Data-Scraping Government Statistics (for the Coding-Challenged)

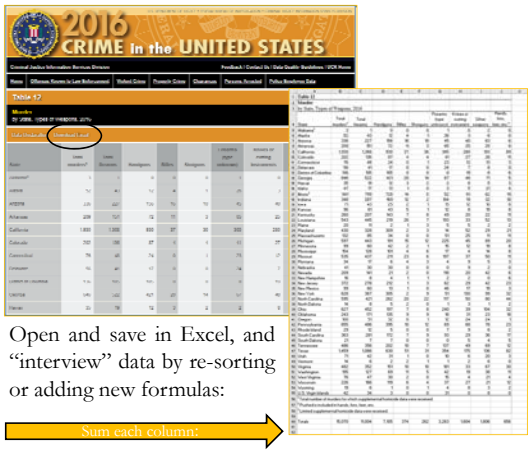
Carl P. Olson, Towson University
colson@towson.edu

What Is Data Scraping?

Data-scraping transfers tabular data from online documents to a spreadsheet format such as Excel or CSV, rather than key in data by hand or harvest a whole site.

What Is The Easiest Way To Scrape Data?

Many government sites offer tables in Excel or CSV. BLS, FRED, and NIH are just a few agencies with multiple data versions.



Open and save in Excel, and "interview" data by re-sorting or adding new formulas:

Sum each column:

Many tables still have no downloads. What then?

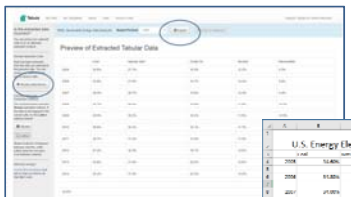
Dozens of free software tools can scrape data. Choose the software that is the safest, most reliable, has the shortest learning curve, and best fits your workflow.

Fancy Data Scraping with free tools

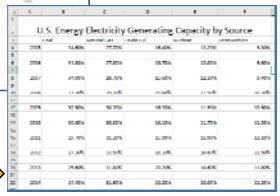
One example is Tabula, a free tool used by journalists for scraping PDF files. Tabula opens in the user's web browser. Its main limitation is that it scrapes only one table at a time. It can be confused by rows that cross a range of cells.

Upload a PDF file to Tabula.

Drag the cursor over the data cells desired.



Use *Revise Selections* to adjust. Export to an Excel table.



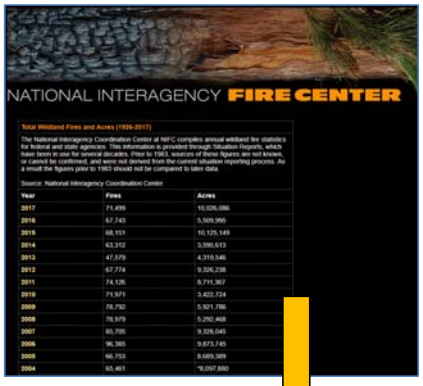
Use the edit functions to add headings or other analysis.

Make it pretty!

Data Scraping with Google Sheets

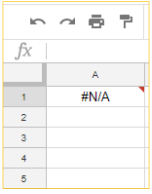
Below is a cool method for scraping tables from an HTML table to Google Sheets, from ProPublica's Lena Groeger (@lenagroeger). It works well with directories or web tables. Database search results or dynamically generated pages do not respond to this formula:

1. Find a table online presented in simple HTML;
2. (ex., https://www.nifc.gov/fireInfo/fireInfo_stats_totalFires.html)
3. Open Google Sheets; Compose the following formula:
4. =IMPORTHTML ("URL","type of element",first element in table)
5. Where: URL=data source, type of element=table, 0 starts the table)
6. Add this formula in cell A1:
7. =IMPORTHTML("[url above]","table",0)
8. The table should populate Google Sheets.

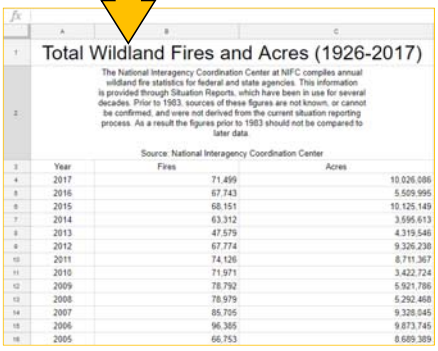


Proof the formula as needed. If you still get an error, flash page tables will require other options.

Use Google Sheets Export functions to create an Excel or PDF file.



#N/A is the error message.



Further Information on Data Scraping for Mortals

1. ProPublica Intro to Data & Code (https://bit.ly/1Kn6Eav);
2. Dan Nguyuen's Intro to Data-Scraping: (https://bit.ly/2Ig5khS);
3. Tabula Data Scraping for PDF (https://tabula.technology/);
4. ParseHub Data Scraping for Web Pages (https://www.parsehub.com/);
4. Web Scraper Chrome Extension (http://webscraper.io/);
5. Outwit Hub Firefox Extension (https://www.outwit.com/);
6. Import.io (https://www.import.io/);

TOWSON UNIVERSITY. | Poster printing provided by the Office of Technology Services | towson.edu/ots