Importing in Excel

In a perfect world, government officials would easily give us the exact data we want in a perfect, clean Excel spreadsheet. (You laughing yet?)

The reality is that journalists often receive data in format that we just can't "double click" and open in Excel. When this happens we need to import it into Excel. We can easily import files into Excel from a variety of file formats including PDFs and some data you find online. Follow this guide to understand more about the different file types and how you can get them into Excel and ready for analysis.

FILE EXTENSIONS

It's easy to tell what kind of file you're working with by looking at the extension following the file name after the period. For example, .doc is a Microsoft Word document and a filename ending with .xls is an Excel file. The most common file types you will be dealing with in data journalism have the following file extensions:

.xls or **.xlsx –** Excel file. These are Excel's native files and you are a master at those by now! Note: .xls is for Excel versions 2003 and older and .xlsx are for versions 2007 and newer.

.txt – Text file, sometimes called a flat file, is a Plain Jane file. There are two types of text files: The data may be aligned in columns, which is called a fixed width text file; or the columns may be separated with a symbol or character such as a comma or Tab, which is called a delimited text file.

.csv – Comma-separated values is a type of delimited text file. Commas separate each column of data. Excel loves .csv and will sometimes open them directly into Excel without having to go through the import process.

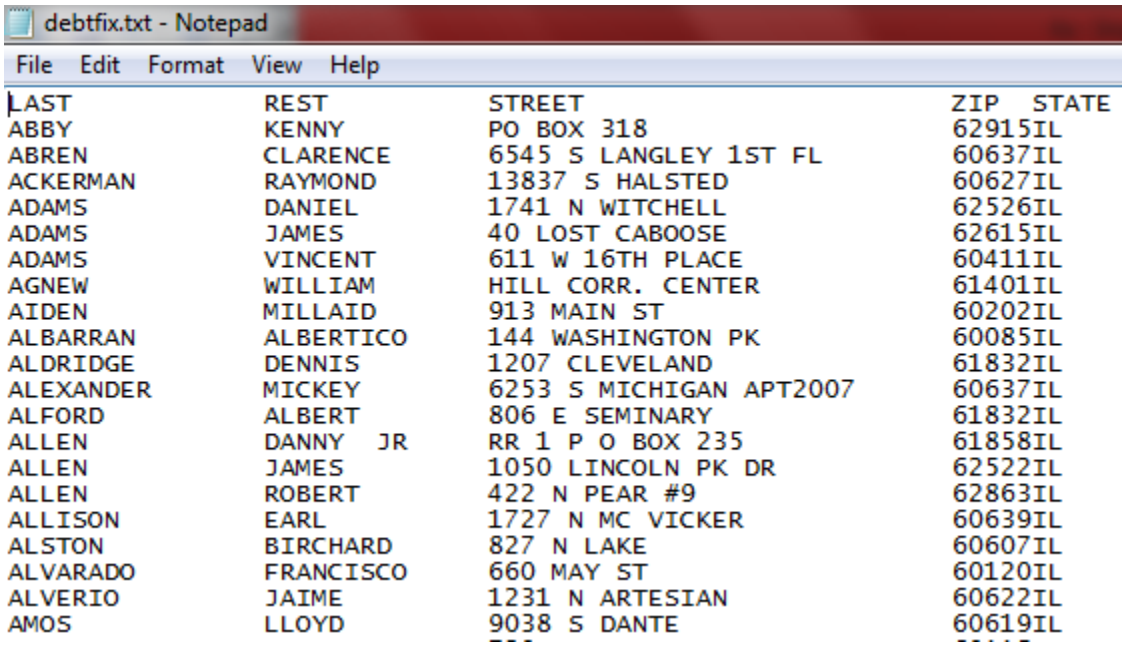
.pdf – PDF. Government agencies are making a habit of releasing data in electronic pdfs. Yes, there is a way to import that into Excel.

Most of the major database management systems that government entities operate on have the ability to export information to .txt, .csv and sometimes .xls. Text files are seen as the universal file when it comes to data. This allows us to work with the government data even when we don't have the same software.

GETTING TO KNOW TEXT FILES

Let's view the two most common formats: fixed width and delimited text files. Double-click on debtfix.txt to open. This is Illinois deadbeat parents data in fixed width format.

Each row represents one court order for back child support due. You see where the columns of data might fit in Excel? This is beautiful on the surface, but somewhat ugly in practice in the data world.

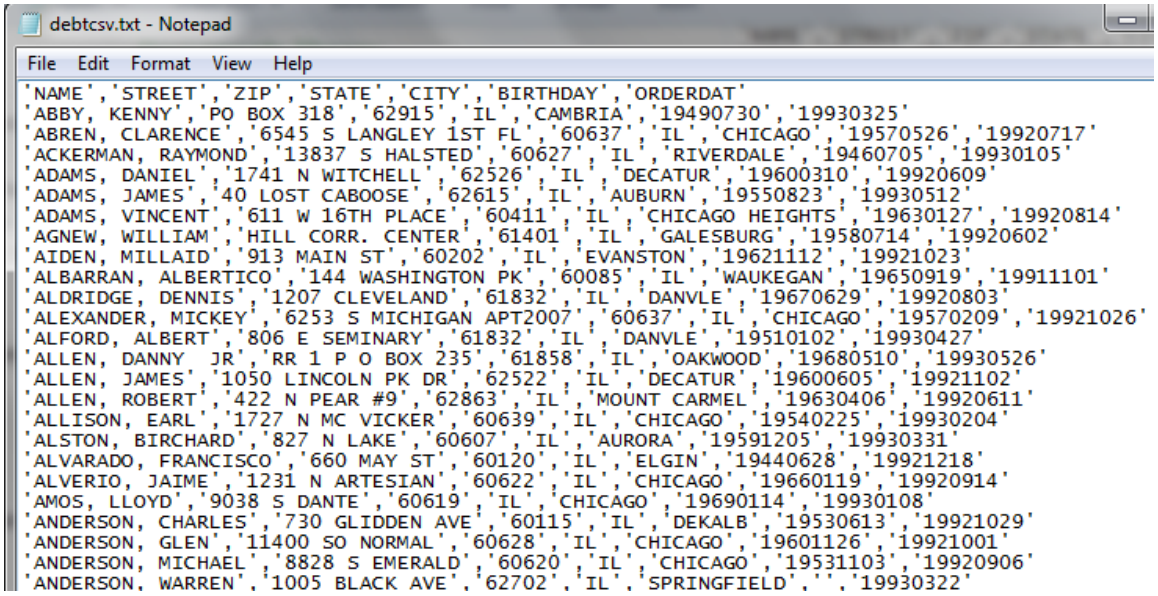


debtfix.txt - Notepad

File Edit Format View Help

LAST	REST	STREET	ZIP STATE
ABBY	KENNY	PO BOX 318	62915IL
ABREN	CLARENCE	6545 S LANGLEY 1ST FL	60637IL
ACKERMAN	RAYMOND	13837 S HALSTED	60627IL
ADAMS	DANIEL	1741 N WITCHELL	62526IL
ADAMS	JAMES	40 LOST CABOOSE	62615IL
ADAMS	VINCENT	611 W 16TH PLACE	60411IL
AGNEW	WILLIAM	HILL CORR. CENTER	61401IL
AIDEN	MILLAID	913 MAIN ST	60202IL
ALBARRAN	ALBERTICO	144 WASHINGTON PK	60085IL
ALDRIDGE	DENNIS	1207 CLEVELAND	61832IL
ALEXANDER	MICKEY	6253 S MICHIGAN APT2007	60637IL
ALFORD	ALBERT	806 E SEMINARY	61832IL
ALLEN	DANNY JR	RR 1 P O BOX 235	61858IL
ALLEN	JAMES	1050 LINCOLN PK DR	62522IL
ALLEN	ROBERT	422 N PEAR #9	62863IL
ALLISON	EARL	1727 N MC VICKER	60639IL
ALSTON	BIRCHARD	827 N LAKE	60607IL
ALVARADO	FRANCISCO	660 MAY ST	60120IL
ALVERIO	JAIME	1231 N ARTESIAN	60622IL
AMOS	LLOYD	9038 S DANTE	60619IL

Even though it looks nice it can be harder to import than its delimited sister. Close that file and open debtcsv.txt. Ugly, right? So this type of file is ugly on the surface, but beautiful in data world because it's easier to import.



debtcsv.txt - Notepad

File Edit Format View Help

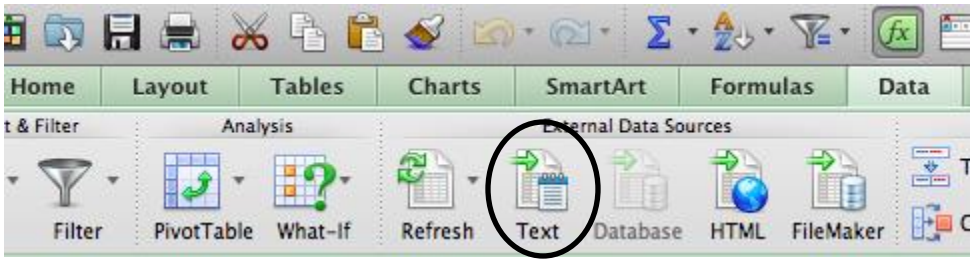
```
'NAME','STREET','ZIP','STATE','CITY','BIRTHDAY','ORDERDAT'
'ABBY, KENNY','PO BOX 318','62915','IL','CAMBRIA','19490730','19930325'
'ABREN, CLARENCE','6545 S LANGLEY 1ST FL','60637','IL','CHICAGO','19570526','19920717'
'ACKERMAN, RAYMOND','13837 S HALSTED','60627','IL','RIVERDALE','19460705','19930105'
'ADAMS, DANIEL','1741 N WITCHELL','62526','IL','DECATUR','19600310','19920609'
'ADAMS, JAMES','40 LOST CABOOSE','62615','IL','AUBURN','19550823','19930512'
'ADAMS, VINCENT','611 W 16TH PLACE','60411','IL','CHICAGO HEIGHTS','19630127','19920814'
'AGNEW, WILLIAM','HILL CORR. CENTER','61401','IL','GALESBURG','19580714','19920602'
'AIDEN, MILLAID','913 MAIN ST','60202','IL','EVANSTON','19621112','19921023'
'ALBARRAN, ALBERTICO','144 WASHINGTON PK','60085','IL','WAUKEGAN','19650919','19911101'
'ALDRIDGE, DENNIS','1207 CLEVELAND','61832','IL','DANVLE','19670629','19920803'
'ALEXANDER, MICKEY','6253 S MICHIGAN APT2007','60637','IL','CHICAGO','19570209','19921026'
'ALFORD, ALBERT','806 E SEMINARY','61832','IL','DANVLE','19510102','19930427'
'ALLEN, DANNY  JR','RR 1 P O BOX 235','61858','IL','OAKWOOD','19680510','19930526'
'ALLEN, JAMES','1050 LINCOLN PK DR','62522','IL','DECATUR','19600605','19921102'
'ALLEN, ROBERT','422 N PEAR #9','62863','IL','MOUNT CARMEL','19630406','19920611'
'ALLISON, EARL','1727 N MC VICKER','60639','IL','CHICAGO','19540225','19930204'
'ALSTON, BIRCHARD','827 N LAKE','60607','IL','AURORA','19591205','19930331'
'ALVARADO, FRANCISCO','660 MAY ST','60120','IL','ELGIN','19440628','19921218'
'ALVERIO, JAIME','1231 N ARTESIAN','60622','IL','CHICAGO','19660119','19920914'
'AMOS, LLOYD','9038 S DANTE','60619','IL','CHICAGO','19690114','19930108'
'ANDERSON, CHARLES','730 GLIDDEN AVE','60115','IL','DEKALB','19530613','19921029'
'ANDERSON, GLEN','11400 SO NORMAL','60628','IL','CHICAGO','19601126','19921001'
'ANDERSON, MICHAEL','8828 S EMERALD','60620','IL','CHICAGO','19531103','19920906'
'ANDERSON, WARREN','1005 BLACK AVE','62702','IL','SPRINGFIELD','','19930322'
```

Can you see that there are commas separating the columns? These commas are known as delimiters. When Excel looks at this kind of a file during an import, every time it sees a

comma it knows that's where a column break should be. While commas are the most common delimiters, other characters can be used to tell Excel "this data goes in this column," such as tab, space, even $. Close this file. Let's import!

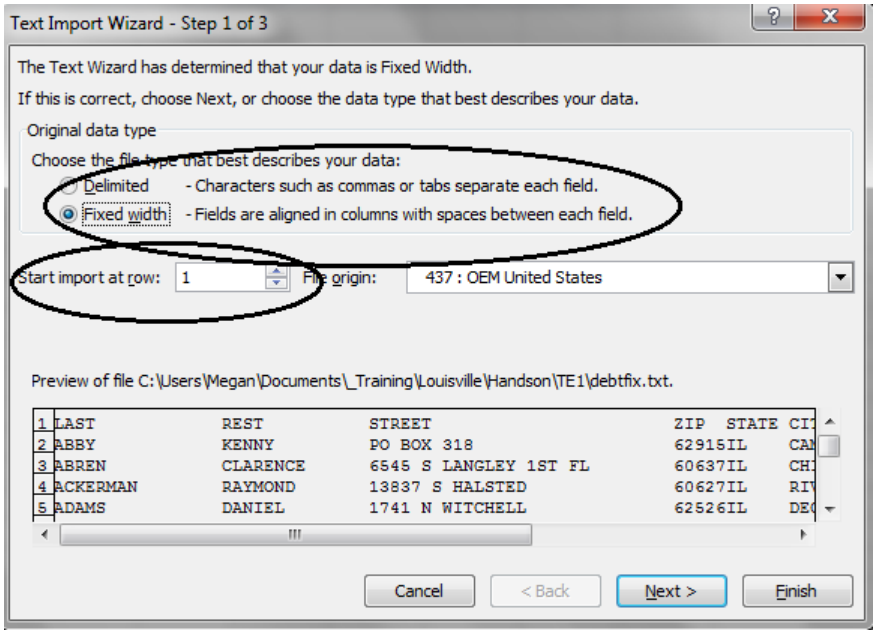
IMPORTING FIXED WIDTH TEXT FILES

Let's import the fixed width file first. Open a blank spreadsheet in Excel. Click on the data tab and choose "From Text".



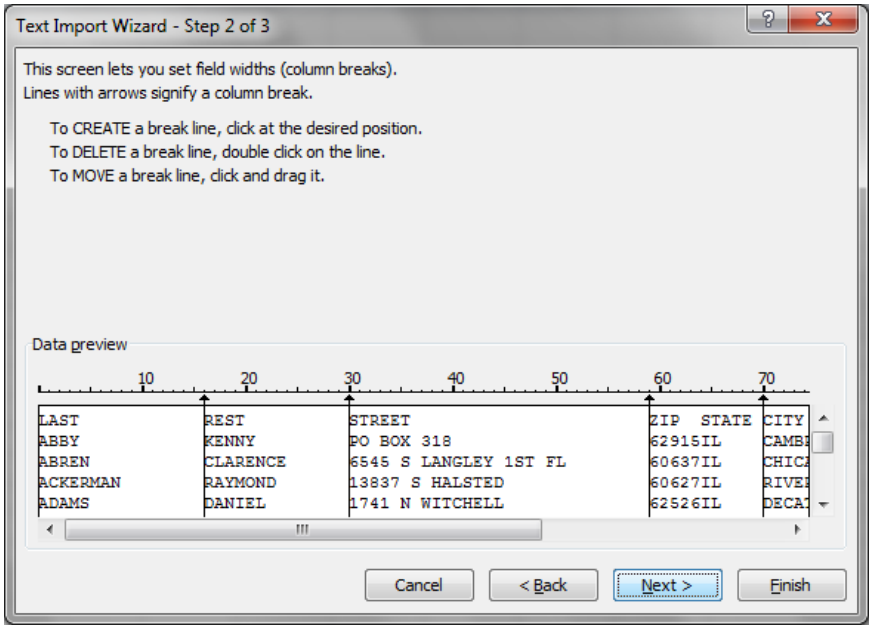
Navigate to the debtfix.txt file and open.

This is the Text Import Wizard. Tada! No magic wand, but it does some cool things. Excel takes a guess about the type of file we are trying to import: Fixed width. It also offers the definitions, in case we forget.



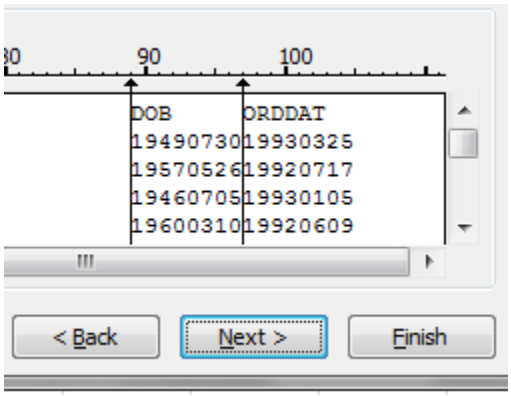
The other item to notice is you can decide what row to start importing at. Sometimes there will be verbiage at the top from the government entity. So you can always skip that and import at a later row. We don't need to change anything on Step 1, so go ahead and click "Next."

This is where fixed width gets ugly. We have to draw lines to tell Excel where the column breaks belong. Excel took a stab at it. Did Excel guess correctly?



No. We need to split ZIP and STATE columns. To do so, click between ZIP and STATE and it will add a line. Don't worry if it's in the in wrong spot, you can click and hold and move the cursor right before the "S" on STATE. If you get super click happy, you can just double-click on the line and it will disappear.

If you scroll to the right, you will also notice that DOB and ORDDAT also need to be split. Go ahead and add a line right before the "O" on ORDDAT. Click "Next."

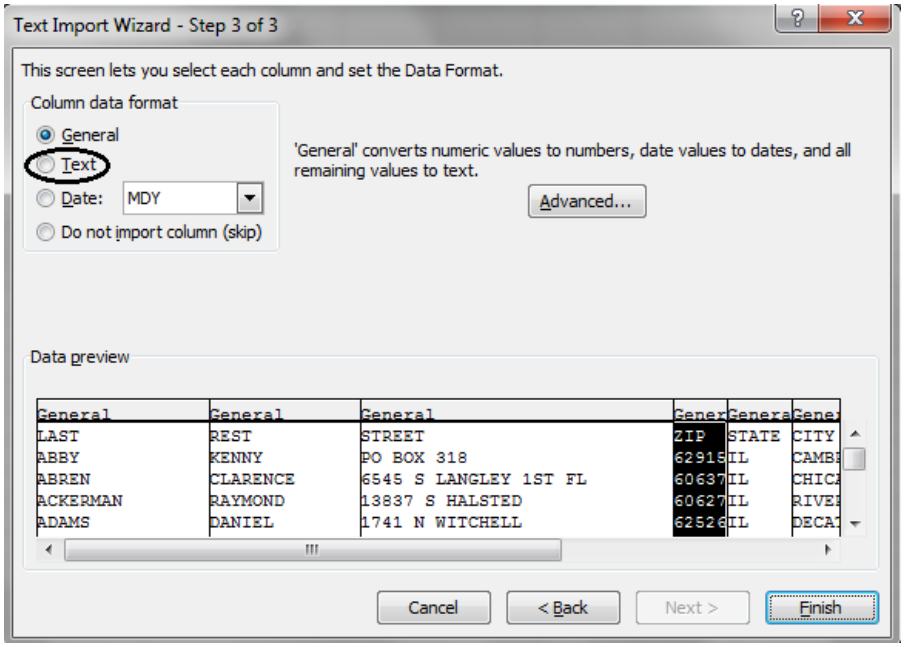


By the way, if you forget to add a line, just click "Back" then add it in and hit "Next" again.

Here is where we tell Excel what *type* of data we are importing. Data types are very important. Essentially, we need to tell Excel which columns contain numbers, which columns contain dates, etc. If we skip this step, we might not be able to sort our data correctly or Excel may change our data inadvertently.

Excel will default all of the columns of data to "General," which means it will treat a number like a number and a letter like a letter. The problem is when you have columns with numbered codes, like ZIP. Would we ever average or sum together a ZIP? No. A good rule of thumb with numbers is this: If you're never going to do math on a number (phone number, address, ZIP, codes, etc.) make the data type text.

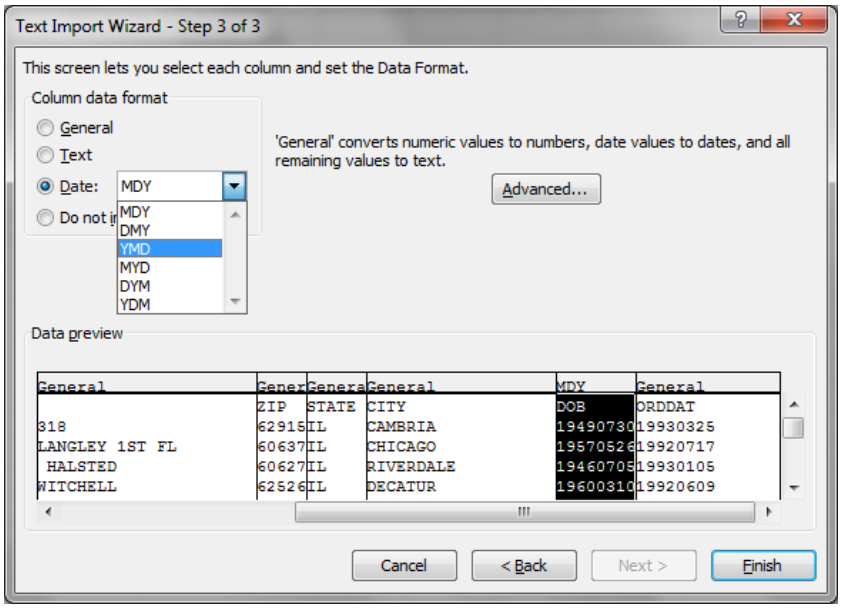
Let's change that column to text so Excel won't read it like a number. Click on ZIP and it will turn black. Choose "Text" in the upper-left corner.



We also need to change the date columns. Scroll over to the right until you see DOB and ORDAT. The DOB is date of birth and ORDAT is the date of the court order. If we leave these columns as "General," Excel will see them as numbers. Then if we want to sort our data by most recent order date it wouldn't sort correctly.

Click on DOB so it turns black. Choose "Date" under "Column data format." Notice the first DOB is "19490730." Our brains can easily see that this is 7/30/1949 but Excel needs more help. We then need to tell Excel what order the different pieces of the date can be

found. If you look closely, it's year, month, day. Choose the drop down menu next to "Date" and choose "YMD". Do the same for ORDDAT and click finish.

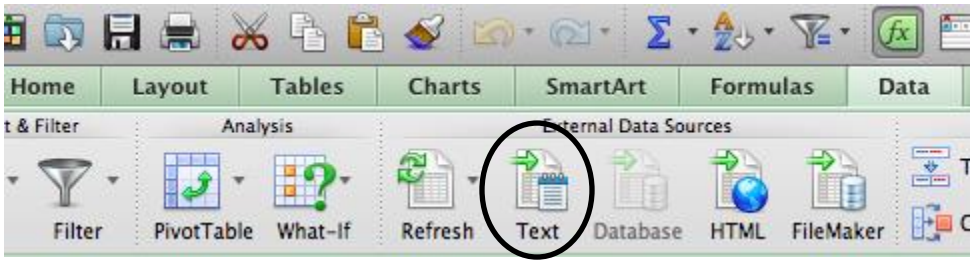


Excel will then ask you where you want to put the data. It should default to =A1, so just click "OK." The last step is to save this file as an Excel workbook. Tada!

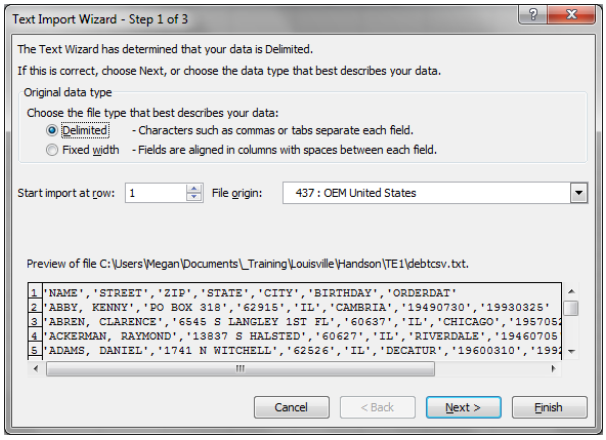
Can you imagine drawing lines for data with 30 columns of data? You see why fixed width might be ugly in data world?

IMPORTING DELIMITED TEXT FILES

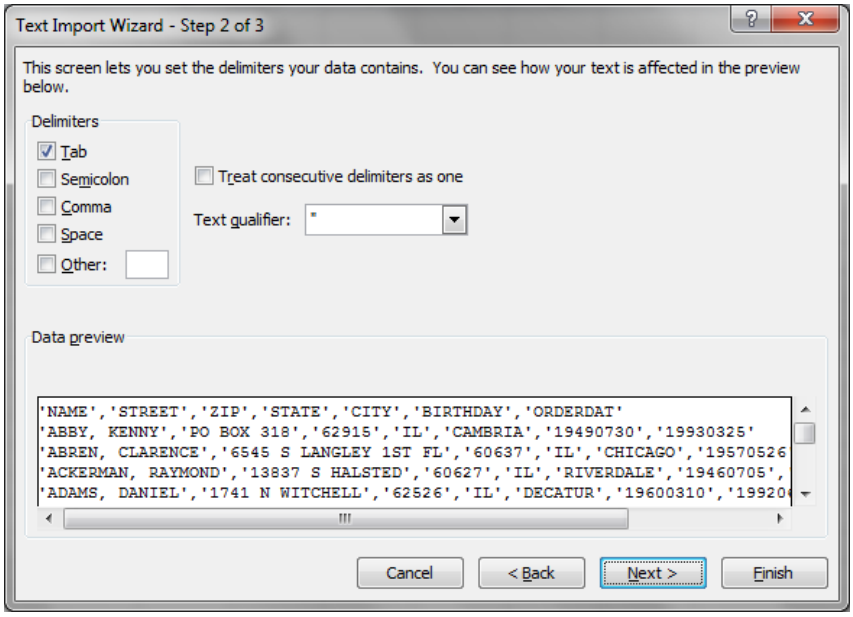
Let's import debtcsv.txt. First, make sure you have a blank spreadsheet in Excel. Then click on the "Data" tab at the top and choose "From Text."



Navigate to the debtcsv file and click "Open". You'll see the same Import Wizard Excel gave us last time.

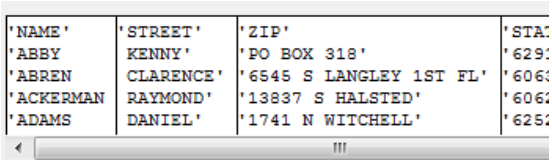


Excel guesses we are importing a "Delimited" file and it's correct. We want to start import at row 1, so click "Next".



Excel takes a guess that "Tab" is our delimiter – what separates the columns. Did Excel guess correctly? No. What is our delimiter? So uncheck "Tab" and check "Comma."

Everything look right?

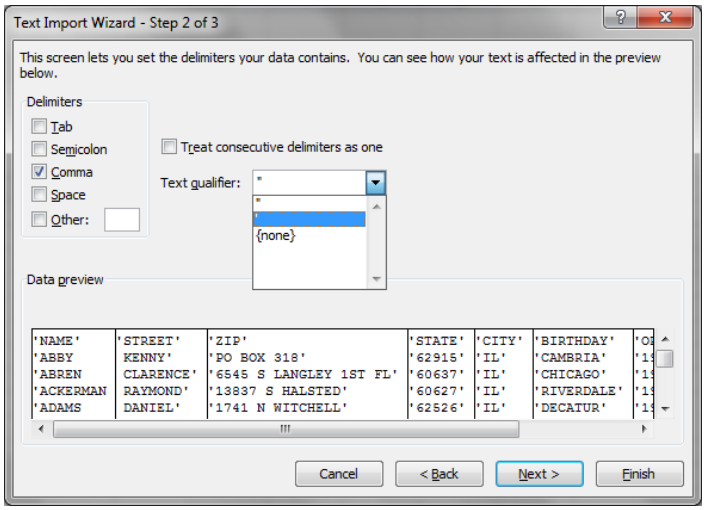


'NAME'	'STREET'	'ZIP'	'STA
'ABBY	KENNY'	'PO BOX 318'	'629
'ABREN	CLARENCE'	'6545 S LANGLEY 1ST FL'	'606
'ACKERMAN	RAYMOND'	'13837 S HALSTED'	'606
'ADAMS	DANIEL'	'1741 N WITCHELL'	'625

No. If you look at the figure above, you can see that the column names do not match up with the data. Under "STATE" are the ZIP codes. Notice the ' around the data. That is called a text qualifier. It basically hugs the data and tells Excel *this* group of data should

be in one column. Text qualifier is needed when you have data, such as the name column, which contains last name, first name. The comma is part of the data, not a delimiter, which is why Excel is confused. It thinks the comma in the NAME column is a delimiter, telling Excel to split the names.

Let's fix this. Choose the drop down menu by "Text qualifier" and choose the single quote. Look better? Click "Next."



Again, we need change the ZIP format to text and the BIRTHDAY and ORDDAT to Date, YMD. Click "Finish" and click "OK." Save the file as an Excel Workbook.

You see why delimited text files are a little easier to deal with?

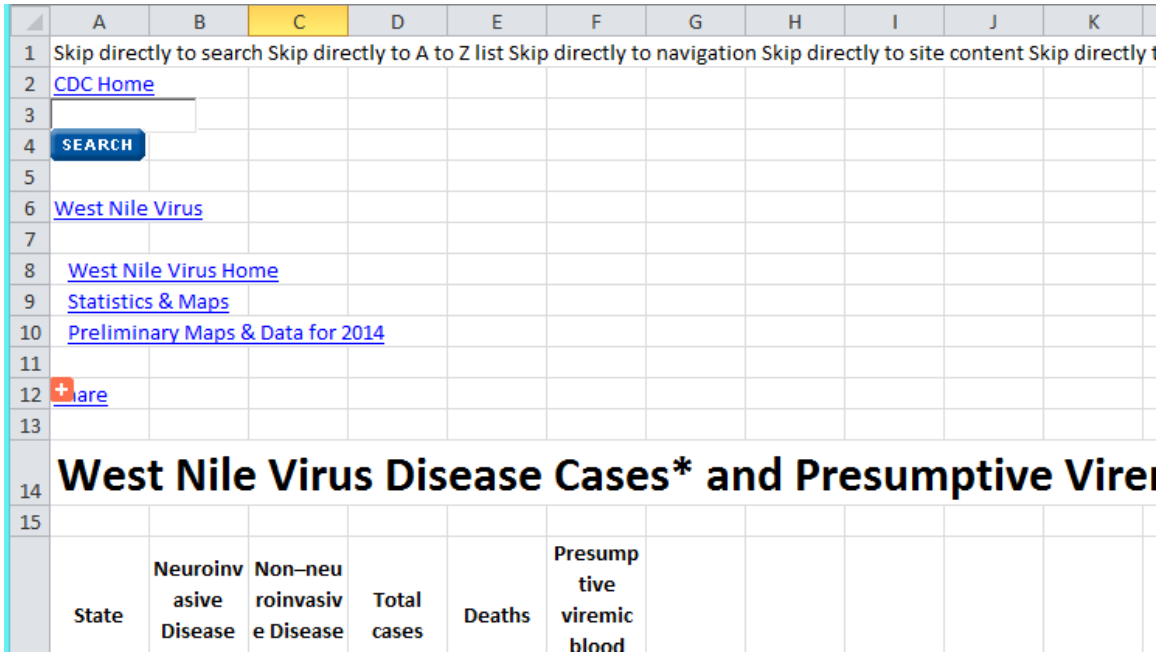
IMPORTING FROM THE WEB

Have you ever seen a data table on a website and tried to copy and paste it into Excel? The majority of time it doesn't work. The PC version of Excel offers an option to import directly from the Web but you can easily do the same thing in the Mac version of Excel in a few simple steps.

Check out this URL:
http://www.cdc.gov/westnile/statsMaps/preliminaryMapsData/histatedate.html

It's the up-to-date number of West Nile Virus cases by state. If we want to analyze this data further in Excel, we first need to get it into a spreadsheet. To do that, first highlight the entire webpage that contains this data using Command+A. Copy the information using Command+C.

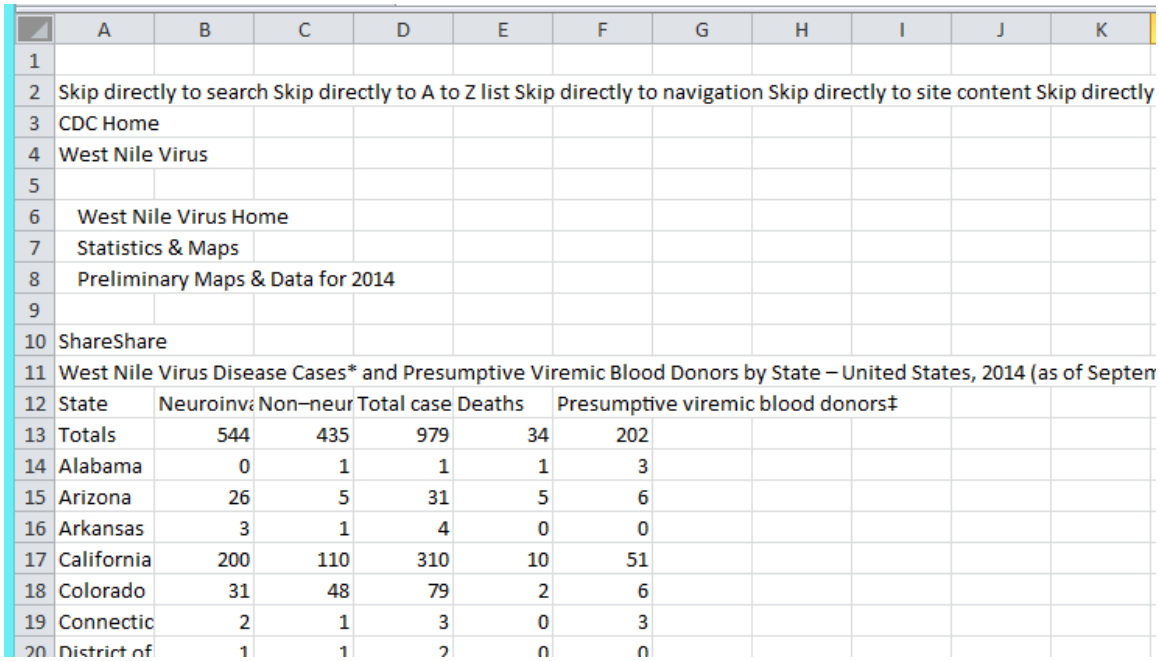
Next, open up a blank Excel workbook. Put your cursor in cell A1 of the first sheet and paste in the information using Command+V. It should look something like the image below.



	A	B	C	D	E	F	G	H	I	J	K
1	Skip directly to search Skip directly to A to Z list Skip directly to navigation Skip directly to site content Skip directly t										
2	CDC Home										
3											
4	SEARCH										
5											
6	West Nile Virus										
7											
8	West Nile Virus Home										
9	Statistics & Maps										
10	Preliminary Maps & Data for 2014										
11											
12	are										
13											
14	**West Nile Virus Disease Cases* and Presumptive Vire**										
15											
	State	**Neuroinvasive Disease**	**Non–neuroinvasive Disease**	**Total cases**	**Deaths**	**Presumptive viremic blood**					

Notice that this is pretty messy. A lot of the information at the top of the page is hyperlinked so if you clicked on it a browser would open. Let's try something different to get around this issue. You still have the copied webpage available on your clipboard so we can paste it again. This time, let's click on Sheet2 and in cell A1, right-click, then select "Paste special," and "Text."

You get the same thing only all of the hyperlinks and formatting have been stripped out of the pasted information – much less messy to deal with.



	A	B	C	D	E	F	G	H	I	J	K
1											
2	Skip directly to search Skip directly to A to Z list Skip directly to navigation Skip directly to site content Skip directly										
3	CDC Home										
4	West Nile Virus										
5											
6	West Nile Virus Home										
7	Statistics & Maps										
8	Preliminary Maps & Data for 2014										
9											
10	ShareShare										
11	West Nile Virus Disease Cases* and Presumptive Viremic Blood Donors by State – United States, 2014 (as of Septem										
12	State	Neuroinva	Non–neur	Total case	Deaths	Presumptive viremic blood donors‡					
13	Totals	544	435	979	34	202					
14	Alabama	0	1	1	1	3					
15	Arizona	26	5	31	5	6					
16	Arkansas	3	1	4	0	0					
17	California	200	110	310	10	51					
18	Colorado	31	48	79	2	6					
19	Connectic	2	1	3	0	3					
20	District of	1	1	2	0	0					

Now, all you need to do is clean this up so all that remains is the data table. You can either delete the rows above and below the table (right-click on the number for a row and select "Delete") or just copy the table and paste it in a new sheet.

The last step is to save this workbook as a new sheet. It's also good to note somewhere when the information was downloaded and what link it came from.

Now you can analyze your data from the Web!

Unfortunately, this method won't work all of the time. Sometimes data online aren't in static tables or won't cleanly import into Excel. In those situations you might want to explore Web scraping. IRE members can search the IRE tipsheet database and the NICAR list archives for information on Web scraping. And remember, you can always submit an open records request to get the information in a usable format.

CONVERTING PDF

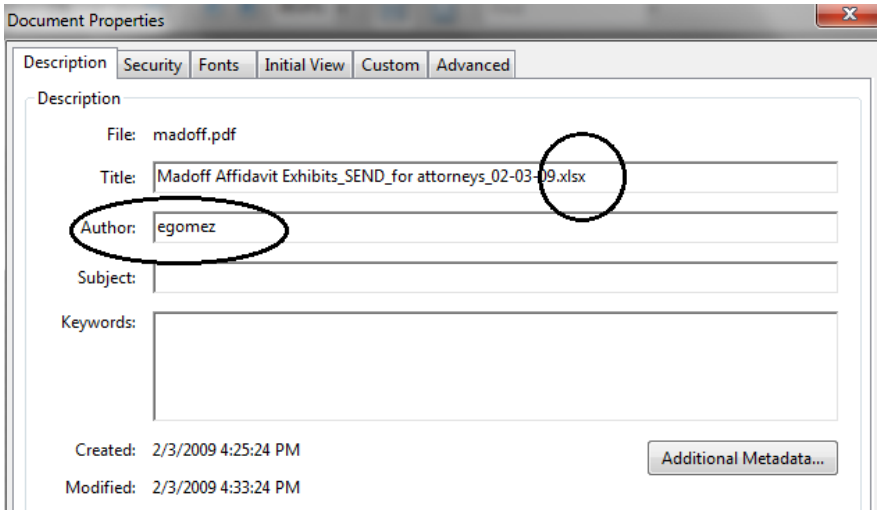
Sigh. This portion of the lesson didn't use to be necessary. Unfortunately, there's a growing trend of government entities releasing data in pdfs. When it's in this format, we can't sort the data, analyze it with PivotTables or use any of the fancy formulas we just learned! Don't worry, there are ways to handle this issue and get PDF files into Excel.

So there are two types of pdfs. Ones that are scanned. (boo!). Others that are saved electronically as a pdf. (much better!). Both are a pain so it might be worth arguing for the actual data. Note: Make sure in your initial records request to ask for .txt or .csv data and feel free to say "not a pdf."

If you are one of the lucky ones who have to deal with the pdf, you are in good company. Let's help you extract that data out!

Open Madoff.pdf. This is a file that the feds gave reporters of Bernie Madoff's clients. You will notice it's 163 pages of a spreadsheet!

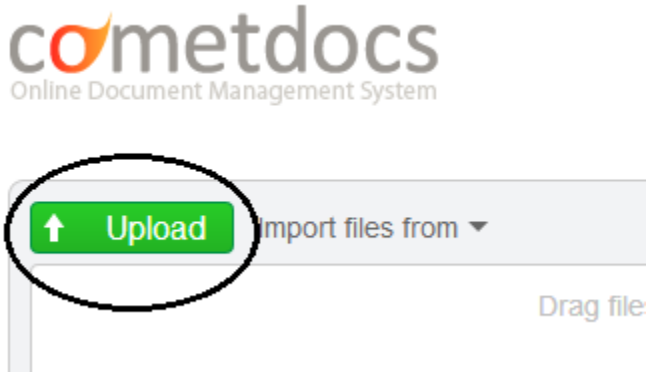
Once you get the pdf, open it and check the "Document Properties" under the "File" menu. This will sometimes show you who created the pdf: "egomez" and the original file type: Excel. So you can go back and say "Tell Gomez, I would like that Excel file he/she was working from."



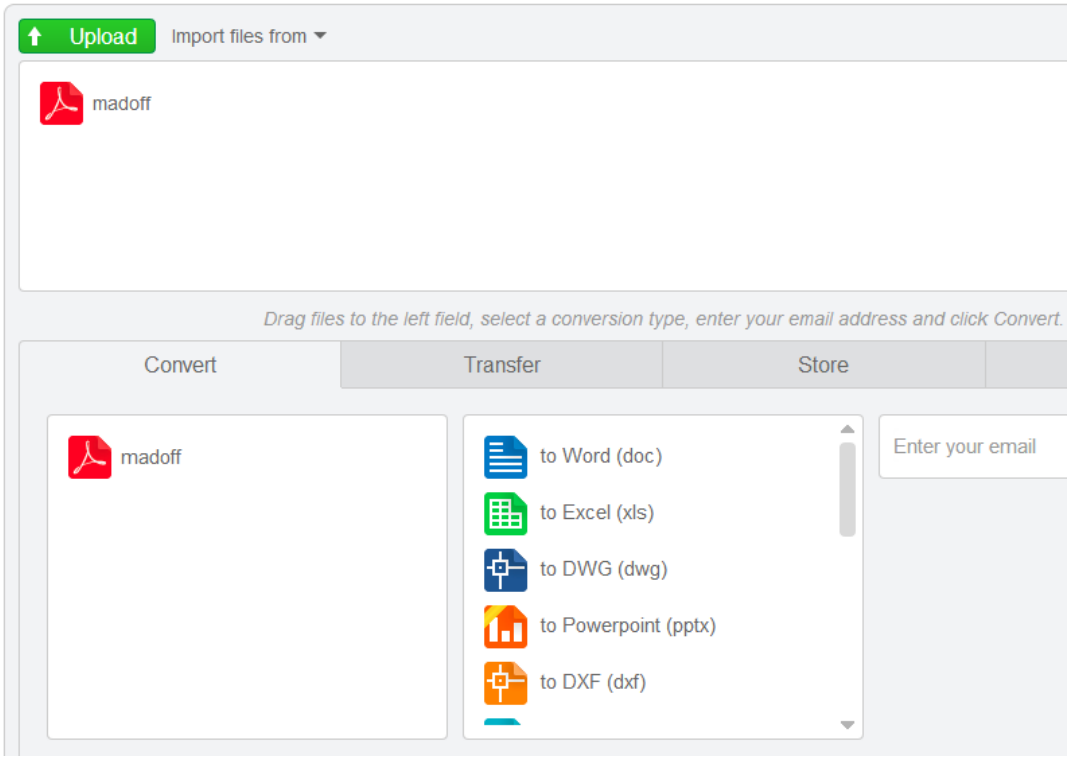
If that doesn't work to pry data from the government's hand, we can convert this pdf to an Excel file. There are a couple of options. One is cometdocs.com, which will be the one we will be using. More on that later. Two other options are Zamzar.com and Tabula. Zamzar is a free online converter site and similar to CometDocs. Tabula is another free option that was created by journalists. Tabula runs on your own computer and is used to extract data tables from PDF files. Find more information here: http://tabula.nerdpower.org/

Cometdocs is a website that allows you to convert files and store them. It's free, to a point. Go to cometdocs.com. Click on the large, blank box, to get the "Upload" button to show.

Click on the "Upload" button and navigate to the Madoff.pdf file. You can also drag the file to the white box.



Once the Madoff file is loaded, grab it and drag it down to where it says “Convert”. Then click on file type. There are several options including .txt. In this case, we will choose Excel. Plug in your email address and click on “Convert.”



In a few minutes you should get an email from Cometdocs with the subject “Your file is ready for download.” Click on the link in the mail and it will download your Excel file.

If you open the file, you will notice that it will need to be cleaned (this is common when you convert PDF files). For example, the page breaks in the pdf will create extra blank rows in the Excel spreadsheet.

Cometdocs only keeps that file for 24 hours. So if you don’t download it within 24 hours of receiving the email, you will need to convert again.

Since it’s a web application, we wouldn’t recommend using this service to convert private files.

For scanned pdfs, Cometdocs requires you be a premium member. IRE members are able to get a premium membership for free. For more information, go here: http://ire.org/blog/ire-news/2013/05/22/ire-announces-partnership-cometdocs/

###