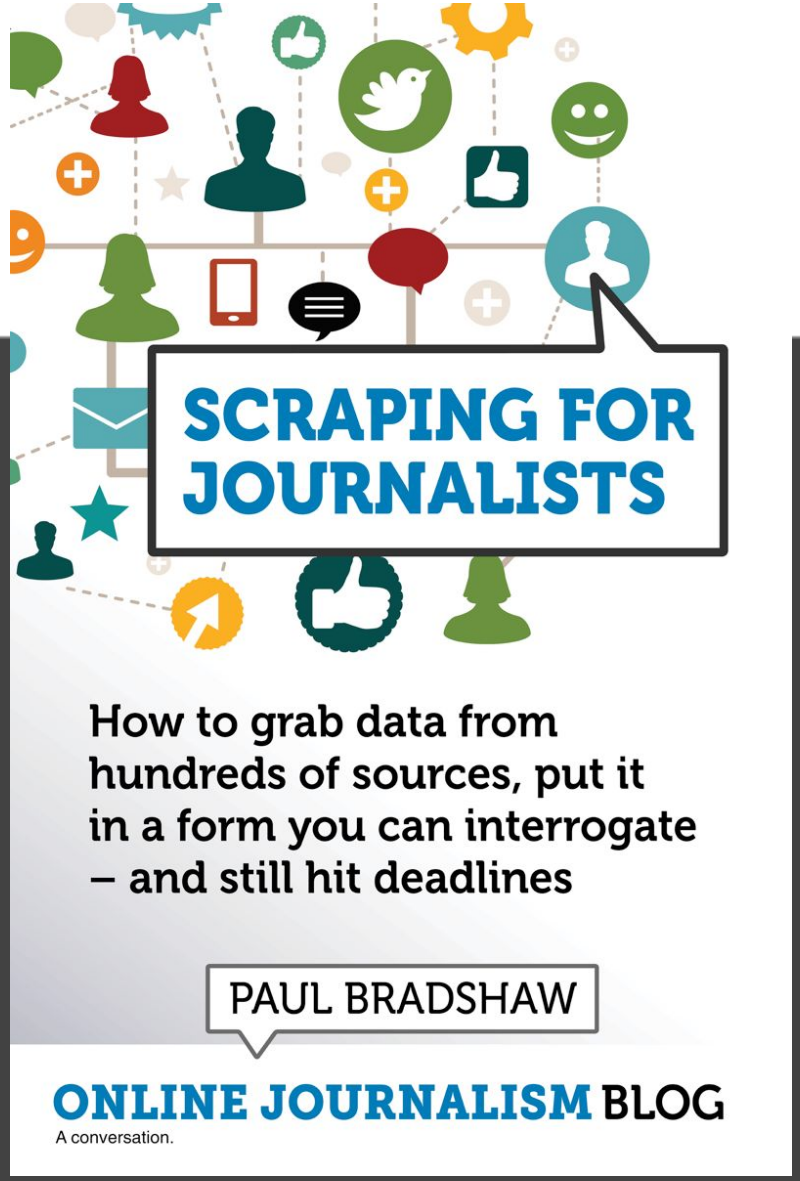
Tricky pages



Paul Bradshaw

Leanpub.com/scrapingforjournalists

What we'll cover

- What to do when your scraper doesn't 'see' what you see
- Using the inspector
- 'Throttling' a scraper to slow it down

Why isn’t it working?

Check the HTML of the webpage that’s scraped - it may be different to what you see

Why might it be different? *Dynamically generated content*

Looking for the source of dynamic content

1. Open the Inspector
2. Go to the Network tab (specifically XHR)
3. Then refresh the page

Try these examples…

Fatal Incident reports

This is a searchable record of our final reports, published on our website once they have been shared with the family of the deceased person, and the coroner's inquest has taken place. Names have been removed to protect the privacy of the individuals concerned. For deaths after 1 March 2015, the name of the deceased remains in the report, but other names have been anonymised.

If you would like an update on a report that you cannot find or if you spot any errors, please email: ppocomms@ppo.gov.uk. Please use the filters to limit the display to specific case types and the sort buttons to reorder by either date of death or date it went on the website.

Hide details

St Leonards

Published: 07/04/2021

Approved premises

Date of death:	25/07/2020
Cause:	Self-inflicted
Gender:	Male
Age:	22-30

PPO Report (PDF, 600 KB)

Action Plan (PDF, 135 KB)

Bure

Published: 07/06/2021

Prison

Date of death:	17/05/2020
Cause:	Self-inflicted
Gender:	Male
Age:	41-50

PPO Report (PDF, 570 KB)

Action Plan (PDF, 129 KB)

Durham

Published: 30/07/2021

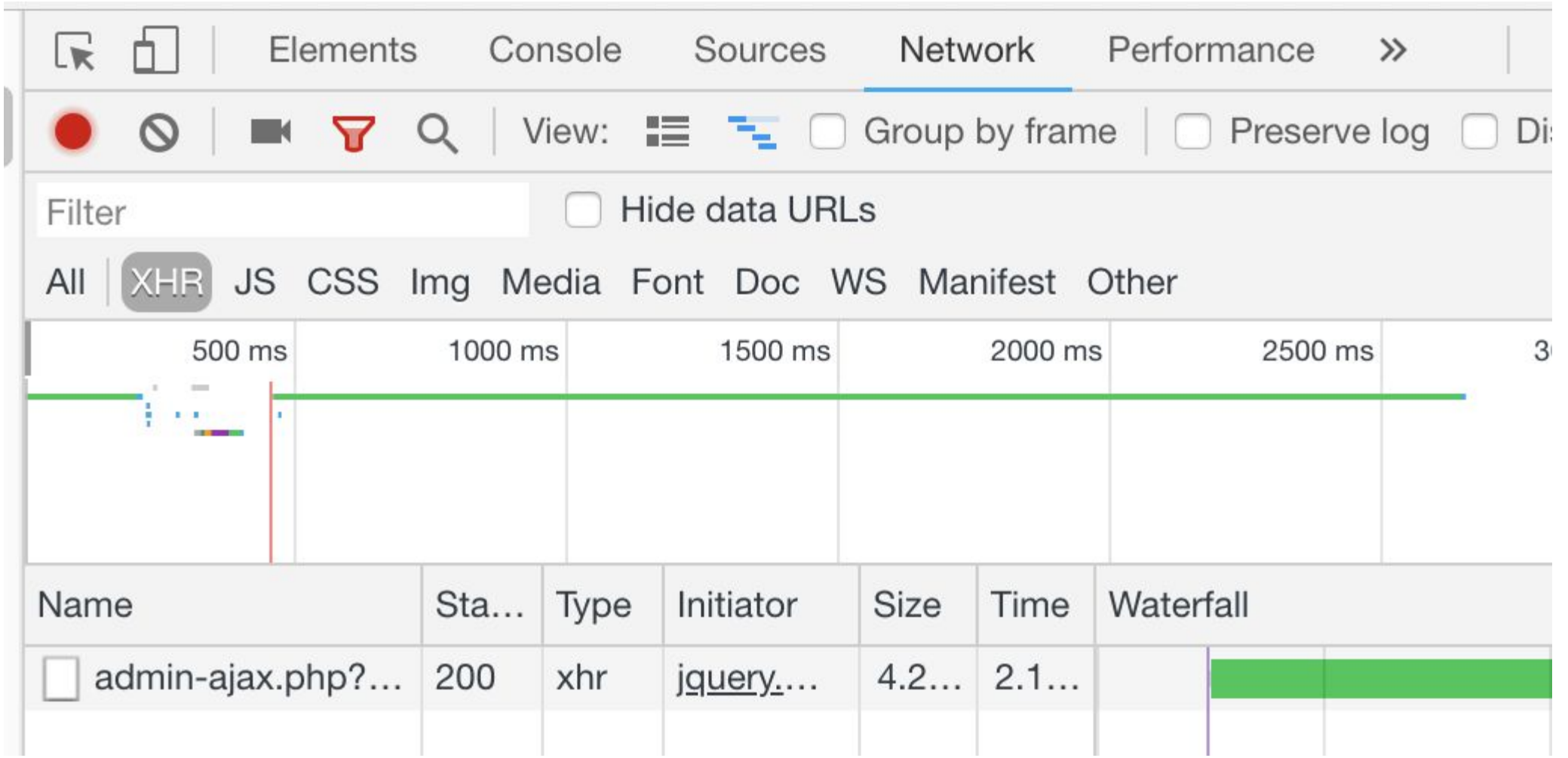
Prison

Date of death:	02/05/2020
Cause:	Self-inflicted
Gender:	Male
Age:	41-50

PPO Report (PDF, 636 KB)

Action Plan (PDF, 147 KB)

https://www.ppo.gov.uk/document/fii-report/


Elements
Console
Sources
Network
Performance
View:
Group by frame
Preserve log
Filter
Hide data URLs
All
XHR
JS
CSS
Img
Media
Font
Doc
WS
Manifest
Other
500 ms
1000 ms
1500 ms
2000 ms
2500 ms
Name
Sta…
Type
Initiator
Size
Time
Waterfall
admin-ajax.php?…
200
xhr
jquery.…
4.2…
2.1…

Altcourse

Published: 28/07/2021

Prison

Date of death: 14/02/2021
Cause: Natural causes
Gender: Male
Age: 61+

- PPO Report (PDF, 544 KB)
- Action Plan (PDF, 64 KB)

Elmley

Published: 16/06/2021

Prison

Date of death: 14/02/2021
Cause: Natural causes
Gender: Male
Age: 61+

- PPO Report (PDF, 521 KB)
- Action Plan (PDF, 95 KB)

Altcourse

Published: 22/07/2021

Prison

Date of death: 13/02/2021
Cause: Natural causes
Gender: Male
Age: 41-50

https://www.ppo.gov.uk/wp/wp-admin/admin-ajax.php?action=update_tiles&queryParams=%7B%22document_type%22%3A%22fii-report%22%2C%22posts_per_page%22%3A50%2C%22paged%22%3A1%2C%22order%22%3A%22DESC%22%2C%22orderby%22%3A%22meta_value%22%2C%22meta_key%22%3A%22fii-death-date%22%7D

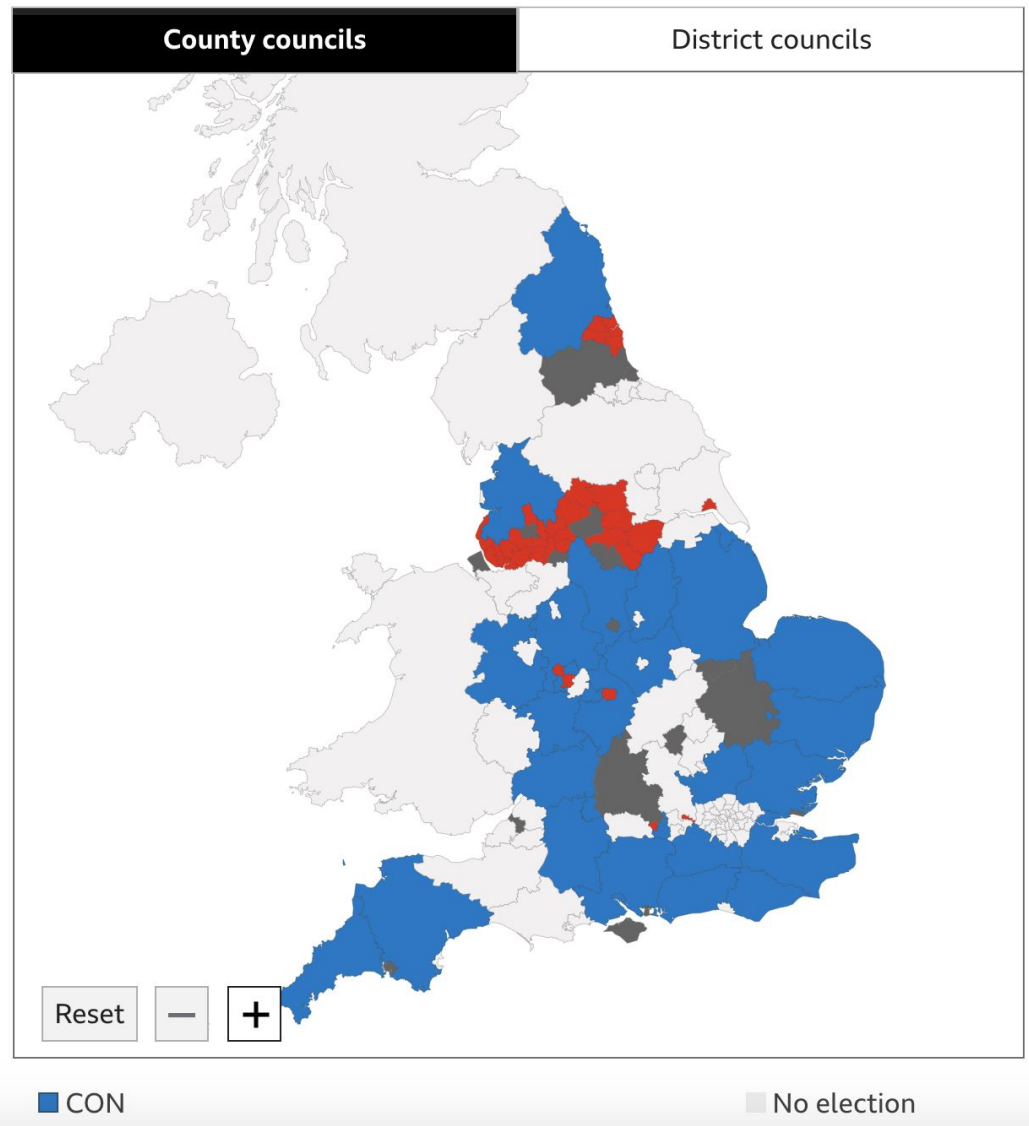
Look for:

- Files which are large (you can sort by size)
- Files with promising names
- Files with ? in (indicates a query to an API)
- Files ending in .json or .xml or .csv

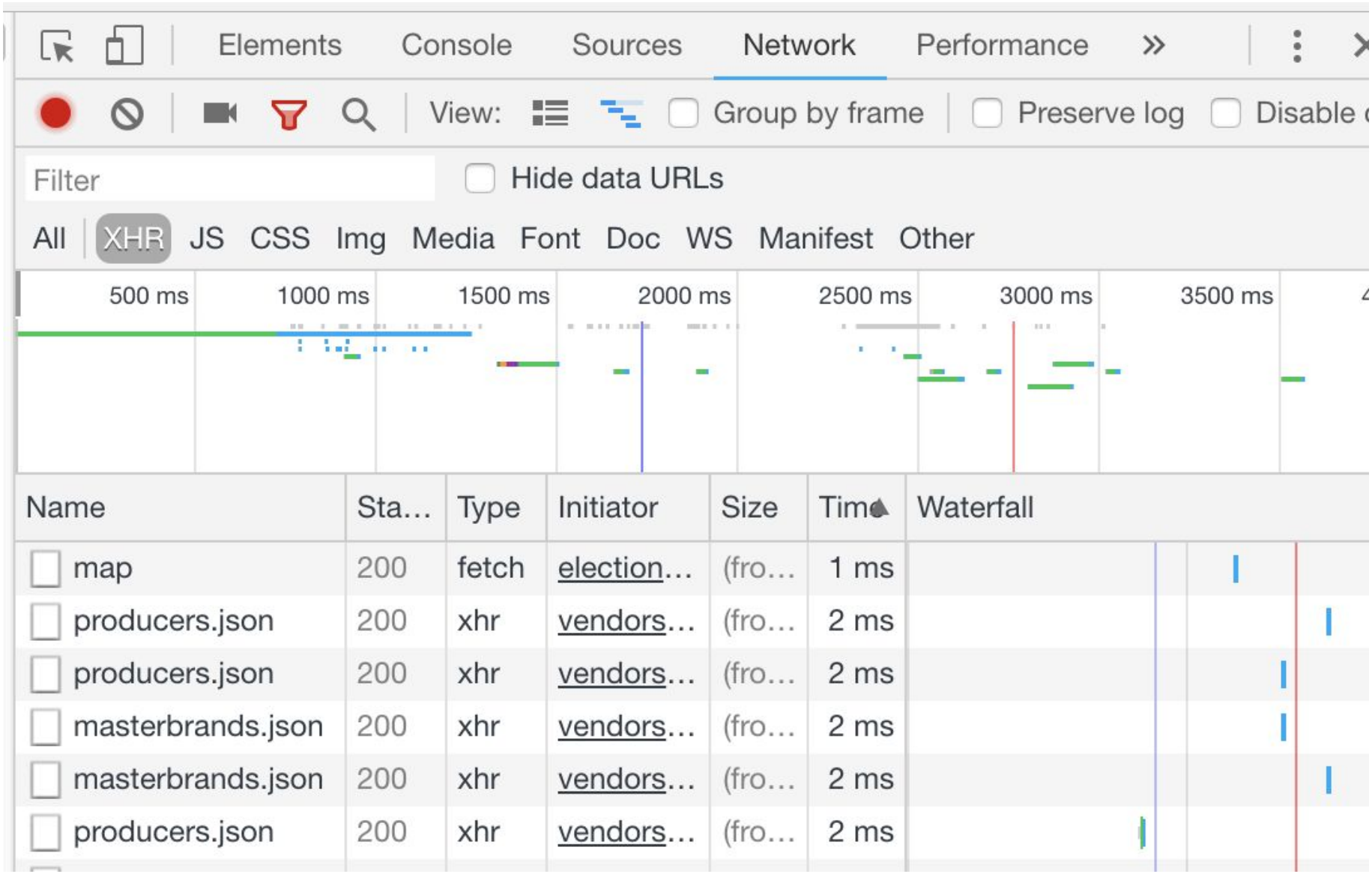
England local elections 2021

Conservatives make council gains across England

The Conservatives have made significant gains across E
extra 235 councillors to their overall tally by Sunday. La
councils, including **Durham County Council**. The party f
as Sadiq Khan was re-elected as **London mayor**, Andy B
Mayor of Greater Manchester, and Tracy Brabin becom
mayor. Labour also won mayoral elections in **Cambridge**
City Region, and **West of England**



https://www.bbc.co.uk/news/topics/c481drqqzv7t/england-local-elections-2021



https://static.files.bbci.co.uk/elections/data/news/election/2021/england/councils/map

```
{
  - partyColours: {
        CON: "#0575C9",
        GRN: "#5FB25F",
        ICHC: "#D32F6C",
        IND: "#FF66A1",
        LAB: "#E91D0E",
        LD: "#EFAC18",
        LIB: "#C7941A",
        MK: "#78721d",
        RA: "#4dadab",
        REF: "#0AD1E0",
        UKIP: "#712F87",
        YP: "#00B8FD"
    },
  - map: {
    - E06000001: {
          wp: "NOC",
          wpp: "NOC",
          flash: "NOC NO CHANGE",
          url: "/news/topics/cv8k1ezy9m9t/hartlepool-borough-council",
          name: "Hartlepool",
          yearLast: 2019
      },
    - E06000006: {
          wp: "LAB",
          wpp: "LAB",
          flash: "LAB HOLD",
          url: "/news/topics/cv8k1ezvmg8t/halton-borough-council",
          name: "Halton",
          yearLast: 2019
      },
```


https://static.files.bbci.co.uk/elections/data/news/election/2021/england/councils/map

https://jsonview.com/

You can import these files like any other

```
import pandas as pd

df = pd.read_json(url)

df.to_csv("mapdata.csv")
```

Throttling!

```
import time


for i in list:
  [scraping code]
  #Sleep for 3 seconds
   time.sleep(3)
```

```
#first, store the URL up to the page number
firsturlpart = "https://www.nhs.uk/service-search/other-servic
#next create a list of page numbers from 1 to 9
pagelist = range(1,10)
#then loop through them and add to the URL
for i in pagelist:
  #convert number to string so it can be combined with URL
  pagenumberasstring = str(i)
  #combine that with URL
  pageurl = firsturlpart+pagenumberasstring
  #scrape the page and store results in df
  df = scrapepage(pageurl)
  print(df)
  #add the new data frame to the existing data frame
  dfhere = dfhere.append(df)
  print(dfhere)
  print("waiting 3 seconds before next scrape")
  #Sleep for 3 seconds before looping again
  time.sleep(3)
```

Recap

- Watch out for dynamically generated pages - the data could be there to grab
- If you’re getting an error after a certain number of pages, try to ‘throttle’ the scraper by adding a delay