAKE 14	Apprenticeship Scheme in the Construction Trades	ESF/07-13/006	144,703.00	72,351.00	
8	None indicated	EMPASSIST 2008	Technical Assistance for the Employment Service	ESF/07-13/007	18,470.00	9,235.00	

importing tables from the web

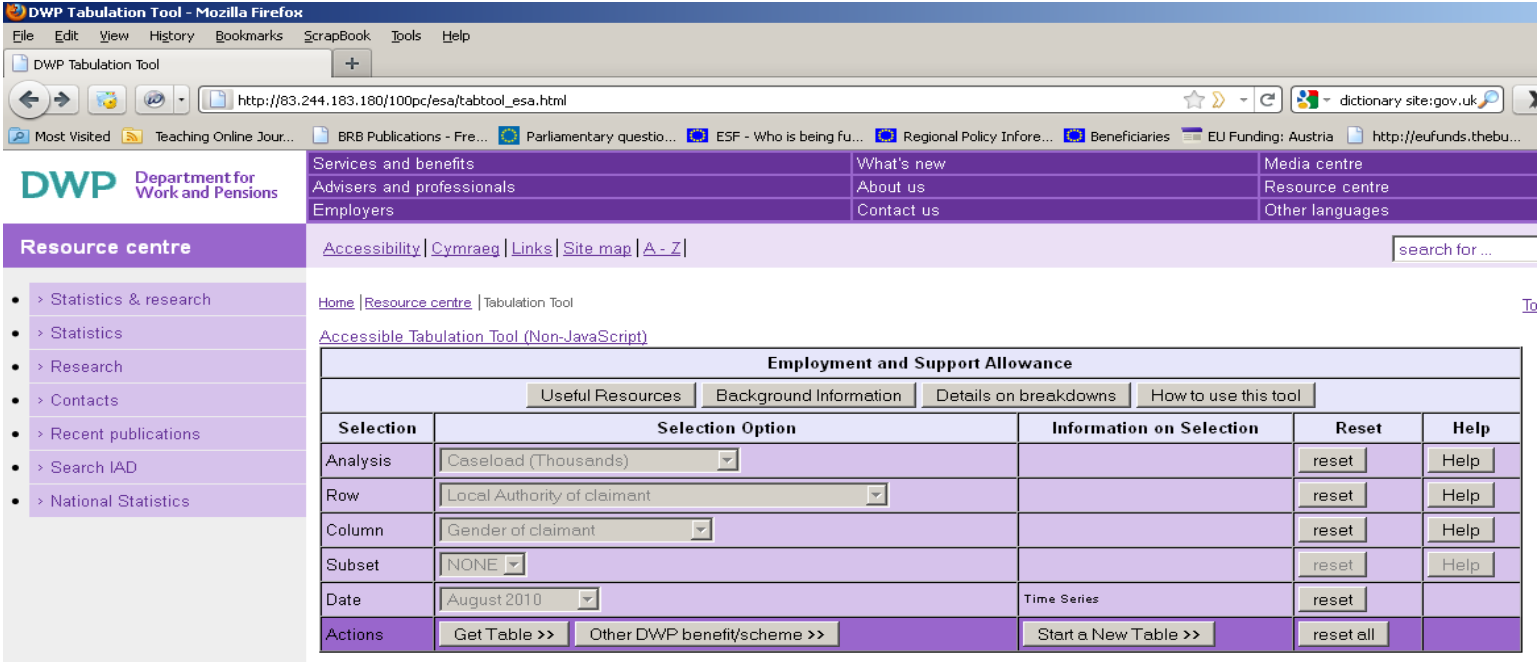
Not all data will be available for download as an Excel or CSV file, but it may be displayed on a website in a table. The cleanest way to acquire data in this case, is to use the import function in Excel, though often the information can be simply cut and pasted into Excel.

We will review importing into Excel by means of the following example:

The Department of Work and Pensions makes some of its data available through its ‘tab too’

http://statistics.dwp.gov.uk/asd/index.php?page=tabtool and http://83.244.183.180/100pc/tabtool.html

Select 'Employment and Support Allowance' and fill in the data like in this example and then click 'get table'.



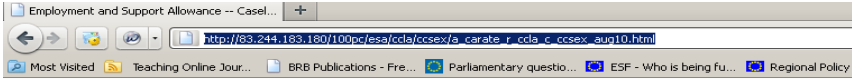
This will bring you to the following url:
http://83.244.183.180/100pc/esa/ccla/ccsex/a_carate_r_ccla_c_ccsex_aug10.html

and show a snapshot of the Department of Work and Pensions caseload for the Employment and Support allowance broken down by local authority.

Getting this data into Excel is very easy. You could cut and paste which often works if the html table is clean, but is not always an option. The best way is to use the 'import' function in Excel.

Open a new workbook and go to the data tab and click 'from web'. This will bring up a pop-up web browser. You can either navigate to the right website through this pop-up, or simply paste the correct url into it.

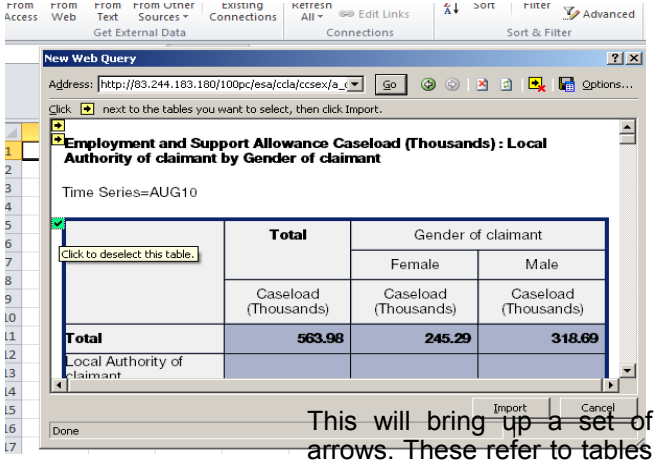
This will bring up a set of yellow boxes containing black arrows. These refer to tables in html code on the website and they allow you to decide which table to select. In this case, click the box highlighted in green. Selecting it turns the arrow into a tick box.



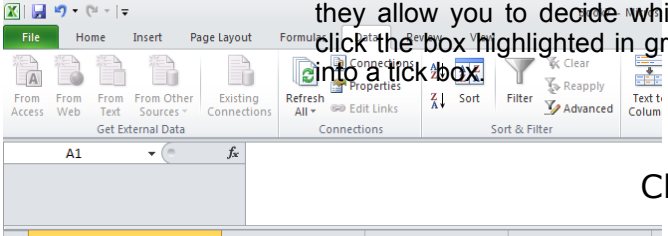
Employment and Support Allowance Caseload (Thousands) : Local Authority of claimant by Gende

Time Series=AUG10

	Total	Gender of claimant	
		Female	Male
	Caseload (Thousands)	Caseload (Thousands)	Caseload (Thousands
Total	**563.98**	**245.29**	**318.6**
Local Authority of claimant			
County Durham	**6.75**	2.82	3.9
Darlington	**1.04**	0.43	0.6
Hartlepool	**1.15**	0.50	0.6
Middlesbrough	**1.80**	0.75	1.0
Northumberland	**2.75**	1.16	1.5
Redcar and Cleveland	**1.42**	0.63	0.7
Stockton-on-Tees	**1.84**	0.78	1.0
Gateshead	**2.57**	1.06	1.5
Newcastle upon Tyne	**3.48**	1.44	2.0



This will bring up a set of y
arrows. These refer to tables in
they allow you to decide which
click the box highlighted in gre
into a tick box.



Clic

	A	B	C	D
1		Total	Gender of claimant	
2			Female	Male
3		Caseload (Thousands)	Caseload (Thousands)	Caseload (Thousands)
4	Total	563.98	245.29	318.69
5	Local Authority of claimant	6.75	2.82	3.94
6	County Durham			
7	Darlington	1.04	0.43	0.61
8	Hartlepool	1.15	0.5	0.65
9	Middlesbrough	1.8	0.75	1.05
10	Northumberland	2.75	1.16	1.59
11	Redcar and Cleveland	1.42	0.63	0.79
12	Stockton-on-Tees	1.84	0.78	1.07
13	Gateshead	2.57	1.06	1.51
14	Newcastle upon Tyne	3.48	1.44	2.05
15	North Tyneside	2.34	1.05	1.29
16	South Tyneside	2.24	0.96	1.28
17	Sunderland	4.06	1.7	2.36

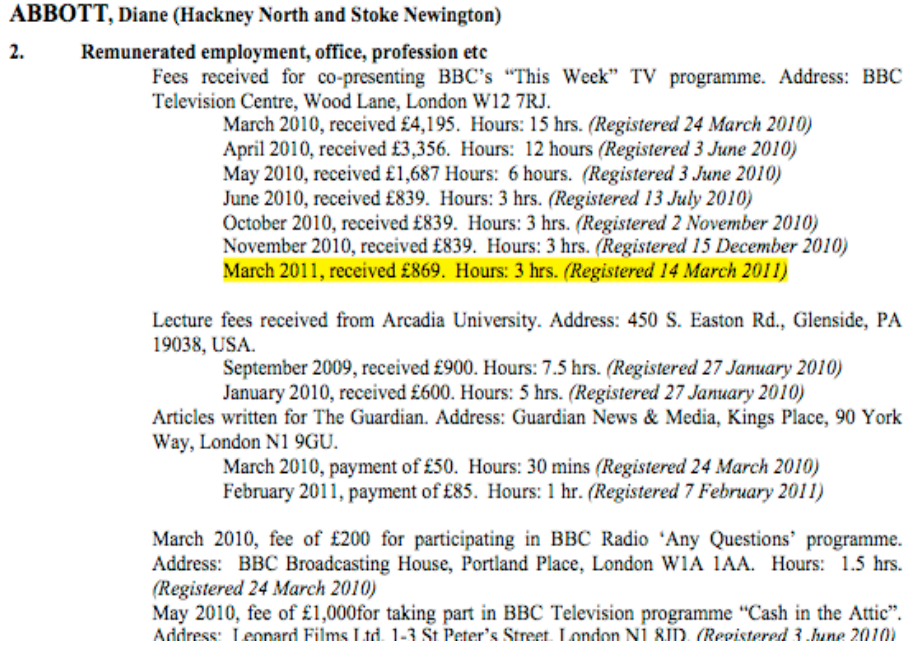
scraping data and importing data from PDF files

Why scraping?
Practicing data journalists find that the government's rhetoric on data transparency does not always match reality. Information is often presented in 'closed' formats such as pdf, presented in badly structured forms or else dispersed across many different websites making it difficult to pick up on trends.

The question to lawmakers is therefore: is transparency only a pro forma exercise or do they really mean it?

Consider, for example, the Register of MP's interests, available on the Parliament.UK website.
http://www.publications.parliament.uk/pa/cm/cmregmem/contents.htm

The individual entries are listed and laid out in a way that can only be quantified with great difficulty.



ABBOTT, Diane (Hackney North and Stoke Newington)

2. Remunerated employment, office, profession etc

Fees received for co-presenting BBC's "This Week" TV programme. Address: BBC Television Centre, Wood Lane, London W12 7RJ.
March 2010, received £4,195. Hours: 15 hrs. *(Registered 24 March 2010)*
April 2010, received £3,356. Hours: 12 hours *(Registered 3 June 2010)*
May 2010, received £1,687 Hours: 6 hours. *(Registered 3 June 2010)*
June 2010, received £839. Hours: 3 hrs. *(Registered 13 July 2010)*
October 2010, received £839. Hours: 3 hrs. *(Registered 2 November 2010)*
November 2010, received £839. Hours: 3 hrs. *(Registered 15 December 2010)*
March 2011, received £869. Hours: 3 hrs. *(Registered 14 March 2011)*

Lecture fees received from Arcadia University. Address: 450 S. Easton Rd., Glenside, PA 19038, USA.
September 2009, received £900. Hours: 7.5 hrs. *(Registered 27 January 2010)*
January 2010, received £600. Hours: 5 hrs. *(Registered 27 January 2010)*
Articles written for The Guardian. Address: Guardian News & Media, Kings Place, 90 York Way, London N1 9GU.
March 2010, payment of £50. Hours: 30 mins *(Registered 24 March 2010)*
February 2011, payment of £85. Hours: 1 hr. *(Registered 7 February 2011)*

March 2010, fee of £200 for participating in BBC Radio 'Any Questions' programme. Address: BBC Broadcasting House, Portland Place, London W1A 1AA. Hours: 1.5 hrs. *(Registered 24 March 2010)*
May 2010, fee of £1,000for taking part in BBC Television programme "Cash in the Attic". Address: Leopard Films Ltd, 1-3 St Peter's Street, London N1 8JD. *(Registered 3 June 2010)*

To answer a question such as 'How much money did news outlets pay Diane Abbott from March 2010 to March 2011?' a journalist would have to manually add together the figures listed in the register which runs over more than a page. If the data were presented in a structured format, a simple SUM formula in Excel would yield the answer in less than 10 seconds. A better and more transparent solution (by no means the only one) would be a structured table, with dates, institutions, amounts of money, such as for example:

MP Name	Type of Interest	Date	Amount	Organisation	Narration	Registered	Organisation Type
Diane Abbott	Remunerated employment, office, profession etc	March 2010	£4,195	BBC	Fees received for co-presenting BBC's 'This Week' TV programme.	*24 March 2010*	Media

The individual entries are listed and laid out in a way that can only be quantified with great difficulty.

There are ways to overcome the problems posed by badly presented data and data published in closed formats such as pdf. In the following section we will discuss two methods to scrape data.

MySociety's TheyWorkForYou website created an xml file containing a structured version of this data. (xml stands for extensible mark-up language and is used to structure, store and share date. It allows coders/users to 'tag' information using their own tags. In the example below <category> is a tag. Using a 'feed' – similar to an rss (really simple syndication) feed – a user can connect to the xml feed and get real-time updated information from the source.)

```
<regmem personid="uk.org.publicwhip/person/10029"
    memberid="uk.org.publicwhip/member/719" membername="Hugh Bayley"
    date="2005-04-11">
    <category type="4" name="Sponsorship or financial or material support">
        <item>I sponsor a parliamentary pass for a research assistant paid by
        the Royal African Society to enable her to work in support of the
        Africa All-Party Parliamentary Group, which I chair.</item>
    </category>
</regmem>
```

What is scraping?

Practicing data journalists find that the government's rhetoric on data transparency does not always match reality. Information is often presented in 'closed' formats such as pdf, presented in badly structured forms or else dispersed across many different websites making it difficult to pick up on trends.

The question to lawmakers is therefore: is transparency only a pro forma exercise or do they really mean it?

Consider, for example, the Register of MP's interests, available on the Parliament.UK website. http://www.publications.parliament.uk/pa/cm/cmregmem/contents.htm

The individual entries are listed and laid out in a way that can only be quantified with great difficulty.

To answer a question such as 'How much money did news outlets pay Diane Abbott from March 2010 to March 2011?' a journalist would have to manually add together the figures listed in the register which runs over more than a page. If the data were presented in a structured format, a simple SUM formula in Excel would yield the answer in less than 10 seconds. A better and more transparent solution (by no means the only one) would be a structured table, with dates, institutions, amounts of money, such as for example:

Webscraping for non-programmers

While some of the more complex websites can only be scraped using custom-built code, there are a number of out-of-the-box tools that journalists can use to scrape sites without having to learn how to code. Here is a selection of these tools:

Downthemall

The task of downloading large amounts of pages from a website is made easy by a Firefox plugin called DownThemAll.

iMacros https://addons.mozilla.org/en-US/firefox/addon/imacros-for-firefox

iMacros, also a Firefox extension, provides a number of templates for scraping and lets you 'record' html-based macros to extract data, fill in forms, and perform a host of other repetitive tasks. The iopus.com site (makers of imacros) have a host of tutorials on how to use the software.

OutWitHub http://www.outwit.com/products/hub/

Another Firefox extension. A very robust tool. , OutwitHub's interface includes a number set data extraction presets such as data, images, emails, but also lets the user make custom adjustments The extracted data can be exported to csv, html, Excel or sql databases.

Needlebase http://needlebase.com/

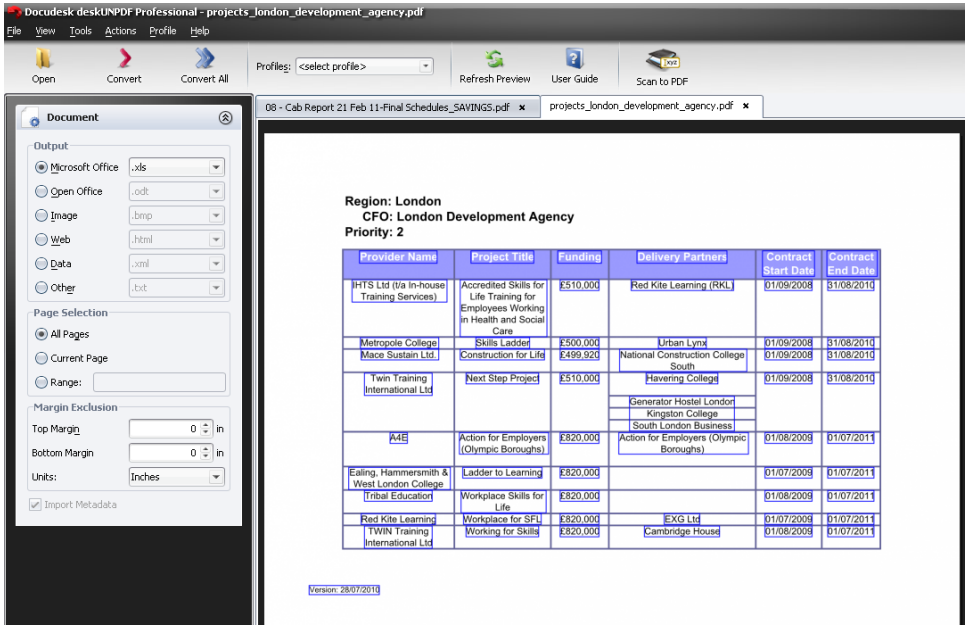
Needlebase - recently acquired by Google - is useful for grabbing data stored on multiple pages and sub pages. The user can "teach" the tool by pointing and clicking through one or two examples and showing it how the site is structured and which data fields you want to extract. Needle then follows this pattern to scrape data. Needle exports into Excel, CSV and XML. The tool can also be used to aggregate data from different sites and create mash-ups. Needlebase is quite pricey (it has a monthly pricing structure). There is a free version however, but it requires the scraped datasets to be public.

PDF scraping

Although some progress has been made in providing information in open formats, too many documents in the UK and elsewhere are still exclusively published as pdfs - certainly not an open format in the data world. It is astonishing how many government agencies still refuse to publish information in a structured form that can be checked and analysed. One official noted that his office had 'previous experience of mischievous manipulation and misrepresentation of Excel and Word documents,' and used this as an excuse to provide data only as a scanned pdf. (To reiterate the point made earlier in this book: misrepresented data and sloppy reporting will detract from the cause of open data and transparency. Be fair and be accurate.)

There are a large number of tools available to unlock and scrape data from PDF documents. It is possible to write code to extract data from pdfs using for example Ruby or Perl - two coding languages - but there are also a number of free or inexpensive, but powerful software tools that accomplish this task.

Essentially, they do the reverse of what a PDF writer tool does. Sometimes, scanned



documents have to be run through an optical character recognition tool, which are often built into PDF extraction tools. This makes cracking open pdfs more difficult and time-consuming, but will not - and should not - deter a determined journalist or citizen.

Some tried and tested tools are AbbyFine reader and UnPdf.

To extract data from a PDF using UnPdf, open the relevant file by selecting 'open' and navigating to the relevant file.

UnPdf recognises the individual cells, and places a blue border around them. It's worth scrolling through the dataset to make sure that the tool has correctly identified the cells. Sometimes, malformed pdfs make it difficult for the software to recognise cells and a manual clean-up is needed.

Case study: Europe's Hidden Billions

In 2010, the Financial Times, in collaboration with the Bureau of Investigative Journalism, tackled an important, but opaque area of EU funding, the European Structural Funds.

Though Structural Funds have existed for decades, there has been little transparency about how the funds are used. At the start of the current funding round in 2007, authorities – for the first time – were obliged to publish a list of beneficiaries of the programme, which represents more than a third of the EU budget.

The investigation started with a simple set of questions: who benefits from the €347bn allocated in the current budget round and does the subsidy achieve what it has set out to do.

Answering this question could not have been done without PDF scraping and database skills.

Contrary to what might have been expected, no central database containing all the beneficiary data existed. A team of multi-lingual reporters and researchers spent months creating a database of beneficiaries in order to analyse it to find the stories contained within. The database is available here: www.ft.com/eu-funds

The primary sources were the lists of beneficiaries of EU Structural Funds, published by authorities across 27 member states. This information was, in principle, freely available, but not in a way that could be meaningfully analysed. Creating the database required the team to scrape nearly 600 pdfs from more than 100 websites and in 21 languages.

Some data was held in online databases, and for the most part we were able to use a scraper to gather this information. However, most of the documents were held as pdf. This required extensive PDF to Excel transformation and included using software such as AbbyFine Reader and UnPdf.

The team used a shared Excel document to list all the documents, including links and translations of the headers. The individual data sets were output as CSV files, which were then uploaded into a master sql database.

Creating your own datasets

If the data does not exist in a structured form, it may be appropriate to create a database from scratch.

The advantage of this is that you can quantify information to underline trends that others may miss or can only refer to anecdotally. It makes for a much stronger story if there are numbers to confirm the trend and even better, the data set will be exclusive to you.

BP cited for safety lapses in North Sea

This story was done by combining a series of Freedom of Information requests with a simple analysis of an Excel table. The reporter asked for offshore inspection letters written by officials from the Department of Energy and Climate Change (DECC), which is in charge of monitoring compliance with companies' approved emergency plans in the North Sea. The letters were mostly scanned documents, and were provided in PDF format and because they were presented in narrative form, they were not laid out in a way that could easily be scraped for analysis.

Reviewing the information, the reporter noted that BP had frequently been cited for not complying with rules on oil spill training. The reporter compiled a spreadsheet listing the regulations and rules referred to in inspection letters and classified the individual infractions by type. This made it possible to pick up trends within the data and it revealed that the records of 11 of the 23 inspections carried out by DECC during the five years until the end of 2009 contain criticism of BP's training processes. Of those, eight inspection records on seven different facilities suggested the necessary training had not taken place.

However, compiling a database from scratch can be prone to error and must be done with caution. It can also be very time consuming.

Here are some tips, if you decide to go this route:

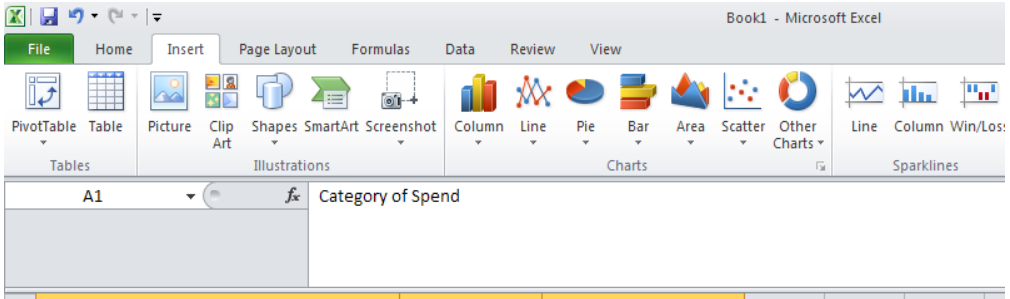
- Make sure nobody else has already done the work. Academics, NGOs or commercial organisations may have already compiled a database from the information you are seeking, but are not using it for journalistic purposes. Asking them to share the data may save a lot of time.
- Be sure that it's worth it and have a top line for the story in mind. It is only worth spending the time to compile a database if you are fairly certain that you will get a story.
- Plan ahead. Look through the information you want to turn into a database and select the type of data you need. Look for trends in what is included, but also pay attention to items that only appear exceptionally. They are often also worth capturing.
- If you are working in a team, make sure you map out all the fields you need and write-up instructions on how the fields need to be populated to keep everything consistent. You don't want to find out right before deadline that you have to go back through all your records and add a field.
- Double-check (triple-check) all entries
- Make the data available online to your readers.

A simple Excel spreadsheet may often be enough. In more complex cases you may need to build a relational database in Access, or SQL.

data visualisation

The goal of data visualisation is to make information easily understandable through graphics and design. Combining data with design should ideally help reveal trends or information that would otherwise be hidden or less evident.

In its simplest forms, data can be visualised as bar charts and line graphs. Stock prices or economic data, for example, have long been graphed in newspapers and graphics can give readers or viewers a quick overview over the major trends.



	A	B	C	D	E	F
1	Category of Spend	First Year	Second Year			
2	Administrative Costs	£ 250,000,000	£ 5,000,000			
3	Schools	£ 200,000,000	£ 250,000,000			
4	Housing	£ 20,000,000	£ 25,000,000			
5	Recreation and Tourism	£ 15,000,000	£ 1,000,000			
6	Salaries of Council Staff	£ 15,000,000	£ 40,000,000			
7	Refuse Collection	£ 25,000,000	£ 15,000,000			
8	Highways	£ 30,000,000	£ 25,000,000			
9	Child Protection Social Services	£ 50,000,000	£ 55,000,000			
10	Adult Social Services	£ 25,000,000	£ 15,000,000			
11						

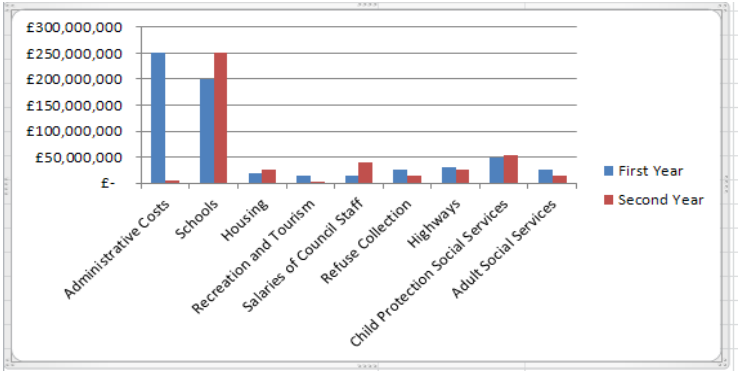
The use of data visualisation to tell stories has become increasingly popular in recent years as tools such as ManyEyes, Google Gapminder and Tableau Public are making it easier to create and share data. News outlets are using multimedia and interactive visualisations to tell sophisticated stories. Visualisations can also be used in the course of investigations, to reveal patterns or interesting outliers that merit further review.

While it is tempting to produce visualisations for the sake of a beautiful design, in a journalistic context, it is important to use clean and reliable data and represent it accurately. What use is an eye-catching visual if the underlying data is gathered from an unreliable source or misleading? Like any article, data visualisations should be accurate, fair and balanced.

Excel can create a wide variety of charts and sometimes this is enough for the purposes of reporting.

To create a chart in Excel, open your spreadsheet and select the data you would like to visualise. Click on the ‘insert’ tab.

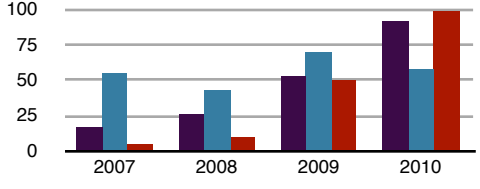
Chose the type of chart best suited to your task. Click on ‘column’ and then “all chart types” and you will get a full list of charts to choose from. The example (right) shows a simple bar chart:



To create a chart in Excel, open your spreadsheet and select the data you would like to visualise. Click on the ‘insert’ tab

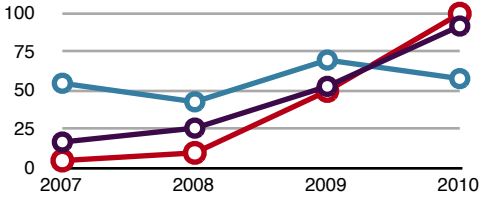
Types of visualisation

Here are some of the most commonly used charts:



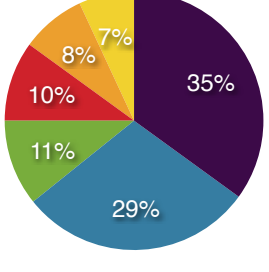
Bar charts

These are frequently used to compare a set of numerical values. To create a bar chart you need at least two columns or rows containing data series and corresponding labels. It can be used for positive and negative values. (See the chart above)



Line charts

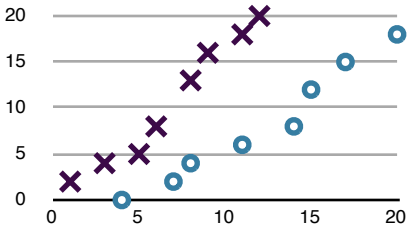
Line graphs are useful for visualising continuous change over time. Most often they are used for time-series data, with the time labels in the x-axis and the values in the y-axis. Line charts can accommodate several lines, showing changes for different categories. Line charts work well for changes over time spans such as months, quarters, or fiscal years. For example: the house price index.



Pie chart

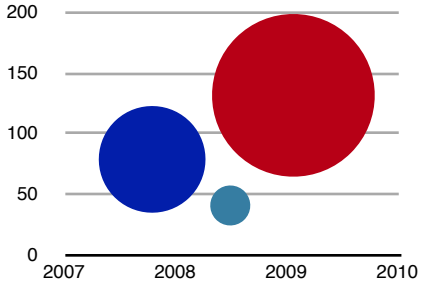
Pie charts are useful when the aim is to show how the size of values within a series relates to the sum, ie which individual categories the total is made up of. For this type of chart the data has to be arranged in one column or row. Values cannot be negative.

2007 2008 2009 2010 2011 2012



Scatter plot

Scatterplots let you show relationships between numeric values in several data series. They are often used comparing values such as scientific or statistical data. You can use it for example to compare pay versus performance.



Bubble charts

A bubble chart is a popular way to display a set of numeric values. They are useful if the difference between the data values graphed is very large. The area of the bubbles represents the value. Bubble charts can be presented in a scatter plot (see case study below), with the first two values representing the position of the bubbles on the x and y axes of the chart, and the third value representing the size of the bubble. However, bubble charts can also just show one value, by means of the size of the bubble and the circles can be packed together tightly to save space.

Word cloud

Tools like Many Eyes and Wordle allow you to import text to reveal word frequencies. To create a word cloud, also known as a tag cloud, you need to import either one or two text files and the tool takes care of the rest. Tag clouds can be a fun way to look at text, but there can be problems such as meaningless words like ‘the’ can be over-emphasised.

An example of a word cloud:
David Cameron’s July 2010 speech on the Big Society

case study: interactive scatterplot/bubble chart

FT analysis of executive pay in the oil and gas industry.

The Financial Times examined executive compensation in the oil and gas industry. When comparing the salaries of executives, readers may weigh elements of compensation and the performance differently and while in a static graph (egthe printed newspaper graphic) it is only possible to pick one of these, an interactive chart allows readers to apply their own judgement to the data.

For the FT visualisation of executive pay the reporters chose a scratterplot, because it allowed them to present the pay of 40 executives at the same time, while also demonstrating their comparative performance.

Placing the pay of executives on the scatterplot also made it possible to easily visualise which executives had the high levels of pay and negative share holder return (top left quadrant) and which one had one of the lowest compensation packages, but a positive shareholder return (bottom right)

The graphic is available at http://www.ft.com/ceo-pay

Using visualisation software tools

Over the past few years a host of new online visualisation tools have made visualising data increasingly accessible for non-coders. In the 2000s, the predominant tool for interactive visualisations has been Adobe Flash, but this is rapidly changing. The many new tools to visualise data are very simple and each of the following sites has extensive tutorials online. Most of them have free versions available. The drawback with the free versions is that all the visualisations are public, so it is not suitable for exploring projects that you are not ready to share. Most of the data visualisation software providers offer paid versions, though the licenses tend to be expensive.
Here are some of the most popular:

Tableau: http://www.tableausoftware.com/public/how-it-works
ManyEyes: http://www-958.ibm.com/software/data/cognos/manyeyes/
GoogleChart: http://code.google.com/apis/chart/
Google Fusion tables: http://www.google.com/fusiontables/
Omniscope: http://www.visokio.com/omniscope
BatchGeo: for mapping data

Common pitfalls

Even in well-established news outlets, graphics containing blatant errors either in captions or in the visualisation have been signed-off.

One of the most common mistakes occurs with bubble charts used to visualise comparisons in size.

The US GDP is about three times the size of China's but in the first visualisation, the bubble showing US GDP is many times larger. Bubbles are correctly sized by area (with radius proportionate to the square root of the radius).

Gross Domestic Product (2009)

United States $14.12
China $ 4.98
Japan $ 5.06
Germany $ 3.33
France $ 2.64

Tips

- A visualisation should tell a story, even if it is a simple bar chart. If all the values are the same in a bar chart, this doesn't make for a compelling visual image.
- Gather the data from reputable sources and check for accuracy.
- Make sure you chose the best visualisation type for the story.
- Label the chart values clearly. £40bn is not the same as £40m.
- Don't try and show too much – this will only muddy the image – the aim is to make trends in the data clearly visible, not to plot every possible data point.
- Be fair and accurate.

Further Reading

There are many more types of charts and many ways to creatively - and playfully - visualise data beyond the simple charts mentioned above. Here are a few resources worth reviewing for inspiration and further information:

The Visual Display of Quantitative Information
One of the many books by Edward Tufte, the godfather of data visualisation

http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html
Statistician Hans Rosling's TED talk on global trends in health and economics is a 'must-view' for anyone interested in the power of data visualisations

Information is beautiful / http://www.informationisbeautiful.net/
David McCandless blog and eponymous book
http://www.improving-visualisation.org/tags
This website shows examples of more than 80 different chart types

http://flowingdata.com/ Excellent blog run by a PhD in statistics

http://www.vizworld.com/tag/infographic/

http://infosthetics.com/

http://www.meryl.net/2008/01/22/175-data-and-information-visualisation-examples-and-resources/

THE CENTRE FOR INVESTIGATIVE JOURNALISM

Centre for Investigative Journalism
is a registered charity (1118602).

This handbook was made possible by a grant from the Open Society Institute www.soros.org

We would really value your thoughts and feedback on this handbook.
Please write to us at info@tcij.org

If you would like to suggest a new topic for a handbook, or know journalists/authors who could help write one, drop us a line at the above address.

For more in the handbook series, please visit our website www.tcij.org

City University London,
Journalism & Publishing,
Northampton Square,
London
EC1V 0HB

info@tcij.org | +44 (0) 207 040 8220 | www.tcij.org