Websites, Twitter & Facebook - oh my!

How to start gathering data from the web

Michael Beckstrand, Ph.D.

Mixed Methods Research Associate

Alicia Hofelich Mohr, Ph.D.

Research Data Manager

Lots of data “out there” to harvest

How can you access it?

How can efficiently download and handle data once accessed?

How can you get downloaded data into a usable format?

Lots of tools "out there" to help you

What to ask yourself to choose:

1. What do you want to do?
2. How much time/effort are you willing to put into learning a tool?
3. How extensible do you need the tool to be?

What do you want to do?

Three major categories:

- Save what you see on a page
- Capture what you see on the screen in a structured format
- Extract specific sets of information from a website

How easy will it be to learn?

Continuum from click-based to fully coded

How extensible?

Can you rely on the listed, built-in feature set or do you need customizability to very specific scenarios

Tool	Use	Ease to Learn	Extensibility
Get Them All	Save what you See	Easy	Limited
HTTrack		Easy	Limited
If this then that		Medium	Some
Scraper	Structure what you see	Medium	Some
Twitter Archiving Google Sheet	Query APIs	Easy	Limited
FacePager		Medium	Some
Python	Structure what you see & query APIs	Difficult	High
R		Difficult	High

Saving what you see

Download a bunch of stuff - ‘batch’ the operations to:

- Get all files linked on a page
- Mirror whole (or sections) of sites
- Create a workflow of tools to store/organize articles

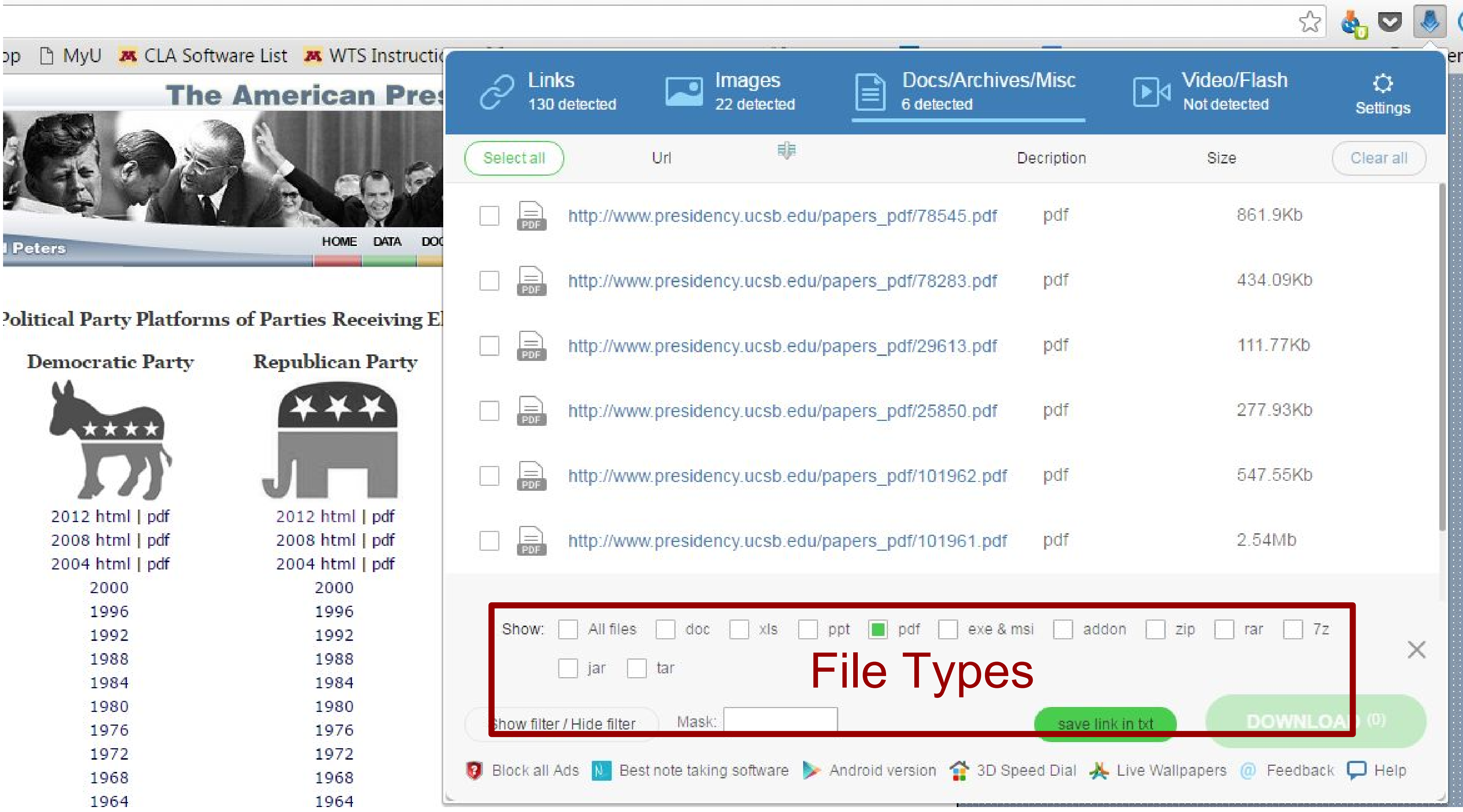
Example Tools for Batching the Web

GetThemAll
Chrome extension for downloading multiple links on a single page all at once ('Chrono Download Manager' also an excellent choice)

HTTrack
Stand-alone crawler that can mirror a whole (or part of) website onto your local computer

IFTTT
Recipe
if this then that
Trigger
Action

IFTTT
'Recipes' for automating actions between web services, like cloud storage, social media, article managers/RSS feeds



op MyU CLA Software List WTS Instructi

The American Pres

Peters

HOME DATA DOC

Political Party Platforms of Parties Receiving El

Democratic Party	Republican Party
2012 html | pdf	2012 html | pdf
2008 html | pdf	2008 html | pdf
2004 html | pdf	2004 html | pdf
2000	2000
1996	1996
1992	1992
1988	1988
1984	1984
1980	1980
1976	1976
1972	1972
1968	1968
1964	1964

Links 130 detected · Images 22 detected · Docs/Archives/Misc 6 detected · Video/Flash Not detected · Settings

Select all · Url · Decription · Size · Clear all

	Url	Decription	Size
□	http://www.presidency.ucsb.edu/papers_pdf/78545.pdf	pdf	861.9Kb
□	http://www.presidency.ucsb.edu/papers_pdf/78283.pdf	pdf	434.09Kb
□	http://www.presidency.ucsb.edu/papers_pdf/29613.pdf	pdf	111.77Kb
□	http://www.presidency.ucsb.edu/papers_pdf/25850.pdf	pdf	277.93Kb
□	http://www.presidency.ucsb.edu/papers_pdf/101962.pdf	pdf	547.55Kb
□	http://www.presidency.ucsb.edu/papers_pdf/101961.pdf	pdf	2.54Mb

Show: □ All files □ doc □ xls □ ppt ☑ pdf □ exe & msi □ addon □ zip □ rar □ 7z
□ jar □ tar

File Types

Show filter / Hide filter · Mask: · save link in txt · DOWNLOAD (0)

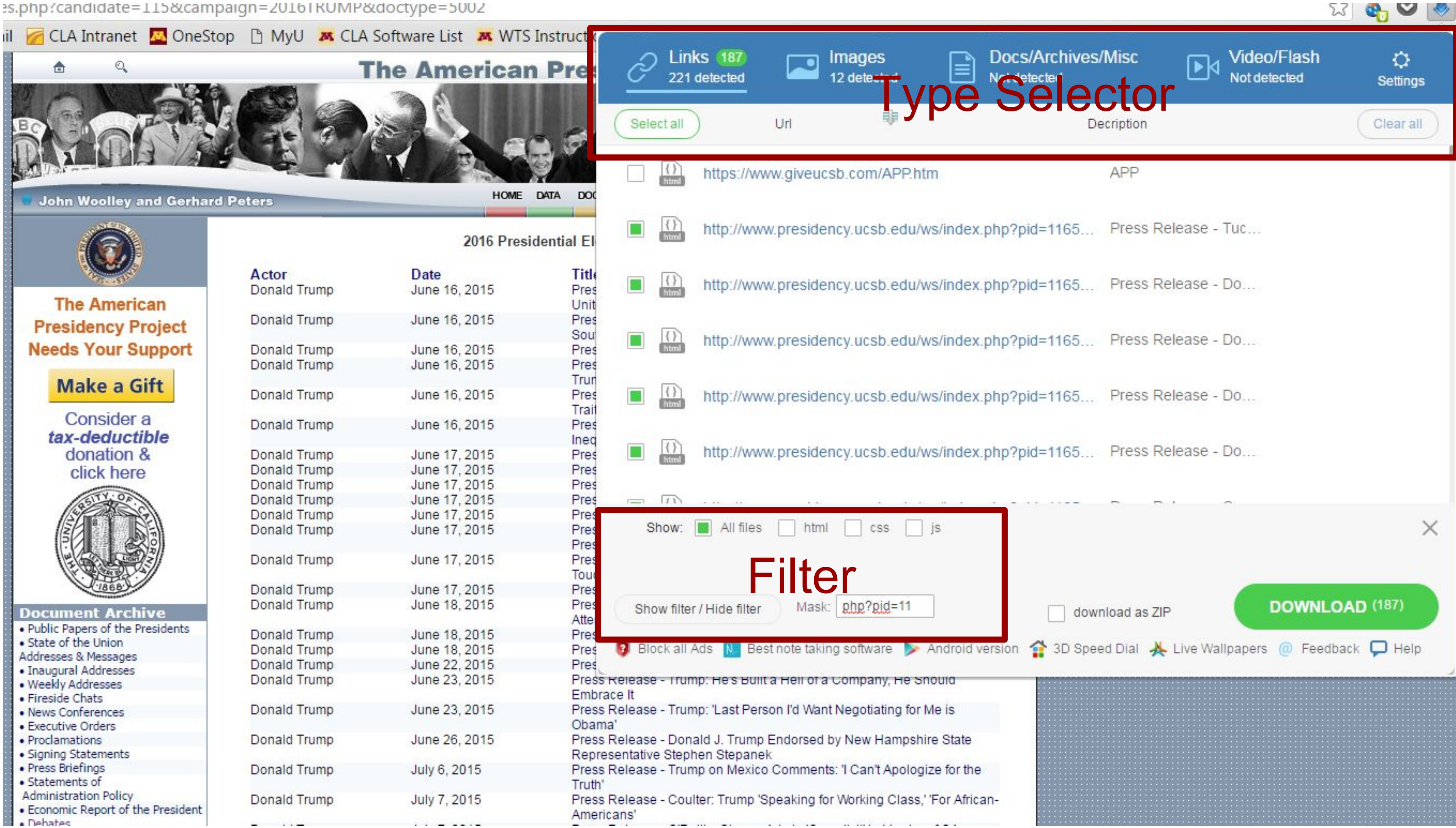
Block all Ads · Best note taking software · Android version · 3D Speed Dial · Live Wallpapers · Feedback · Help

Turning what you see (and don’t see) into structured data

‘Web scraping’ relies on underlying HTML

This hierarchical/marked-up data can be converted to spreadsheets and tables

Big caveat: ‘scraping’ only grabs what is on screen or has been loaded by the browser



Mask: php?pid=11

	Url	Decription
□	https://www.giveucsb.com/APP.htm	APP
☑	http://www.presidency.ucsb.edu/ws/index.php?pid=1165...	Press Release - Tuc...
☑	http://www.presidency.ucsb.edu/ws/index.php?pid=1165...	Press Release - Do...
☑	http://www.presidency.ucsb.edu/ws/index.php?pid=1165...	Press Release - Do...
☑	http://www.presidency.ucsb.edu/ws/index.php?pid=1165...	Press Release - Do...
☑	http://www.presidency.ucsb.edu/ws/index.php?pid=1165...	Press Release - Do...

The American Presidency Project Needs Your Support

Make a Gift

Consider a *tax-deductible* donation & click here

Document Archive
- Public Papers of the Presidents
- State of the Union Addresses & Messages
- Inaugural Addresses
- Weekly Addresses
- Fireside Chats
- News Conferences
- Executive Orders
- Proclamations
- Signing Statements
- Press Briefings
- Statements of Administration Policy
- Economic Report of the President
- Debates

2016 Presidential El

Actor	Date	Title
Donald Trump	June 16, 2015	Pres Unit
Donald Trump	June 16, 2015	Pres Sou
Donald Trump	June 16, 2015	Pres
Donald Trump	June 16, 2015	Pres Trun
Donald Trump	June 16, 2015	Pres Trai
Donald Trump	June 16, 2015	Pres Ineq
Donald Trump	June 17, 2015	Pres
Donald Trump	June 17, 2015	Pres
Donald Trump	June 17, 2015	Pres
Donald Trump	June 17, 2015	Pres
Donald Trump	June 17, 2015	Pres
Donald Trump	June 17, 2015	Pres Pres
Donald Trump	June 17, 2015	Pres Tou
Donald Trump	June 17, 2015	Pres
Donald Trump	June 18, 2015	Pres Atte
Donald Trump	June 18, 2015	Pres
Donald Trump	June 18, 2015	Pres
Donald Trump	June 22, 2015	Pres
Donald Trump	June 23, 2015	Press Release - Trump: He's Built a Hell of a Company, He Should Embrace It
Donald Trump	June 23, 2015	Press Release - Trump: 'Last Person I'd Want Negotiating for Me is Obama'
Donald Trump	June 26, 2015	Press Release - Donald J. Trump Endorsed by New Hampshire State Representative Stephen Stepanek
Donald Trump	July 6, 2015	Press Release - Trump on Mexico Comments: 'I Can't Apologize for the Truth'
Donald Trump	July 7, 2015	Press Release - Coulter: Trump 'Speaking for Working Class,' 'For African-Americans'

Example Tools for Scraping

Scraper

Chrome extension with point & click interface & regular expression-style queries

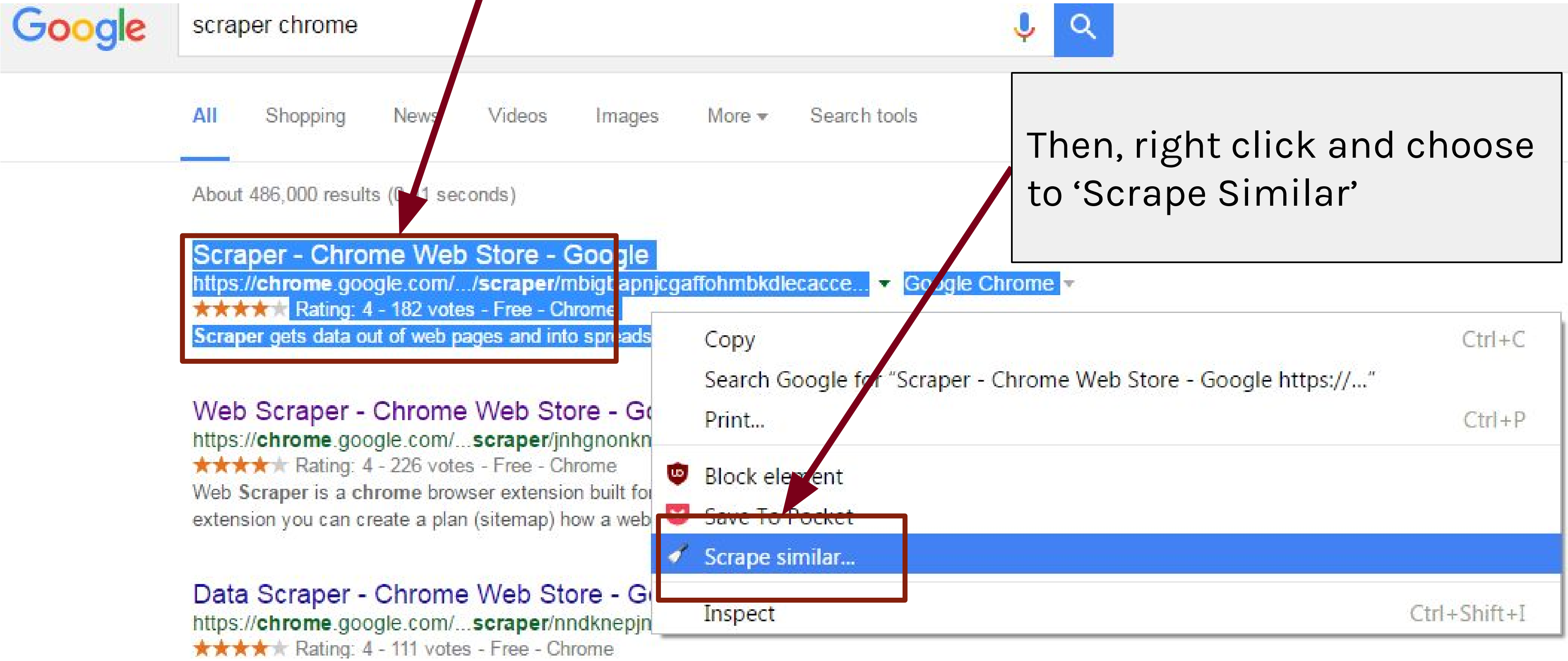
Python (among other coding solutions)

Programming language with libraries for extracting data from the web (Beautiful Soup)

R

Packages for scraping and harvesting webpages including RSelenium, Rvest, scrapeit

First, select an example of the content you want to scrape



scraper chrome

All Shopping News Videos Images More Search tools

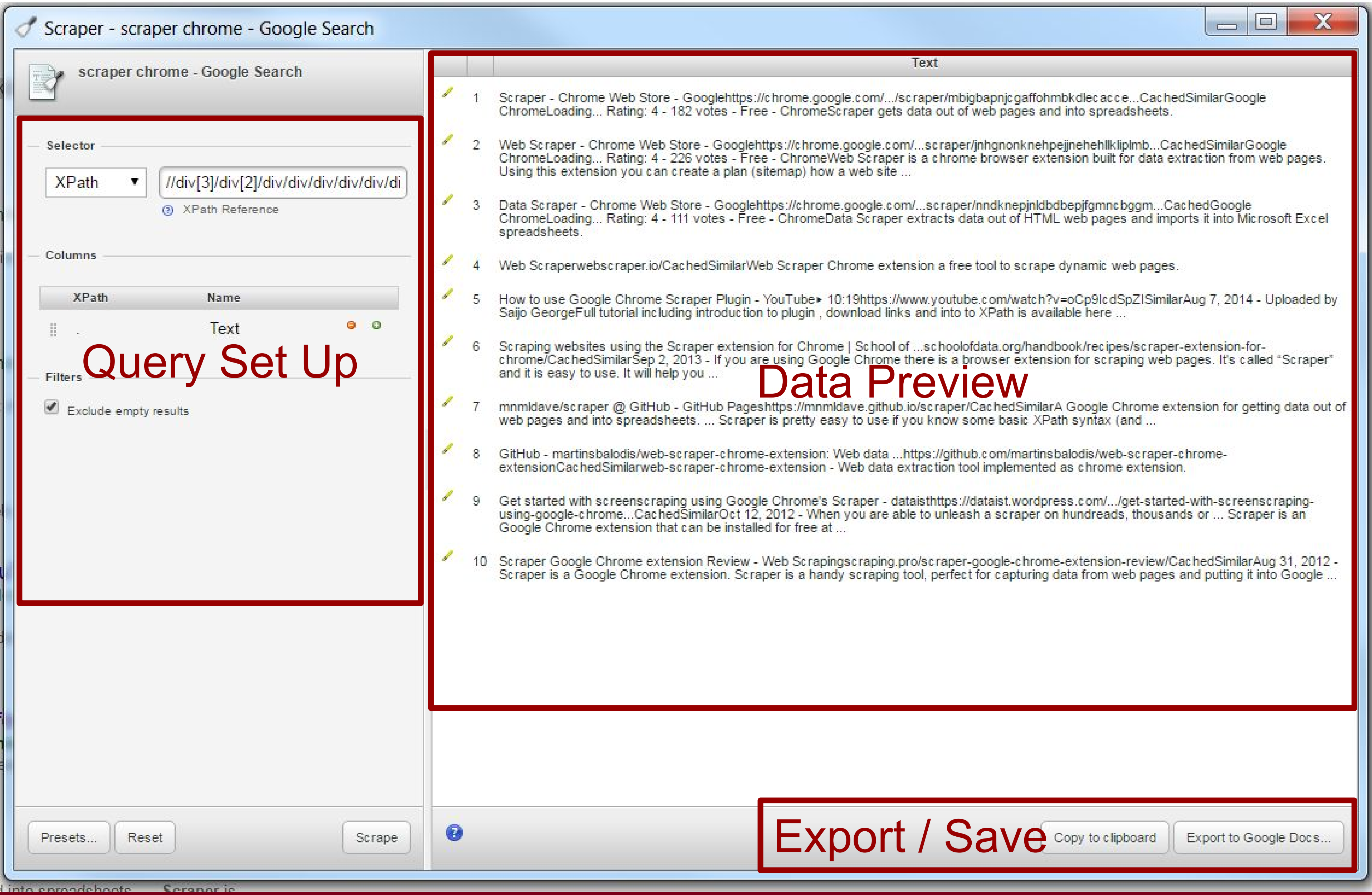
About 486,000 results (0.41 seconds)

Scraper - Chrome Web Store - Google
https://chrome.google.com/.../scraper/mbigbapnjcgaffohmbkdlecacce... Google Chrome
★★★★☆ Rating: 4 - 182 votes - Free - Chrome
Scraper gets data out of web pages and into spreads

Web Scraper - Chrome Web Store - G
https://chrome.google.com/...scraper/jnhgnonkn
★★★★☆ Rating: 4 - 226 votes - Free - Chrome
Web Scraper is a chrome browser extension built fo
extension you can create a plan (sitemap) how a web

Data Scraper - Chrome Web Store - G
https://chrome.google.com/...scraper/nndknepjn
★★★★☆ Rating: 4 - 111 votes - Free - Chrome

Then, right click and choose to 'Scrape Similar'



Query Set Up

Selector

XPath ▼ //div[3]/div[2]/div/div/div/div/div/di

Columns

XPath	Name
.	Text

Filters

☑ Exclude empty results

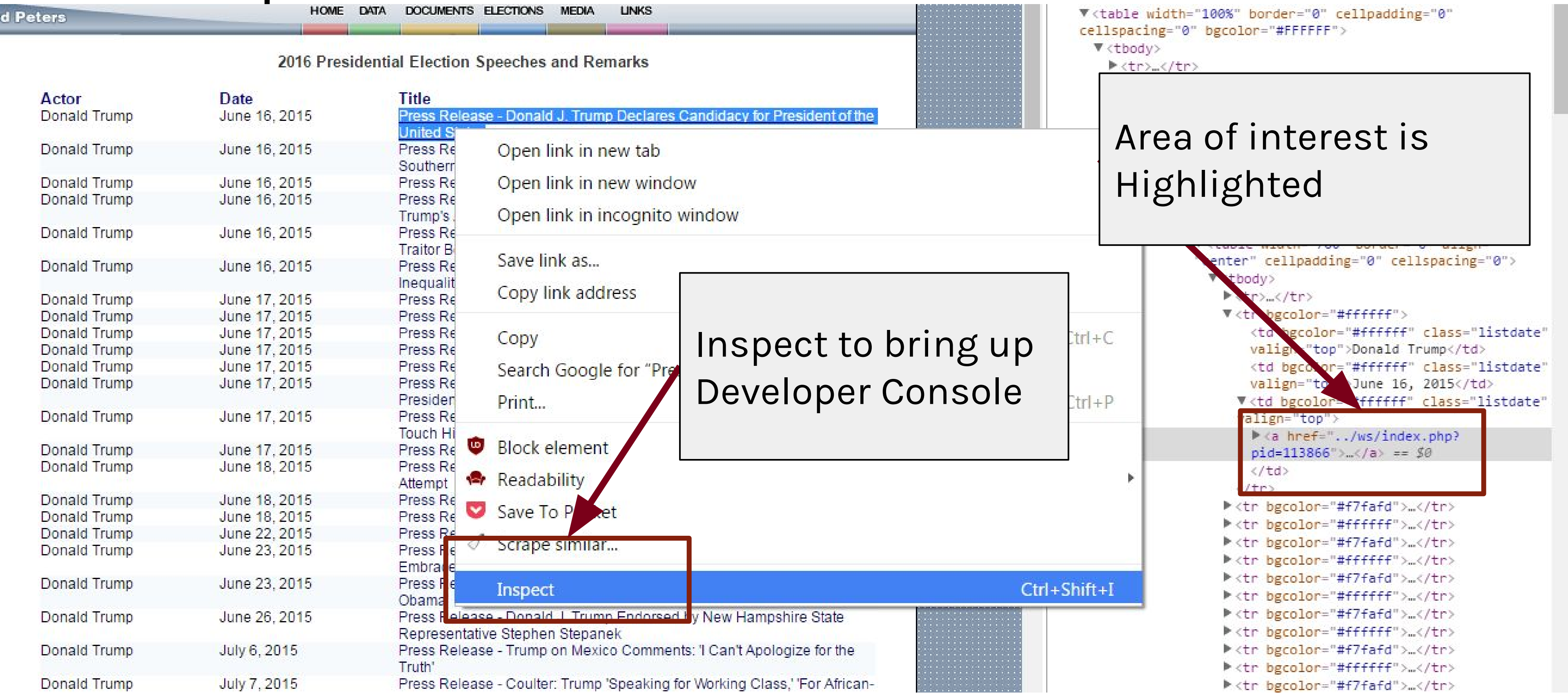
Data Preview

	Text
1	Scraper - Chrome Web Store - Googlehttps://chrome.google.com/.../scraper/mbigbapnjcgaffohmbkdlecacce...CachedSimilarGoogle ChromeLoading... Rating: 4 - 182 votes - Free - ChromeScraper gets data out of web pages and into spreadsheets.
2	Web Scraper - Chrome Web Store - Googlehttps://chrome.google.com/...scraper/jnhgnonknehpejjnehehllkliplmb...CachedSimilarGoogle ChromeLoading... Rating: 4 - 226 votes - Free - ChromeWeb Scraper is a chrome browser extension built for data extraction from web pages. Using this extension you can create a plan (sitemap) how a web site ...
3	Data Scraper - Chrome Web Store - Googlehttps://chrome.google.com/...scraper/nndknepjnldbdbepjfgmncbggm...CachedGoogle ChromeLoading... Rating: 4 - 111 votes - Free - ChromeData Scraper extracts data out of HTML web pages and imports it into Microsoft Excel spreadsheets.
4	Web Scraperwebscraper.io/CachedSimilarWeb Scraper Chrome extension a free tool to scrape dynamic web pages.
5	How to use Google Chrome Scraper Plugin - YouTube▸ 10:19https://www.youtube.com/watch?v=oCp9lcdSpZISimilarAug 7, 2014 - Uploaded by Saijo GeorgeFull tutorial including introduction to plugin , download links and into to XPath is available here ...
6	Scraping websites using the Scraper extension for Chrome | School of ...schoolofdata.org/handbook/recipes/scraper-extension-for-chrome/CachedSimilarSep 2, 2013 - If you are using Google Chrome there is a browser extension for scraping web pages. It's called "Scraper" and it is easy to use. It will help you ...
7	mnmldave/scraper @ GitHub - GitHub Pageshttps://mnmldave.github.io/scraper/CachedSimilarA Google Chrome extension for getting data out of web pages and into spreadsheets. ... Scraper is pretty easy to use if you know some basic XPath syntax (and ...
8	GitHub - martinsbalodis/web-scraper-chrome-extension: Web data ...https://github.com/martinsbalodis/web-scraper-chrome-extensionCachedSimilarweb-scraper-chrome-extension - Web data extraction tool implemented as chrome extension.
9	Get started with screenscraping using Google Chrome's Scraper - dataisthttps://dataist.wordpress.com/.../get-started-with-screenscraping-using-google-chrome...CachedSimilarOct 12, 2012 - When you are able to unleash a scraper on hundreads, thousands or ... Scraper is an Google Chrome extension that can be installed for free at ...
10	Scraper Google Chrome extension Review - Web Scrapingscraping.pro/scraper-google-chrome-extension-review/CachedSimilarAug 31, 2012 - Scraper is a Google Chrome extension. Scraper is a handy scraping tool, perfect for capturing data from web pages and putting it into Google ...

Export / Save

For more advanced selection

Select target area, right click, choose ‘Inspect’



2016 Presidential Election Speeches and Remarks

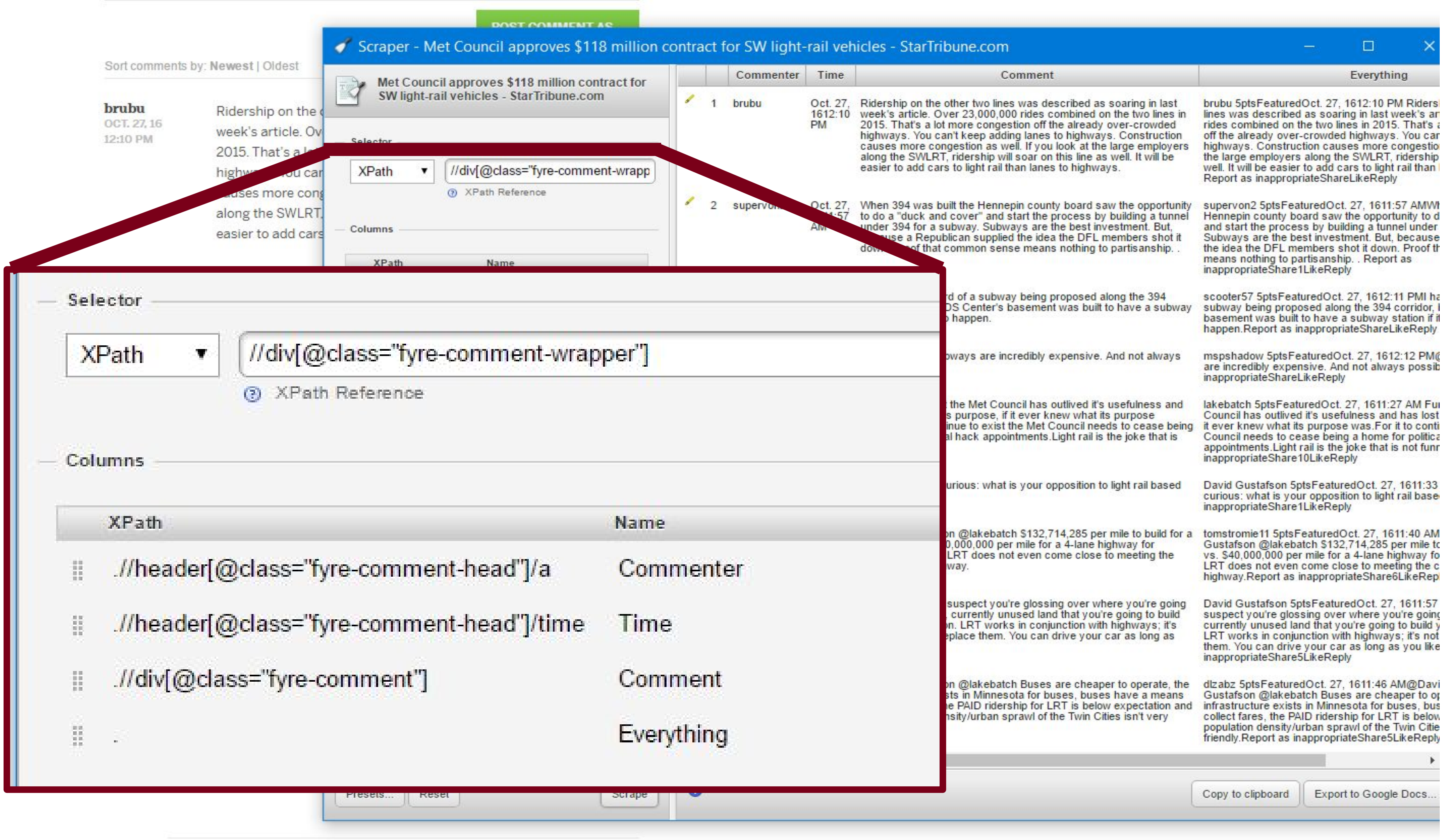
Actor	Date	Title
Donald Trump	June 16, 2015	Press Release - Donald J. Trump Declares Candidacy for President of the United S
Donald Trump	June 16, 2015	Press Re Southern
Donald Trump	June 16, 2015	Press Re
Donald Trump	June 16, 2015	Press Re Trump's
Donald Trump	June 16, 2015	Press Re Traitor B
Donald Trump	June 16, 2015	Press Re Inequalit
Donald Trump	June 17, 2015	Press Re
Donald Trump	June 17, 2015	Press Re
Donald Trump	June 17, 2015	Press Re
Donald Trump	June 17, 2015	Press Re
Donald Trump	June 17, 2015	Press Re
Donald Trump	June 17, 2015	Press Re Presiden
Donald Trump	June 17, 2015	Press Re Touch Hi
Donald Trump	June 17, 2015	Press Re
Donald Trump	June 18, 2015	Press Re Attempt
Donald Trump	June 18, 2015	Press Re
Donald Trump	June 18, 2015	Press Re
Donald Trump	June 22, 2015	Press Re
Donald Trump	June 23, 2015	Press Re Embra
Donald Trump	June 23, 2015	Press Re Obama
Donald Trump	June 26, 2015	Press Release - Donald J. Trump Endorsed by New Hampshire State Representative Stephen Stepanek
Donald Trump	July 6, 2015	Press Release - Trump on Mexico Comments: 'I Can't Apologize for the Truth'
Donald Trump	July 7, 2015	Press Release - Coulter: Trump 'Speaking for Working Class,' 'For African-

```
▼<table width="100%" border="0" cellpadding="0"
cellspacing="0" bgcolor="#FFFFFF">
  ▼<tbody>
    ▶<tr>…</tr>
      enter" cellpadding="0" cellspacing="0">
        ▼<tbody>
          ▶<tr>…</tr>
          ▼<tr bgcolor="#ffffff">
              <td bgcolor="#ffffff" class="listdate"
              valign="top">Donald Trump</td>
              <td bgcolor="#ffffff" class="listdate"
              valign="top">June 16, 2015</td>
            ▼<td bgcolor="#ffffff" class="listdate"
            valign="top">
              ▶<a href="../ws/index.php?
              pid=113866">…</a> == $0
              </td>
            </tr>
          ▶<tr bgcolor="#f7fafd">…</tr>
          ▶<tr bgcolor="#ffffff">…</tr>
          ▶<tr bgcolor="#f7fafd">…</tr>
          ▶<tr bgcolor="#ffffff">…</tr>
          ▶<tr bgcolor="#f7fafd">…</tr>
          ▶<tr bgcolor="#ffffff">…</tr>
          ▶<tr bgcolor="#f7fafd">…</tr>
          ▶<tr bgcolor="#ffffff">…</tr>
          ▶<tr bgcolor="#f7fafd">…</tr>
          ▶<tr bgcolor="#ffffff">…</tr>
          ▶<tr bgcolor="#f7fafd">…</tr>
```

Area of interest is Highlighted

Inspect to bring up Developer Console

Use the details to specify a path



Scraper - Met Council approves $118 million contract for SW light-rail vehicles - StarTribune.com

Met Council approves $118 million contract for SW light-rail vehicles - StarTribune.com

Selector

XPath ▾ //div[@class="fyre-comment-wrapper"]

XPath Reference

Columns

XPath	Name
./header[@class="fyre-comment-head"]/a	Commenter
./header[@class="fyre-comment-head"]/time	Time
./div[@class="fyre-comment"]	Comment
.	Everything

	Commenter	Time	Comment	Everything
1	brubu	Oct. 27, 1612:10 PM	Ridership on the other two lines was described as soaring in last week's article. Over 23,000,000 rides combined on the two lines in 2015. That's a lot more congestion off the already over-crowded highways. You can't keep adding lanes to highways. Construction causes more congestion as well. If you look at the large employers along the SWLRT, ridership will soar on this line as well. It will be easier to add cars to light rail than lanes to highways.	brubu 5ptsFeaturedOct. 27, 1612:10 PM Riders lines was described as soaring in last week's ar rides combined on the two lines in 2015. That's a off the already over-crowded highways. You car highways. Construction causes more congestio the large employers along the SWLRT, ridership well. It will be easier to add cars to light rail than Report as inappropriateShareLikeReply
2	supervo	Oct. 27, 1611:57 AM	When 394 was built the Hennepin county board saw the opportunity to do a "duck and cover" and start the process by building a tunnel under 394 for a subway. Subways are the best investment. But, because a Republican supplied the idea the DFL members shot it down. Proof that common sense means nothing to partisanship. .	supervon2 5ptsFeaturedOct. 27, 1611:57 AMWh Hennepin county board saw the opportunity to d and start the process by building a tunnel under Subways are the best investment. But, because the idea the DFL members shot it down. Proof th means nothing to partisanship. . Report as inappropriateShare1LikeReply

Copy to clipboard | Export to Google Docs...

Sort comments by: Newest | Oldest

brubu
OCT. 27, 16
12:10 PM

Ridership on the other two lines was described as soaring in last week's article. Over 23,000,000 rides combined on the two lines in 2015. That's a lot more congestion off the already over-crowded highways. You can't keep adding lanes to highways. Construction causes more congestion as well. If you look at the large employers along the SWLRT, ridership will soar on this line as well. It will be easier to add cars to light rail than lanes to highways.

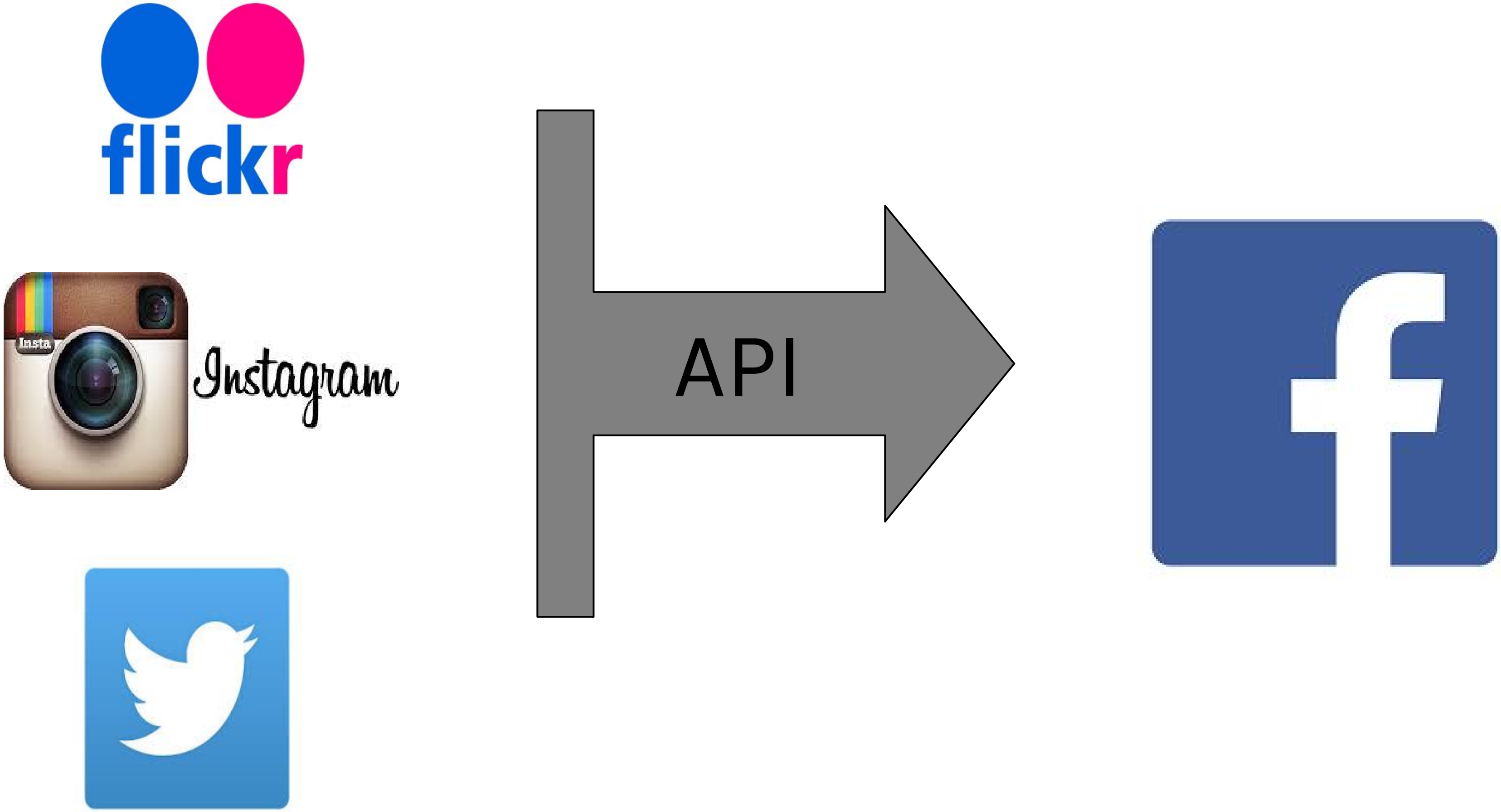
Forget what you see - ask the computer to get your data

Using an Application Programming Interface (API)

- Structure that defines how programs can communicate with each other
  - Sharing information
  - Taking in information

- Structured requests; structured responses

Application Programming Interface (API)



Defined Queries & Limitations

- Set actions with defined parameters
  - Twitter: https://dev.twitter.com/rest/public
  - Facebook: https://developers.facebook.com/docs/graph-api/
- Set Limitations
  - Privacy: can't get everything you may be able to see
  - Rate limitations: can't make unlimited calls
  - Return limitations: can't get all the data at once

Example Tools for Querying APIs

FacePager

Point & click interface for querying Twitter, Facebook, and general APIs

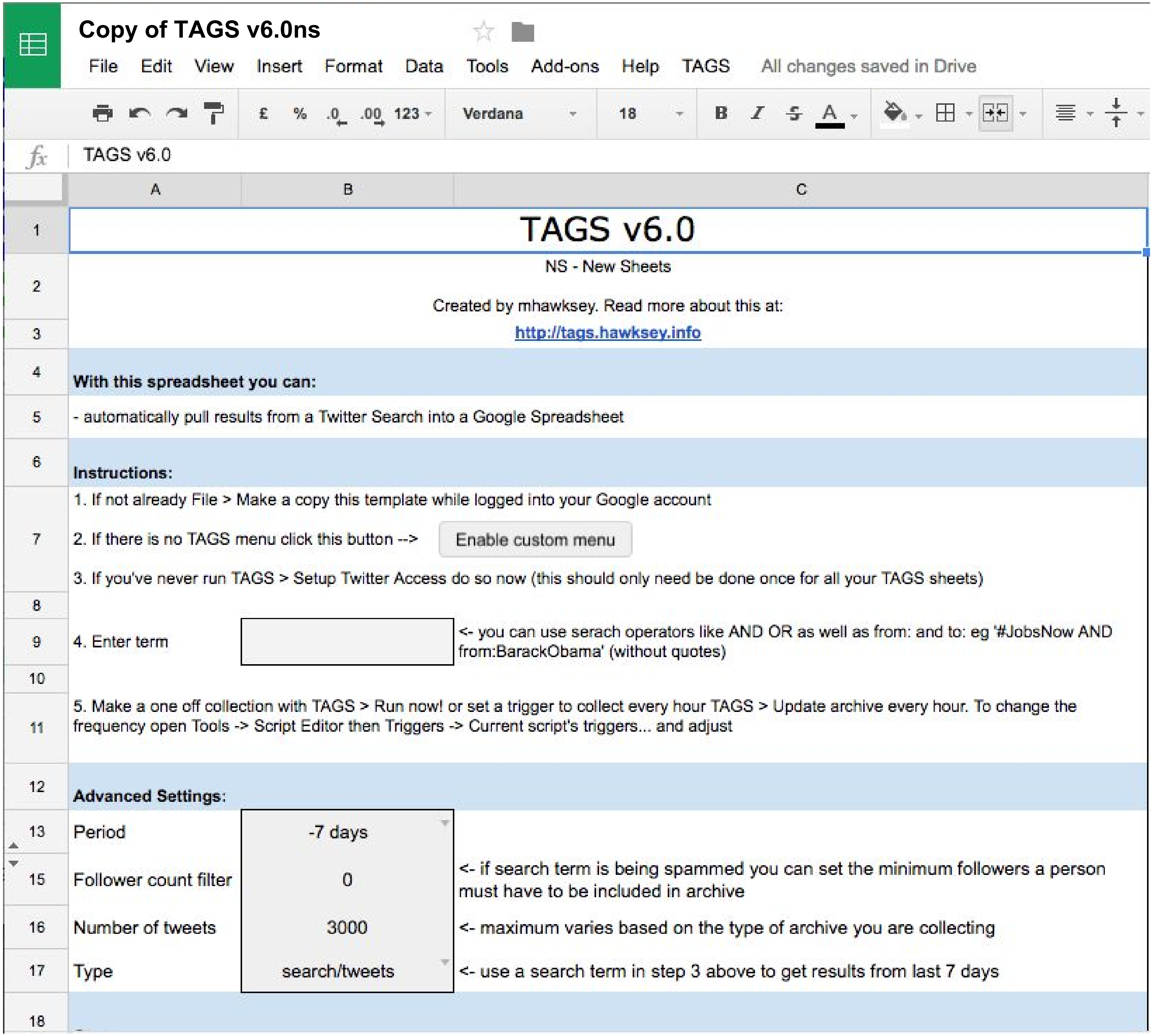
Twitter Archiving Google Spreadsheet

Very easy-to-use script for Google sheets that accesses Twitter's streaming API to collect and archive tweets

R

Packages for accessing APIs for Twitter, Facebook, Elsevier, and general sites (TwitteR, Rfacebook, rscoipus, httr)

TAGS


Copy of TAGS v6.0ns
File Edit View Insert Format Data Tools Add-ons Help TAGS All changes saved in Drive
£ % .0 .00 123
Verdana
18
B I S A
fx
A
B
C

TAGS v6.0

NS - New Sheets

Created by mhawksey. Read more about this at:
http://tags.hawksey.info

With this spreadsheet you can:

- automatically pull results from a Twitter Search into a Google Spreadsheet

Instructions:

1. If not already File > Make a copy this template while logged into your Google account

2. If there is no TAGS menu click this button --> Enable custom menu

3. If you've never run TAGS > Setup Twitter Access do so now (this should only need be done once for all your TAGS sheets)

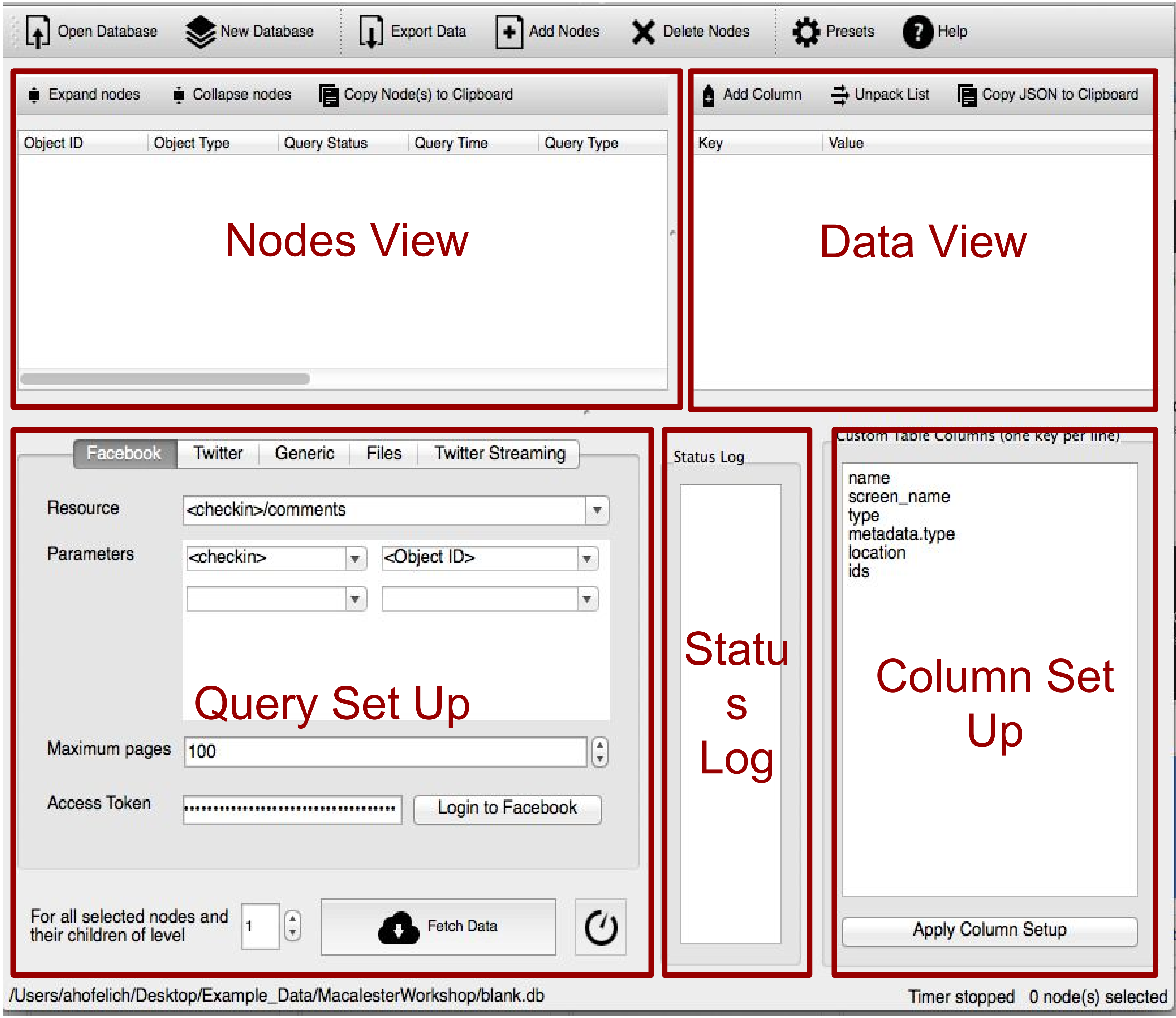
4. Enter term <- you can use serach operators like AND OR as well as from: and to: eg '#JobsNow AND from:BarackObama' (without quotes)

5. Make a one off collection with TAGS > Run now! or set a trigger to collect every hour TAGS > Update archive every hour. To change the frequency open Tools -> Script Editor then Triggers -> Current script's triggers... and adjust

Advanced Settings:

Period	-7 days	
Follower count filter	0	<- if search term is being spammed you can set the minimum followers a person must have to be included in archive
Number of tweets	3000	<- maximum varies based on the type of archive you are collecting
Type	search/tweets	<- use a search term in step 3 above to get results from last 7 days

Face Pager



Resource <checkin>/comments

Parameters <checkin> <Object ID>

Maximum pages 100

Access Token ••

For all selected nodes and their children of level 1

Custom Table Columns (one key per line)

name
screen_name
type
metadata.type
location
ids

/Users/ahofelich/Desktop/Example_Data/MacalesterWorkshop/blank.db

Timer stopped 0 node(s) selected

Tool	Use	Ease to Learn	Extensibility
Get Them All	Save what you See	Easy	Limited
HTTrack		Easy	Limited
If this then that		Medium	Some
Scraper	Structure what you see	Medium	Some
Twitter Archiving Google Sheet	Query APIs	Easy	Limited
FacePager		Medium	Some
Python	Structure what you see & query APIs	Difficult	High
R		Difficult	High

Thank you! Questions?

Michael Beckstrand

mjbeckst@umn.edu

Alicia Hofelich Mohr

hofelich@umn.edu