Getting data from the Internet

Peter Smyth
Cathie Marsh Institute

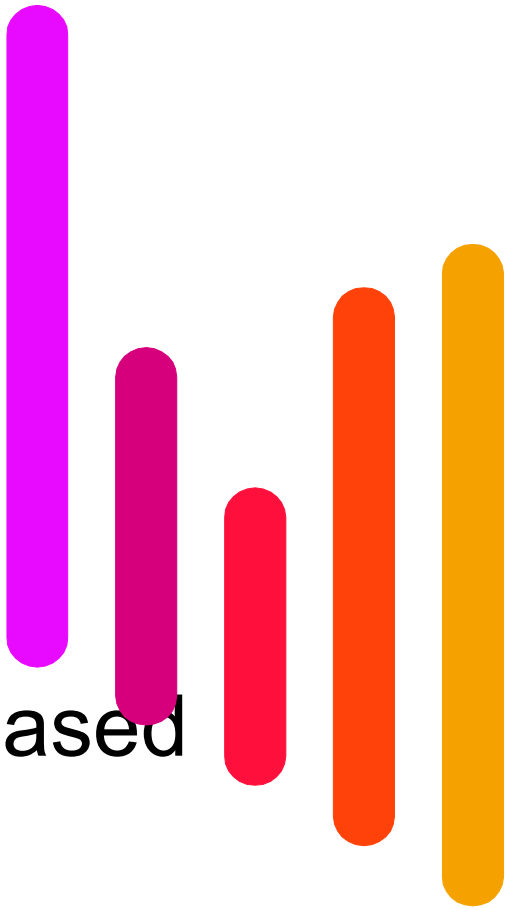
April 2020

UK Data Service

Overview of Seminar

- Copy & Paste
- Downloading files
- Using APIs
- Web scraping

- All very much demonstration based



Disclaimer & Explanation

- This Webinar is NOT about finding data on the Internet
- It is about using different techniques and software tools to download data available in different formats on the Internet
- With the exception of Excel all of the tools are open source

Copy & Paste

- For small discrete amounts of data, the simplest approach
- But there are pitfalls – What You Get isn't necessarily What You See!
- Different browsers can give different results
- For Tables Excel may have a better alternative.

DEMO

Copy & Paste Demo

- Copy the premier league football table from the BBC website and paste it into Excel
- Does the Browser used make a difference?
- How to download a Web page table using Excel

Downloading files or different types

- Click on the download button – if available
- Finding the URL of the file
- Use software tools or code
  - Wget and Curl
  - Python code

Using code provides better documentation / reproducibility than point & click

DEMO

Multiple file downloads

- Automation is not always possible
  - UUIDs in path or other randomness
  - HTML calls JavaScript function

- Can automate if paths & filenames are structured and you can anticipate
- Can also work with API type calls for a download

Using an API

- https://www.football-data.org/

- Always read the API documentation
  - How it works!
  - Need for keys
  - Call restrictions

Using an API

- Demo
- Football match results

Real Web Scraping

- Need suitable software tools, such as the BeautifulSoup package in Python
- Need some (but not a lot of) understanding of how HTML works
- Most importantly you need to be able to match what you see in your Browser screen to the underlying HTML which puts it there. – There is a tool for that!

HTML Tags (examples)

```
<h1>Basic HTML tags</h1>


<h3>Lists can be nested</h3>

<ol>
  <li> Item 1 </li>
    <ul>
                    <li> Item 1 </li>
                    <li> Item 2 </li>
    </ul>
  <li> Item 3 </li>
</ol>
```


HTML Tags (examples)

Other elements are used almost entirely to aid formatting and presentation

Some spanned text

Some text in between `span` tags

HTML Tags (examples)

```
<table>
   <thead>
     <tr>
       <th>Item</th>
       <th>Amount</th>
     </tr>
   </thead>
   <tbody>
     <tr>
       <td>Bananas</td>
       <td>$8</td>
     </tr>
   </tbody>
 </table>
```

There are more than one way to specify a table

Real Web Scraping

- Demo
- Tesco store locations

Questions

Peter Smyth

Peter.smyth@manchester.ac.uk

ukdataservice.ac.uk/help/

Subscribe to the UK Data Service news list at https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE

Follow us on Twitter https://twitter.com/UKDataService or Facebook https://www.facebook.com/UKDataService