Getting data into Excel

There are a number of ways to get data into Excel, but for all methods you have to start with a plan. You have to have a vision about your table – what columns does it have – what patterns can you see when you look at the information that you want to get into Excel.

**1. Entering data**

Before looking at methods for getting data automatically into Excel, you have to consider the possibility of having to enter the data in by hand. Sometimes cutting and pasting, or scanning and ocr-ing will not just work. And when entering data by hand there are a few thing to think about:

* Think about what story you want to write and make sure you have columns that covers that
* Be consistent in entering your data, make sure that you have a rule for interpreting certain cases and stick to that rule
* Be open for the possibility that you will have to change your rule – or add new columns as your work continues. And in that case, you have to go back and re-enter the previous data
* If things are very sensitive, consider entering it twice – or getting a friend to help out
* If there is a lot of data, maybe Mechanical Turk is the way to go

**2. Using Excel’s function for getting data from a html table**

If you find a table online that you want to get into Excel, you might want to start with just selecting the table, using crtl-C to copy and ctrl-V to paste the table. Sometimes that works just fine. Another way is to use Excel’s external data function. Google wikipedia US states to find the wikipedia entry that has a list of all states and a lot of information for each state: <https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States>

Select the url and copy it to memory. Then open up Excel and open a new sheet. Go to the Data tab and click on get data from the web. A window opens for importing data and you can paste the url into the box on top of the screen. Then press Go. You’ll see your page in the window below. Keep a look out for small yellow arrows. If you are lucky you’ll find a small yellow arrow next to your table, but even if you don’t, it’s ok as long as you find one. For the US states there is one at the top of the screen. Click on that, turning it into a green check mark instead. Click on Import in the bottom of the window and just click OK when Excel asks you where to put the data. After a couple of seconds all the text from that wikipedia page ends up in your Excel sheet.

When you scroll down you see that our table is very neatly put into columns in Excel, and all we really need to do is delete all the rows that we don’t need, both before and after the table. You might want to clean up your data a little bit, get rid of the columns that you don’t want.

**3. Textfiles and csvs**

Sometimes data will come in the form of a textfile. Usually the textfile has some kind of separator between columns, such as comma, semicolon or a vertical bar. A csv-file is just a textfile where the columns are separated by a comma. It’s important to know that even though windows and Excel will treat a csv as a spreadsheet, it’s just a textfile.

One thing that Excel does is that it assumes what data format your data has. The program tries to identify weather your column is a text column like names, places other information – or a numeric column with numbers. You’ll usually see that a text column is aligned to the right, where as a numeric column is aligned to the left. In some cases Excel’s assumptions will distort our data.

We’re going to look at FEC data - donations to SuperPACs. FEC has a page for downloading this data, however that file is way too big to use for Excel. I have made a smaller copy for us to use. FEC does the date that the donation was made in a peculiar way, so March 8, 2016 will be written 03082016. The problem with uploading the textfile into Excel is that Excel considers that a number and will change the data to 3082016 instead.

Go to [www.helenabengtsson.se/material](http://www.helenabengtsson.se/material). In the middle of the page you see a file called SuperPac2, right click on the file and download it to your computer. Then open Excel.

Choose File and Open and navigate to the folder where you saved the file. Then change the extension so that Excel looks for .txt files instead. When you click on the file a window for importing data opens. First thing to decide is if the file has fixed length columns or separated columns. In this case it’s a separated file. Also take a look at the file format – you can choose a number of formats and that could really help if you have a file where some characters doesn’t import correctly. Foreign letters for instance works usually better if you choose UTF-8 as your file format. Click on next.

Now we tell Excel the separator – and the FEC has chosen a vertical bar as that. As you can see there is a number of separators to choose from, but you could also enter one yourself. And that is what we have to do. Put a vertical bar in the box and look at the data so that you see that it changes into columns. And click on next.

Last step in the import – but the most important. Here you can set the datatype for each column. Excel will put everything as General, but you can change that. Use the slider to get to the transaction\_date column and select it. Then change the data type from General to Text. Click Finish to import your data. Everything looks fine? No, the zip code seems to have been imported as a number. It might not be a problem, but if you want to be sure you might want to import the file again, making both the zip and the date into text. In fact, if I know that I have a problematic file that I want to work with I might import all columns as text to start with. I can always change the text data type into a number, but it’s harder to fix the different assumptions that Excel does.

**4. Cutting and pasting data from the internet**

So, what if the information you’ve found isn’t in columns to start with? Go to [www.senate.gov/senators/contact/](http://www.senate.gov/senators/contact/) to find a list of all US senators. Let’s ignore the fact that there is a xml download for this and use it as an example anyway. Select the whole list and copy it to memory. Open up Word and make sure that you paste it into word using the unformatted text option. Examine your data – look for patterns. A tip is to click on ”show hidden characters” (¶) to see all spaces, tabs and enters that you don’t normally see.

We see that every senator has four lines of information, the name, the address, the phone number and the email. And we also see that each senator starts with the words ”horizontal line”. That is a pattern we can use. Let’s think about our goal. We would like to have all information about a senator on one row, with different information in separate columns. We have a start for each row and we need to delete all the new line characters, except for the one following ”horizontal line”.

So we can actually start by taking out all the new line characters. Open Find and Replace and write ^p in the box for Find. Then write ^t in the box for Replace. This means take all new lines and replace them with tabs. Click Replace all. About 500 characters are replaced and the document looks a little scary.

Next step is to add in the new line where we want it. Open Find and Replace again and enter horizontal line in the Find box and ^p in the Replace box. Click Replace all, and you should get 100 replacements made. Close the Find and Replace window and look at your data. Looks better, but we see that each row starts with a tab. Let’s eliminate that before copying our data to Excel. We do that by using Find and Replace again.

How do we make sure that we just delete the tab in front of the name, and no other tabs. Again we have to look for the pattern of the data. That tab starts a line. And that is the pattern we can use. Open up the Find and Replace window and write ^p^t in the Find box. In the Replace box we write just ^p to keep the new line, but get rid of the tab. Click Replace all.

Select all text in your word document with ctrl-A and copy it with ctrl-C. Then open Excel. Click in cell A1 and paste the text into Excel with ctrl-V. And there is your table. You can of course use Text to columns to separate out the state and party. And you might want to get rid of the ”Contact:” in the last column.

For extra credit: Go to <http://www.europarl.europa.eu/> and click on MEPs. Then choose a country, for example UK and look at the list of all the members. Use copy and paste unformatted text into Word to get all 73 members from the UK. Then look at the data and try to find a pattern that will help you separate each member. Make a table of all the information about the members.

(Tip: Make sure you look at the hidden characters, maybe there is a pattern there that you can use)

Helena Bengtsson

The Guardian

helena.bengtsson@theguardian.com