Data Scraping Exercise

Contact: mikereilley1@gmail.com | @journtoolbox

bit.ly/spjdatascrape

ProPublica's Lena Groeger (@lenagroeger) shared this cool trick for importing data A works for both HTML tables and long lists. Huge timesaver. Just cut and paste the formula, URL and table/list number in the first cell on the Google Sheet and hit return.

Data is moved into the sheet. Start by opening up a new spreadsheet in Google Drive, then build the script using the instructions below.

Google News Initiative Training Center: https://newsinitiative.withgoogle.com/training/

Journalist's Toolbox (scraping and data viz tools):
https://www.journaliststoolbox.org/

Journalist's Toolbox e-newsletter: https://journaliststoolbox.substack.com/

Here's a video how to from Google and the Knight Massive Open Online Course 2019: https://www.youtube.com/watch?time_continue=10&v=Mu8yLpJsqDs

Video: Reilley demos scraping with =importhtml and Tabula (.PDFs):
https://www.youtube.com/watch?v=vIKwsQPvdHA&feature=youtu.be

***--Searching for datasets? Use Google Dataset Search:**
https://datasetsearch.research.google.com/

Visualize data without coding on Google Public Data Explorer:
https://www.google.com/publicdata/directory

***-More data scraping in Mike and Samantha Sunne's book, "Data + Journalism" buy it:** http://bit.ly/buydatabook | https://dataplusjournalism.com/

VisualPing.io: https://visualping.io/

—------

**** DO NOT TYPE ON STEPS 1-3 JUST COPY NUMBER 5 ****

1.Start with this formula:

=IMPORTHTML("URL","table", 0)

2. Add this URL to the formula. This is the page we'll scrape: Country by Country COVID-19 vaccinations

=IMPORTHTML("https://www.fdic.gov/resources/resolutions/bank-failures/failed-bank-list/",“ELEMENT HERE”, NUMBER OF ELEMENT ON PAGE HERE)

3. Add this type of element to the formula: table and 0

=IMPORTHTML("https://www.fdic.gov/resources/resolutions/bank-failures/failed-bank-list/","table",0)

4. **Code you enter into cell A1 on the Goog**

le Sheet. Just COPY and paste this into a Google Sheet:

=IMPORTHTML("https://www.fdic.gov/resources/resolutions/bank-failures/failed-bank-list/","table",0)

****-Note: Depending on your Google Sheets language settings, the delimiter in the function could be "," or ";" It's usually a comma**

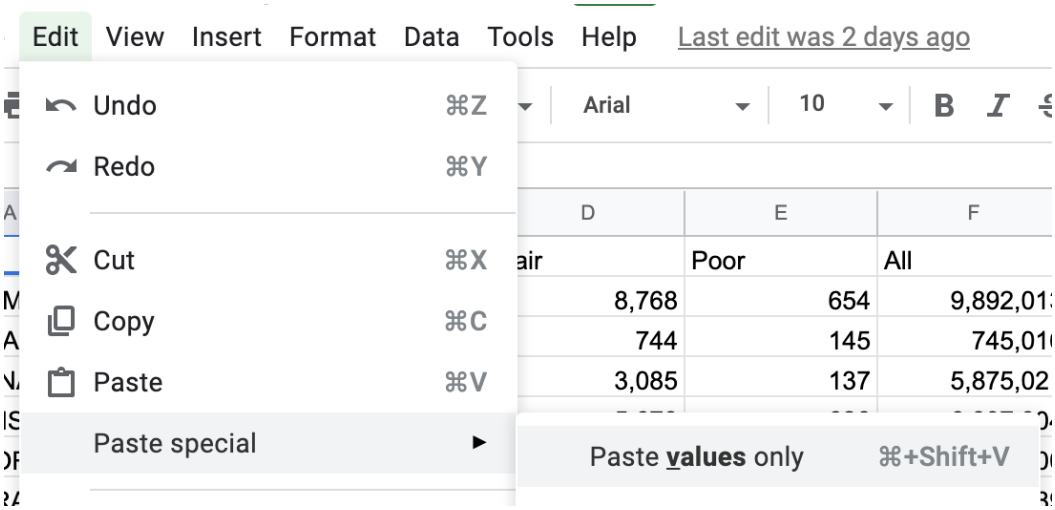
5. One problem: If you go to type on the screen after you scrape, your data disappears. This is because the spreadsheet is linked to the web page. When the web page updates, so does the sheet.

=IMPORTHTML("https://www.fdic.gov/resources/resolutions/bank-failures/failed-bank-list/","table",0)

6. To remedy the issue, highlight all the data on the screen and copy it. Then click on the Plus Sign in the lower left corner of the Google Sheet to open a new tab in the spreadsheet. Click on the “Sheet 2” language on the tab to rename it “EDITS” have himor whatever you want.

> **Pro tip:** You always, always, always work off a COPY of your dataset in case you mess something up.

7. To paste the data into the new sheet, click into cell A1, then go to the **Edit > Paste Special > Paste values only** menu at the top and click **(See figure 1.1 below).** Presto! Your data appears in the new sheet and is now editable. You still have your original scraped sheet in the first tab.



Exercise on your own: Here are some more pages you can practice scraping, using this formula for webpages. Just plug ONE of the web address into the area that says URL in one of the formulas below, copy and paste the entire formula into cell A1 on a Google Sheet and hit return:

https://www.nifc.gov/fire-information/statistics/wildfires

https://www.fhwa.dot.gov/bridge/nbi/no10/condition19.cfm

https://ourworldindata.org/covid-vaccinations#source-information-country-by-country

https://www.bls.gov/news.release/cpi.t02.htm

https://www.fdic.gov/resources/resolutions/bank-failures/failed-bank-list/

https://www.prisonpolicy.org/origin/ny/zipcodes.html

https://www.prisonpolicy.org/origin/ny/2020/zipcode.html

https://www.prisonpolicy.org/origin/ny/2020/tract.html

https://www.inmo.ie/Trolley_Ward_Watch

https://www.transtats.bts.gov/AverageFare/

https://www.bls.gov/news.release/cpi.t02.htm

https://www.geostat.ge/en/modules/categories/195/business-statistics

https://www.census.gov/quickfacts/MA

https://www.census.gov/quickfacts/fact/table/orangecitycityflorida,ormondbeachcityflorida,daytonabeachshorescityflorida,daytonabeachcityflorida,asburylakecdpflorida,volusiacountyflorida/PST045221

https://www.census.gov/quickfacts/fact/table/orangecitycityflorida,ormondbeachcityflorida,daytonabeachshorescityflorida,daytonabeachcityflorida,asburylakecdpflorida,volusiacountyflorida/PST045221

http://www.dllr.state.md.us/employment/warn.shtml

http://www.espn.com/mlb/attendance/_/year/2019

http://www.dllr.state.md.us/employment/warn.shtml

https://en.wikipedia.org/wiki/List_of_Formula_One_Grand_Prix_winners#By_driver
(select table 7 and verify data as this comes from a Wikipedia page)

https://www.cision.com/2015/01/top-50-rich-media-social-influencers-to-follow-on-twitter/

https://artbabridgereport.org/state/profile/MI
(Change table number to pull different locations)

If you want to scrape multiple tables on a page ...

** **Note:** Be sure to fact-check any data off a Wikipedia page

Washington D.C. COVID testing sites: https://coronavirus.dc.gov/testing (Select table 1, 2, 3 in the number of element on the page)
E.g.: =IMPORTHTML("https://coronavirus.dc.gov/testing","table",2)

You can drop these into Google MyMaps to create a map, then export as a KML file and import into Google Earth Studio to create a small tour. Very cool.

New Google Sheet: sheets.new

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

=IMPORTHTML("URL","table",0)

—----

READ MORE

More info. on other things you can do with sheets:

https://www.sheetgo.com/importhtml-formula-google-sheets/

https://support.google.com/docs/answer/3093339?hl=en

· Can be a table or a list

· Can also use IMPORTXML, IMPORTDATA, IMPORTFEED ----------for web data:

· IMPORTRANGE: Imports a range of cells from a specified spreadsheet.

· IMPORTFEED: Imports a RSS or ATOM feed.

· IMPORTDATA: Imports data at a given url in .csv (comma-separated value) or .tsv (tab-separated value) format.

· support.google.com/docs/search?q=import

———————

Google Chrome Scraper Extension

· In Chrome, visit the URL you want to scrape

· Install extension: http://mnmldave.github.io/scraper/

http://mnmldave.github.io/scraper/

Then try scraping this table:

http://www.nifc.gov/fireInfo/fireInfo_stats_totalFires.html

https://egov.uscis.gov/processing-times/historic-pt

Steps: > Right-click and choose 'Scrape similar…' to scrape the relevant content of the page > Inspect content to determine relevance > Adjust copied text if not relevant

· Save it in Google Docs!

Other pages to try the scraper extension in:

Google Chrome Store scrapers to try: Web Scraper, Scrapely, Scrape.it, Agently, ocscraper, Data Scraper.

Download Web Scraper in the Google Chrome Store: **Video how-to:** https://www.youtube.com/watch?time_continue=4&v=G74Xg1-QpD4

** It's good for catching embedded tables with beveled edges, scrollbars and tags in them. Takes some time to learn it, though.

Scrape PDFS

Tabula: http://tabula.technology/

Free tool. Download to computer's hard drive. Scrape tables out of .PDFs! Great security and privacy with this tool.

*-**Training video:** https://www.youtube.com/watch?v=eVQ93FTtph0

PDFtoExcel.com: https://www.pdftoexcel.com/
Scrape native and scanned .PDFs. Be careful with security as you are loading your document to a live website that could get hacked.

TinyWow: TinyWow
Dozens of free file conversion tools, including PDF to Excel, converting spreadsheets, video, etc.

***-Training video:** https://www.youtube.com/watch?v=AvdpqJaMCQ8

***-Folder of PDFs to Scrape**

CometDocs
https://www.cometdocs.com/

QuickCode, SensibleCode and .PDF table scraper tools. Python library:
https://github.com/scraperwiki/scraperwiki-python

Abbyy Fine Reader OCR .PDF Scraper
https://www.abbyy.com/en-us/finereader/
Top-end OCR .pdf converter

OnlineOCR .PDF Scraper
https://www.onlineocr.net/
Good for smaller files.

Import.io: https://www.import.io/

Costly but offers some great data extraction tools for journalists and academics.

Google Keep: https://keep.google.com/

Google's note-taking app lets you export text out of an image. Just click on "Grab Image Text" and pops it into the text of your note. Hat tip to Samantha Sunne from Tools for Reporters on this tip.

XML to CSV file (scrape off an XML web page): XML Grid óplhttp://xmlgrid.net/xml2text.html

You can also google XML to CSV converter and find many other free tools to do this.

BeautifulSoup: https://www.crummy.com/software/BeautifulSoup/

Parse data from websites.

Open Refine: http://openrefine.org/

Clean up dirty data.

10 Best Web Scraping Tools: https://www.hongkiat.com/blog/web-scraping-tools/

Social media scraping: https://www.journaliststoolbox.org/category/twitter-resources/

I don't do a lot of this, but you can find some resources on The Journalist's Toolbox Twitter/social media page

Conversations with Data Newsletter: Scraping Tools

http://r.mediapusher.eu/mk/mr/s6Ww5rsbAaT4IN2kJCLr0YgkMl64ilhRorv4Y1xmSZ1IaXSG3ezdi4bVdA1-fGxBv5-fkGDQ_axbP7nMkT5rpiniz5QUrT3aZzO0xy0

A great collection of data scraping tools, tips and tricks.

Read about the Ethics of Web Scraping:
https://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/

SEARCHING FOR DATASETS

*** Here are some tools Mike showed you how to use**

Video: Find Data with Advanced Google Search (filetype: operator)

https://www.youtube.com/RXJTYwatch?time_continue=3&v=Axf0vZ

Google Dataset Search (aggregator): https://toolbox.google.com/datasetsearch

Video how-to with Dataset Search: https://www.youtube.com/watch?v=lCXoyZprPn8

Try these in Google.com:

Cut and paste into the search field and hit return: filetype:csv u.s. mass shootings

* Finds the raw data files on topic across the web

Cut and paste into the search field and hit return: filetype:xlsx site:worldbank.org education

* Finds the raw data files within that website

** To find a list of files you can search on, go to Advanced Search in the lower right corner of Google.com and find the filetype pulldown menu.

Google Public Data Explorer: https://www.google.com/publicdata/directory

Video how-to: https://www.youtube.com/watch?time_continue=6&v=kplu466Y8rg

* Build and embed animated charts from datasets provided in software. No coding necessary. Great for census stories.