FROM HTML TO XLS – TWO QUICK WAYS TO CAPTURE DATA

Robert Gebeloff

[rgebeloff@nytimes.com](mailto:rgebeloff@nytimes.com)

These days, the Web is filled with “open” data – information provided by government agencies in ready-to-use form. You’ll often encounter a data table on an agency Web site that has a download feature, where you can usually get the data you’re looking at in Excel format or as a text file that is easily importable into Excel.

But sometimes the data you want does not come with any download options. For these cases, let’s look at two simple methods you can use to grab the information for analysis.

WEB QUERIES:

Web queries are a feature in Excel that allow you to embed an HTML table from the Web inside a spreadsheet. Better yet, the data updates every time you close and open your file.

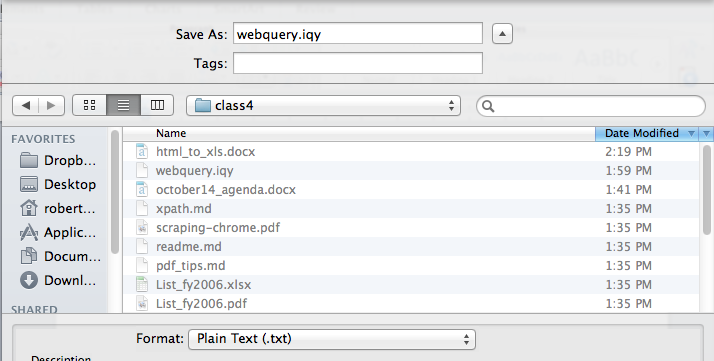
Two notes, I’m going to show you the process for a Mac. The process for a PC is slightly different – you can see it [here](http://www.techrepublic.com/article/pull-data-into-microsoft-excel-with-web-queries/).

Let’s say you wanted to track the stock prices of a few companies on your beat. One option would be to create a spreadsheet and embed a stock price feed from an online provider such as Yahoo!.

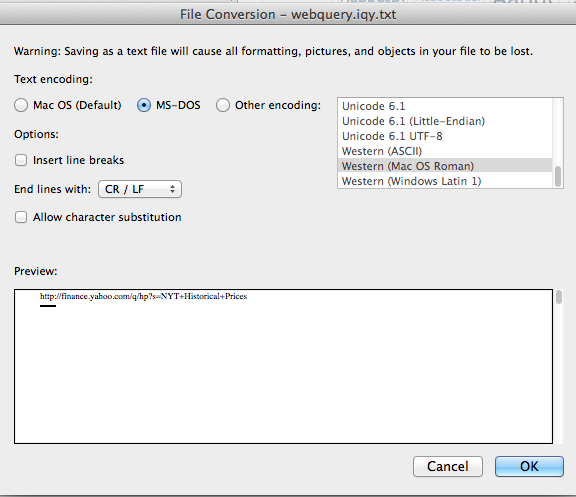
To start with, let’s go to Yahoo! and create the feed we want. Yahoo! finance is at <http://finance.yahoo.com>. Click on quote lookup, and then enter the symbol of our favorite media company, NYT. Then click on “Historical Prices”.

In Windows Excel, we’d be able to plug this URL into a Web query tool that is built into the software (Data/New Web Query), but on a Mac, we need to proceed down a different path.

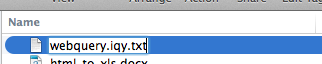
Copy the url, open up Microsoft Word, and paste the url into the first line. Then save the file as plain text. THE KEY: When saving, change the .txt extension to .iqy.

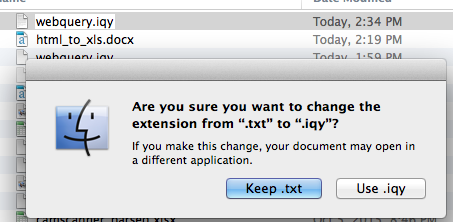


A dialog box appears – make sure you change the encoding to MS-DOS or your web query will not work.



One final step in preparing the query file – Word will want to append the .txt back onto your file. You’ll need to go to Finder and get rid of it.





Ok, now let’s get our data into Excel. Open up a new spreadsheet file, click on Data/Get External Data/Run Saved Query. Select the Web query file you just created. The data will load into your worksheet. Note that if you click on properties, there are various options for this move which you can explore on your own.

HTML CAPTURING PLUGINS

Web queries are great if you have data with a distinct URL and if you want to be able to automatically update the data. But sometimes you’re working with data generated by an online system that does not provide you with a distinct URL. In this case, it’s often easiest to turn to a browser plugin that is designed to capture the data.

The problem for us with data tables on Web pages is that they don’t copy and paste cleanly. When you copy from an HTML document, you’re not just copying what you see, but all of the underlying HTML code, and that can confuse Excel.

Let’s consider the case of the U.S. Federal Reserve’s Survey of Consumer Finance. Go to:

<http://www.federalreserve.gov/econresdata/scf/scfindex.htm>

The SCF is conducted every three years with the goal of capturing detailed information about the finances of U.S. households. The survey has a much smaller sample than the U.S. Census but is much more detailed, with more of a focus on long-term assets (such as stock and real estate holdings) than on annual income. This gives reporters a chance to analyze “wealth” – a different concept then the more common reporting on income.

The easiest way to access the data is through the SDA online query tool, a tool developed at UC-Berkeley that you’ll encounter from time to time when using survey data online. (Two other examples – the General Social Survey hosted by the University of Chicago and various Census microdata sets hosted by the Minnesota Population Center).

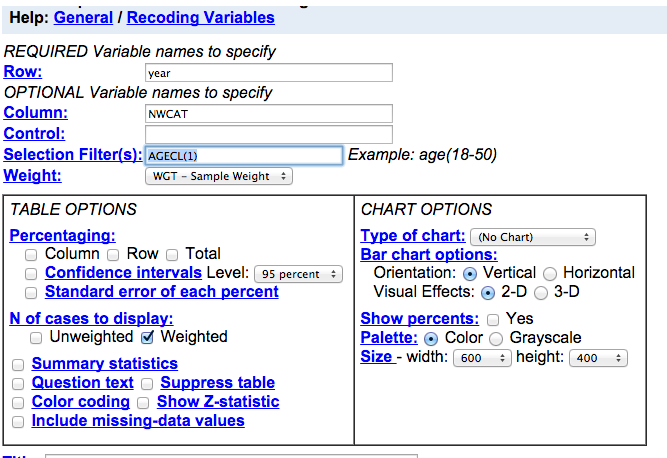
Before we begin, let’s make sure we have our browser plugin ready to grab the data we’re about to create.

In Chrome (we’ll go over a Firefox alternative in a minute), go to chrome://extensions/ and search for an extension called “Table Capture” and add it to Chrome.

Going back to the SCF page, look for a link to the SDA analysis tool about halfway down the page and click on it.

This just opened up a whole world of fun, which you can explore on your own, but let’s for this exercise run my sample query. Let’s look at the distribution of households headed by somebody under 35 by Net Worth percentile over time and see if the share of young people on top of the economic ladder has changed much.

1. In the form field marked Row, enter the variable YEAR
2. In the field marked Column, enter the variable NWCAT, which stands for “net worth category”.
3. In the field marked Filter, enter AGECL(1), which means only include households where the head is in age group 1 (under 35).
4. Uncheck every box in Table Options and set the chart options to “no chart”



Run the query. The result is a data table that we’d like to analyze, but because there is no unique URL, we can’t save it as a Web query. Let’s use our plug in.

Click on the red Table Capture icon in your URL bar:



A menu appears that describes all of the HTML tables detected on the page. If you hover over the choices, it shows you its intentions. Let’s click on Table 3, 13x1, and copy that to your clipboard. Now paste it into a new Excel worksheet.

Spend a few minutes analyzing the data on your own and we’ll discuss your findings.

One last note: In Firefox, there’s a plugin called Table2Clipboard that I find works more consistently than Table Capture, and I sometimes use Firefox as my browser when I know I’m going to be working with an online query tool like this – YMMV.