Supporting, delivering and facilitating high quality opportunities and training by engaging with students, staff and partners to positively impact the social science community in Scotland

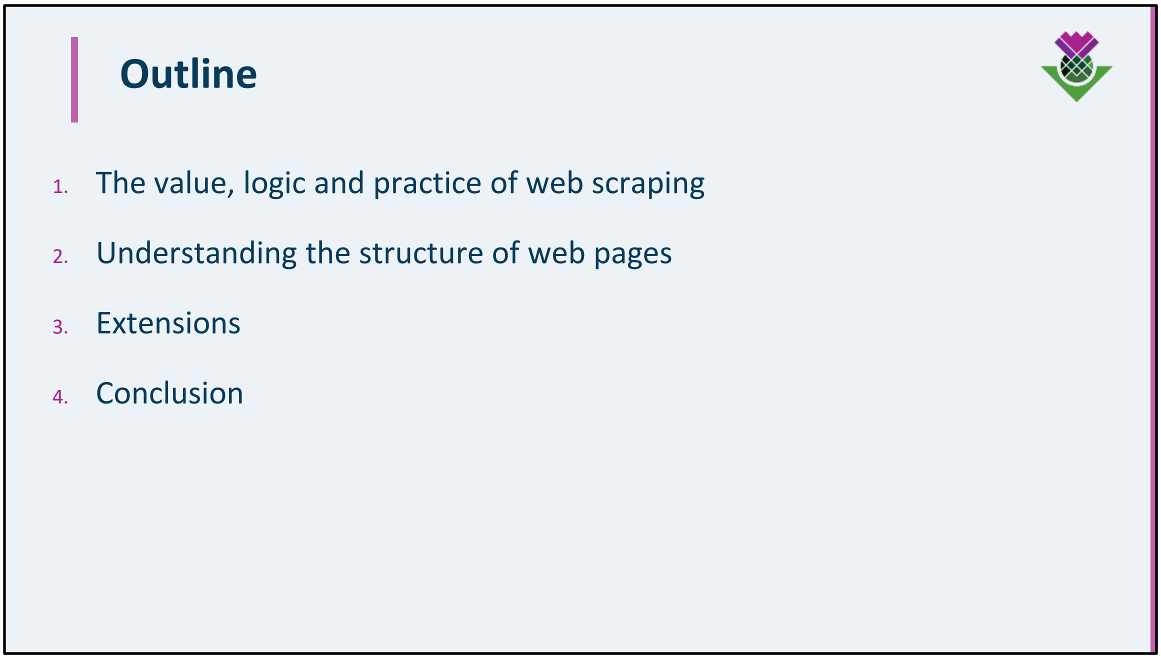
Practical Computational Methods for Social Scientists

2025-03-07
Dr Diarmuid McDonnell

Scottish Graduate School of Social Science
Sgoil Cheumnaichean Saidheans Sòisealta na h-Alba

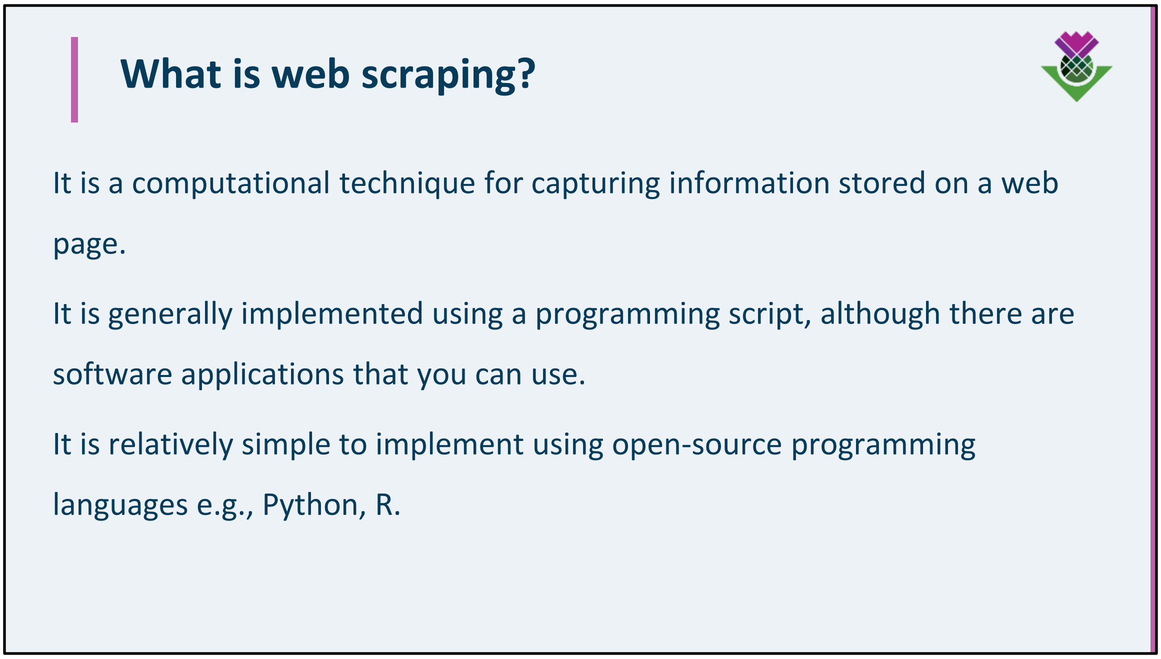
UKRI Economic and Social Research Council

Scottish Funding Council
Promoting further and higher education



Outline

1. The value, logic and practice of web scraping
2. Understanding the structure of web pages
3. Extensions
4. Conclusion



What is web scraping?

It is a computational technique for capturing information stored on a web page.

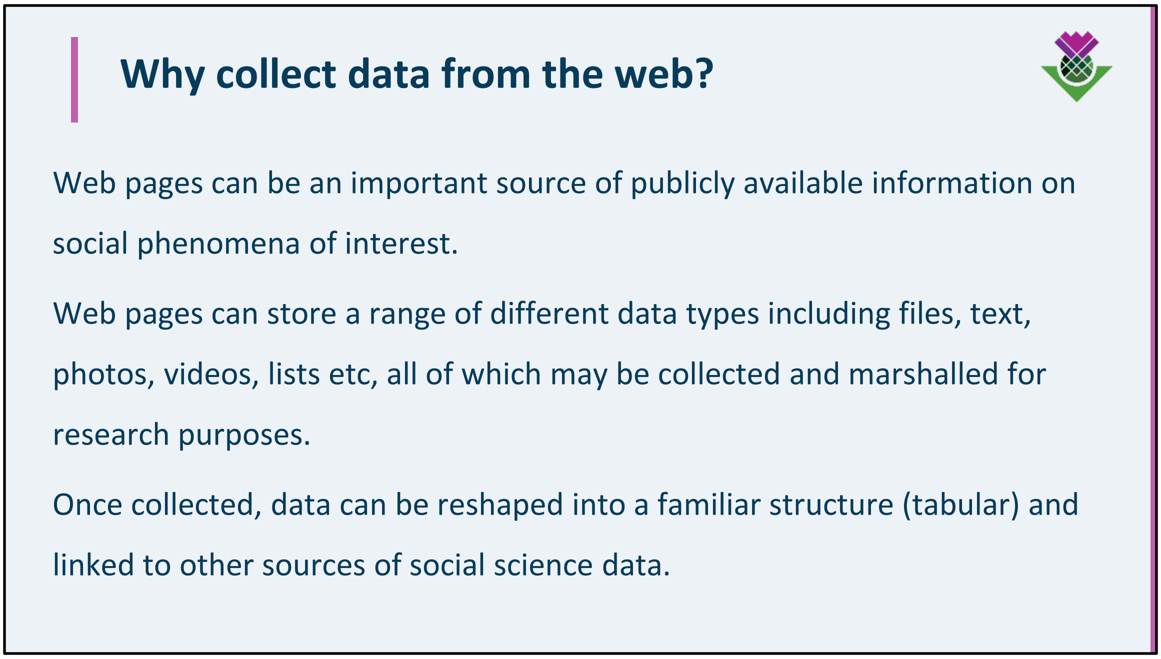
It is generally implemented using a programming script, although there are software applications that you can use.

It is relatively simple to implement using open-source programming languages e.g., Python, R.

Computational is the key difference: copy-and-pasting information from web pages is usually allowed (or undetectable!), though the manual approach carries considerable disadvantages in terms of accuracy and labour resource.

I always advocate the flexibility of writing your own code (nevermind the intellectual benefits that accrue from learning such skills), however the ultimate aim is to collect data, so there are out-of-the-box solutions you can avail of e.g., Excel, Scrapy.

You do not need to be highly computationally literate, nor write screeds of code: this is a popular and mature computational method, with tons of documentation and examples for you to learn from.



Why collect data from the web?

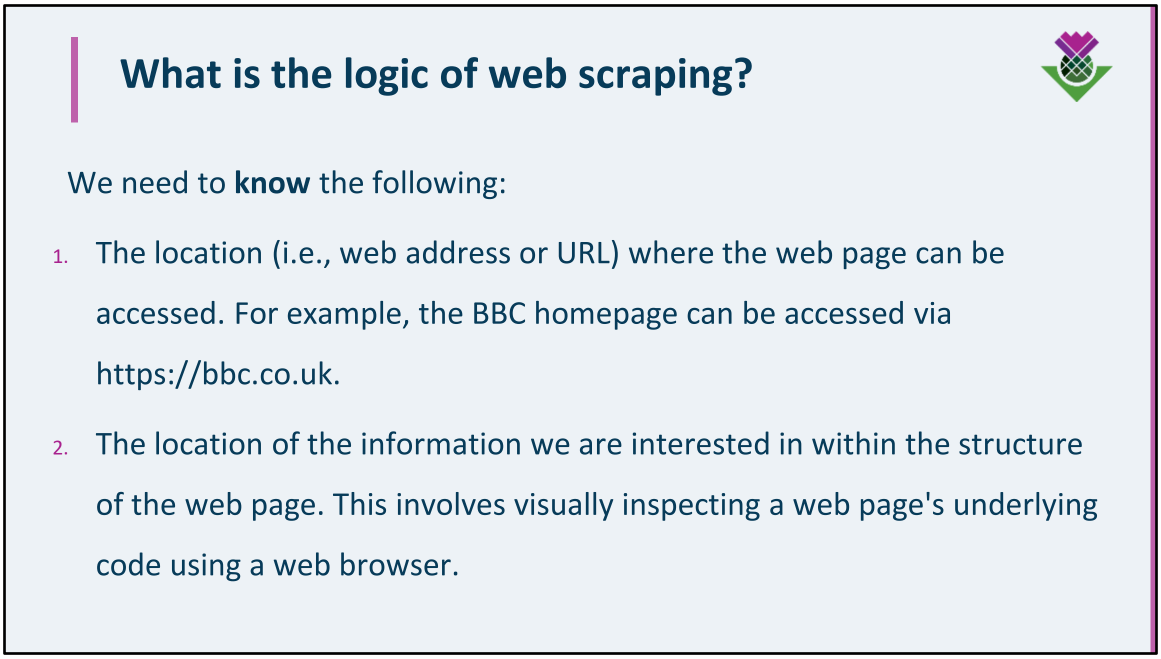
Web pages can be an important source of publicly available information on social phenomena of interest.

Web pages can store a range of different data types including files, text, photos, videos, lists etc, all of which may be collected and marshalled for research purposes.

Once collected, data can be reshaped into a familiar structure (tabular) and linked to other sources of social science data.

Coming from a social research perspective here, though it is of course commercially valuable also e.g., Google or price comparison sites.

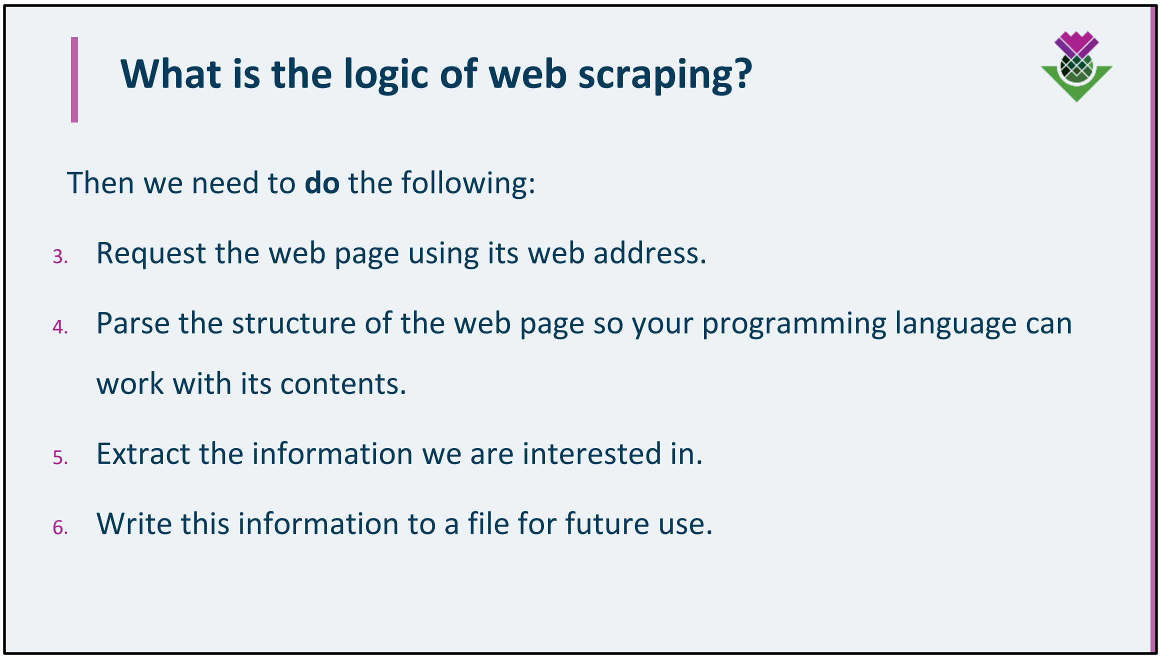
However, the data stored on websites are typically not structured or formatted for ease of use by researchers: for example, it may not be possible to perform a bulk download of all the files you need (think of needing the annual accounts of all registered companies in London for your research...), or the information may not even be held in a file and instead spread across paragraphs and tables throughout a web page (or worse, web pages). Luckily, web-scraping provides a means of quickly and accurately capturing and formatting data stored on web pages.



What is the logic of web scraping?

We need to **know** the following:

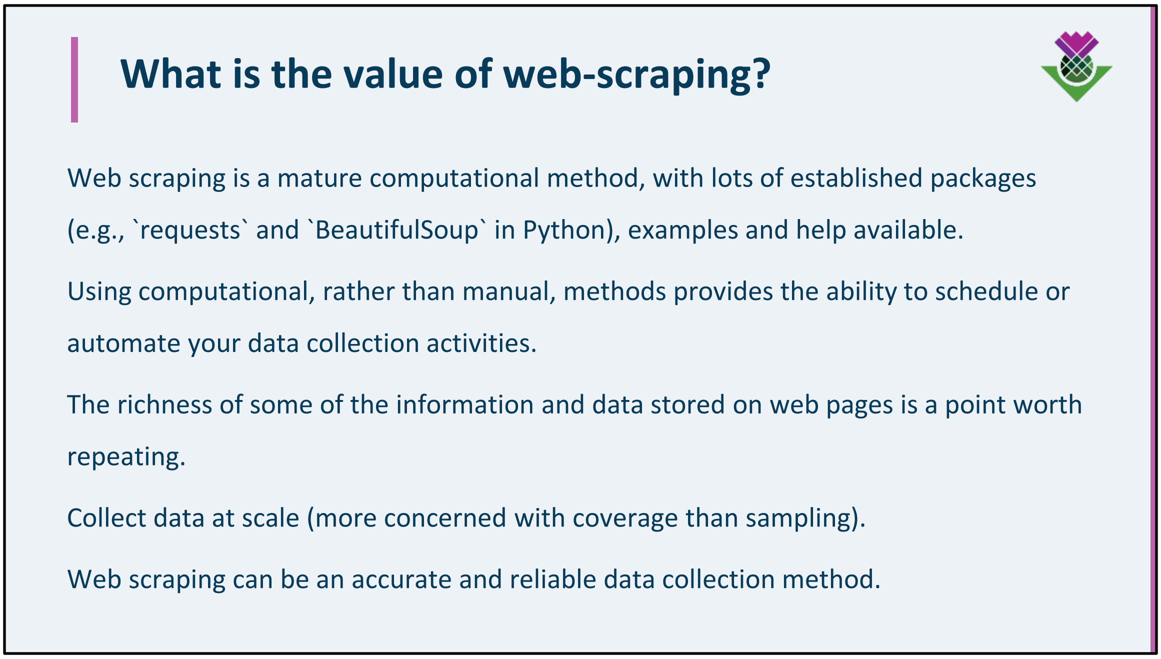
1. The location (i.e., web address or URL) where the web page can be accessed. For example, the BBC homepage can be accessed via https://bbc.co.uk.
2. The location of the information we are interested in within the structure of the web page. This involves visually inspecting a web page's underlying code using a web browser.



What is the logic of web scraping?

Then we need to **do** the following:

3. Request the web page using its web address.
4. Parse the structure of the web page so your programming language can work with its contents.
5. Extract the information we are interested in.
6. Write this information to a file for future use.



What is the value of web-scraping?

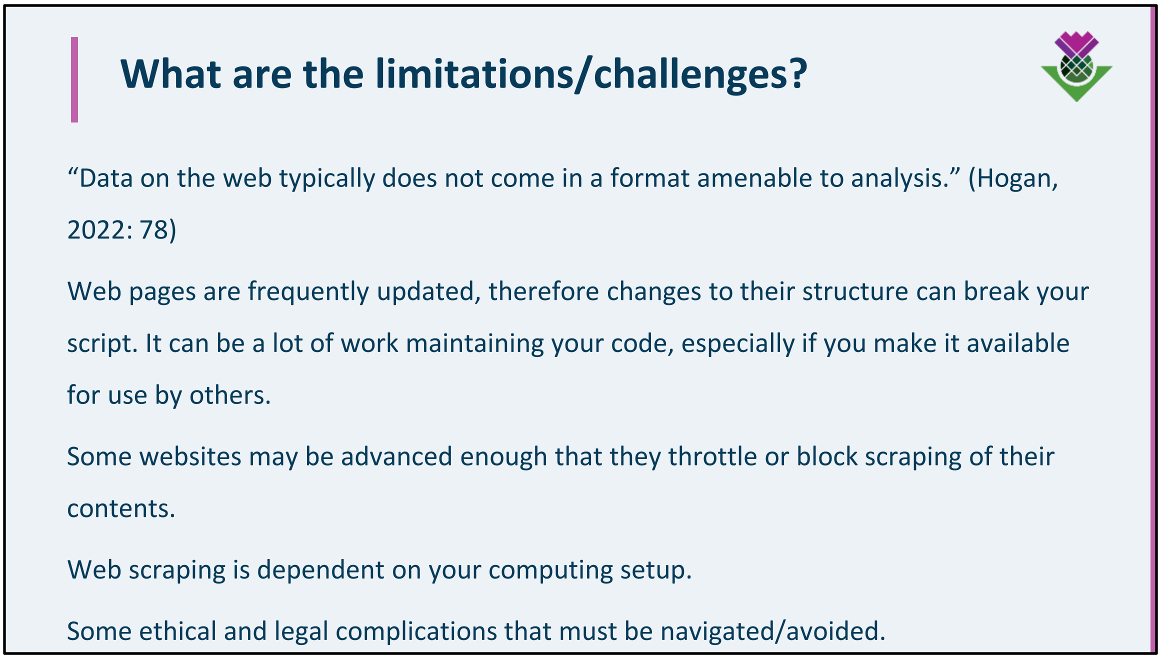
Web scraping is a mature computational method, with lots of established packages (e.g., `requests` and `BeautifulSoup` in Python), examples and help available.

Using computational, rather than manual, methods provides the ability to schedule or automate your data collection activities.

The richness of some of the information and data stored on web pages is a point worth repeating.

Collect data at scale (more concerned with coverage than sampling).

Web scraping can be an accurate and reliable data collection method.



What are the limitations/challenges?

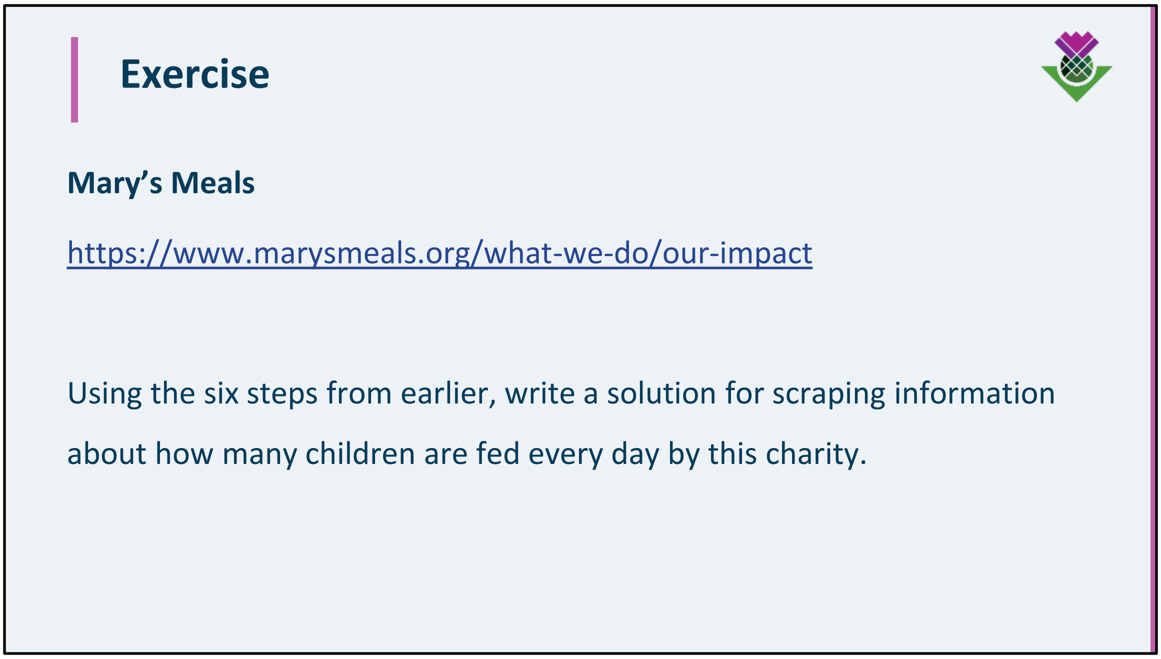
“Data on the web typically does not come in a format amenable to analysis.” (Hogan, 2022: 78)

Web pages are frequently updated, therefore changes to their structure can break your script. It can be a lot of work maintaining your code, especially if you make it available for use by others.

Some websites may be advanced enough that they throttle or block scraping of their contents.

Web scraping is dependent on your computing setup.

Some ethical and legal complications that must be navigated/avoided.

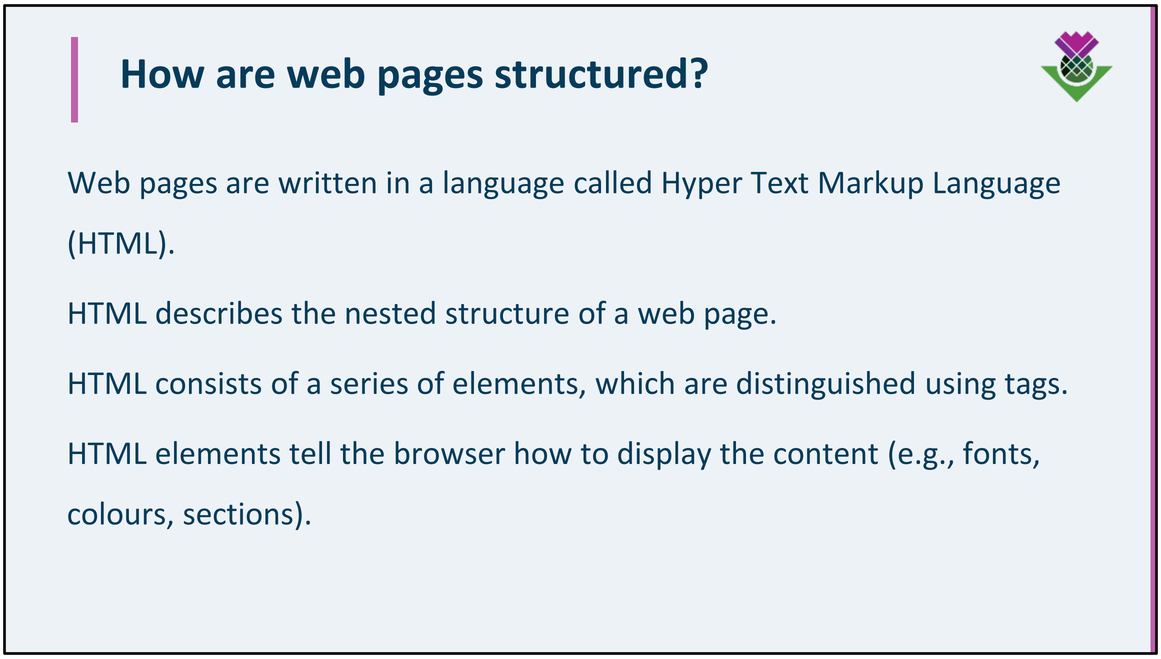


Exercise

Mary's Meals

https://www.marysmeals.org/what-we-do/our-impact

Using the six steps from earlier, write a solution for scraping information about how many children are fed every day by this charity.



How are web pages structured?

Web pages are written in a language called Hyper Text Markup Language (HTML).

HTML describes the nested structure of a web page.

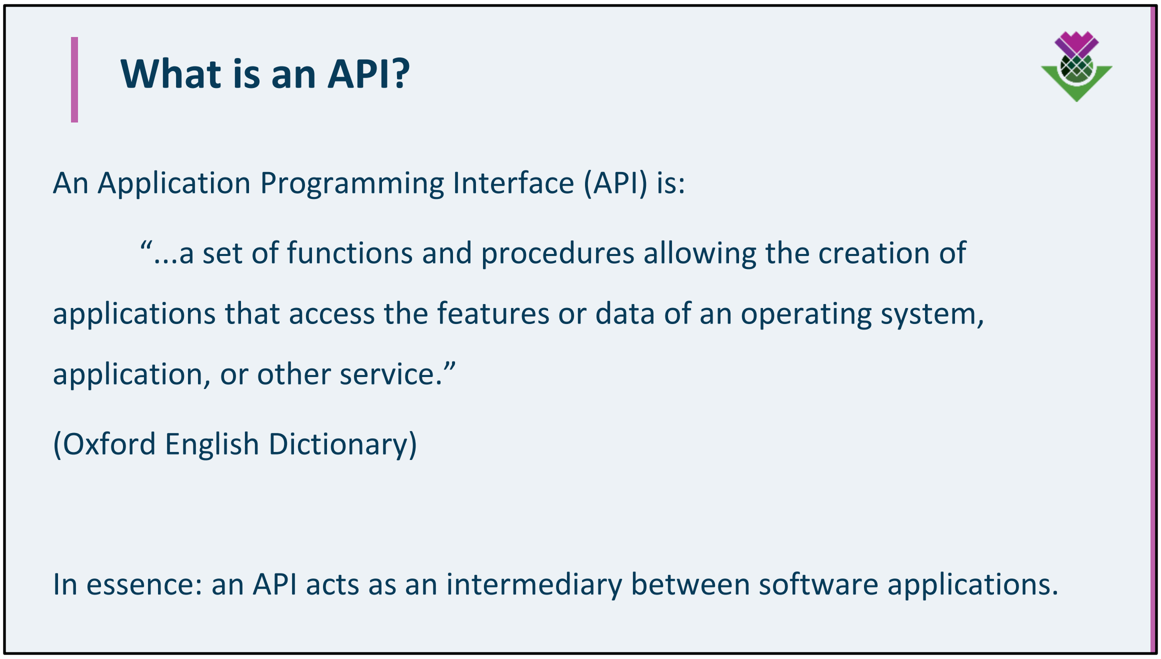
HTML consists of a series of elements, which are distinguished using tags.

HTML elements tell the browser how to display the content (e.g., fonts, colours, sections).

Markup languages typically use tags to open and close levels of the hierarchy.

Tags enclose data ("values") and also have attributes (e.g., fonts, emphasis, ids).
There are also self-closing tags that can have useful attribute information.

Best way to learn is to examine a web page’s structure.



What is an API?

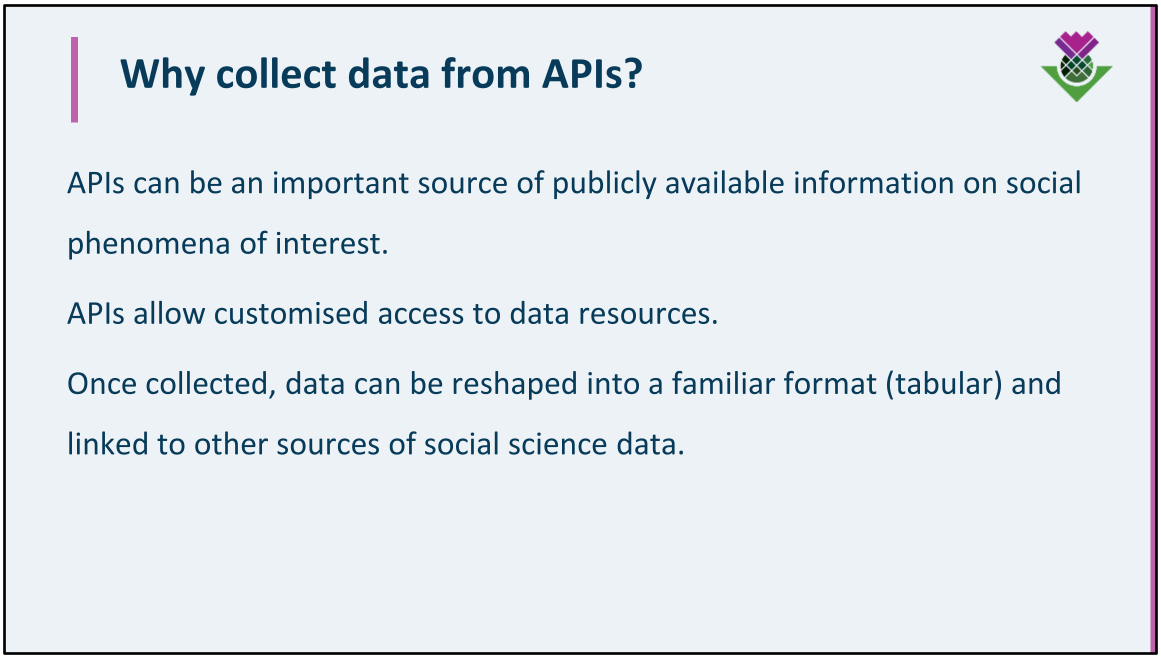
An Application Programming Interface (API) is:

> “...a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service.”

(Oxford English Dictionary)

In essence: an API acts as an intermediary between software applications.

How does an API work?



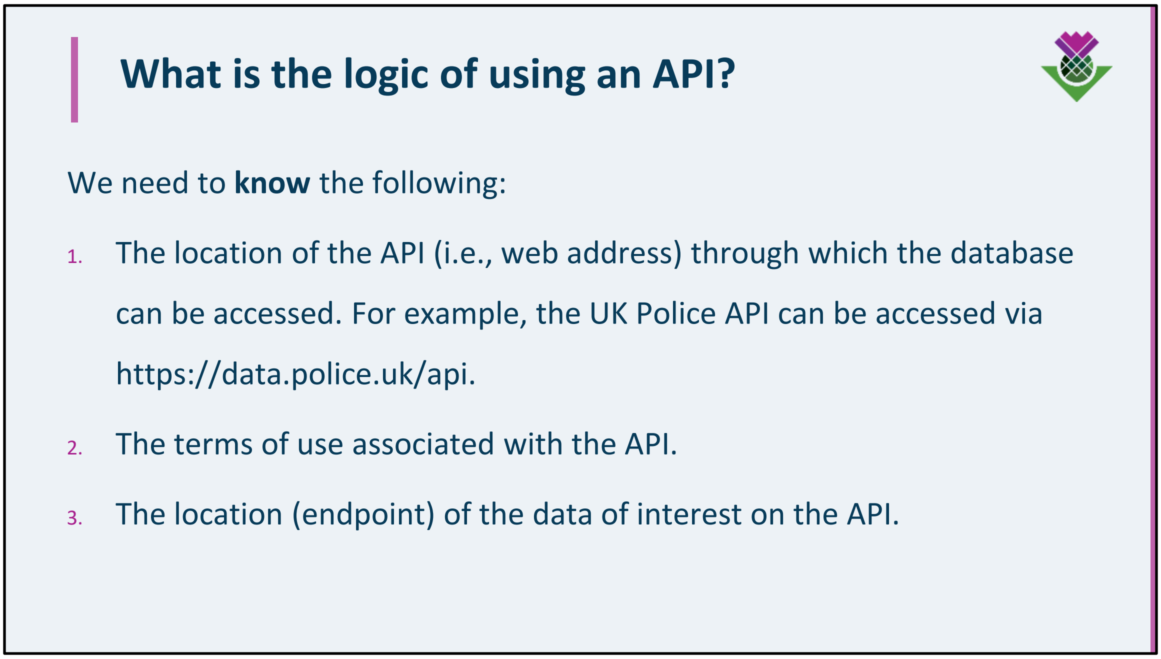
Why collect data from APIs?

APIs can be an important source of publicly available information on social phenomena of interest.

APIs allow customised access to data resources.

Once collected, data can be reshaped into a familiar format (tabular) and linked to other sources of social science data.

Similar reasons to web-scraping, but there is definitely an increasing trend towards sharing data via APIs instead / alongside of providing data downloads.



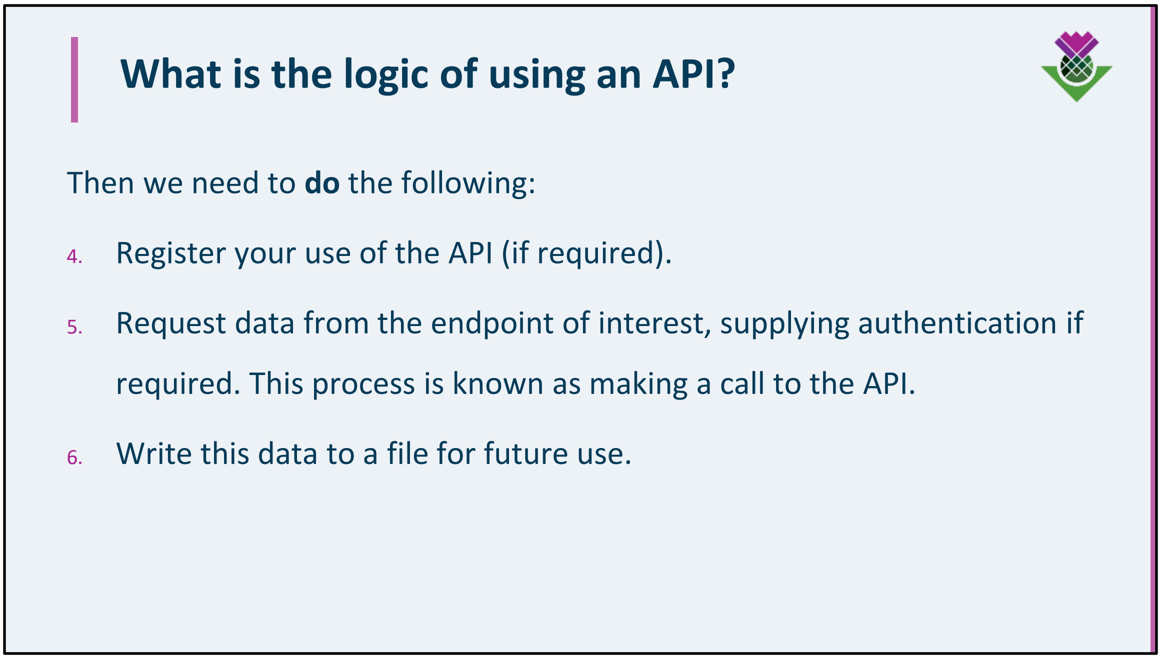
What is the logic of using an API?

We need to **know** the following:

1. The location of the API (i.e., web address) through which the database can be accessed. For example, the UK Police API can be accessed via https://data.police.uk/api.
2. The terms of use associated with the API.
3. The location (endpoint) of the data of interest on the API.

For example, the UK Police API does not require you to provide authentication but restricts the number of requests for data you can make (*rate limit*).

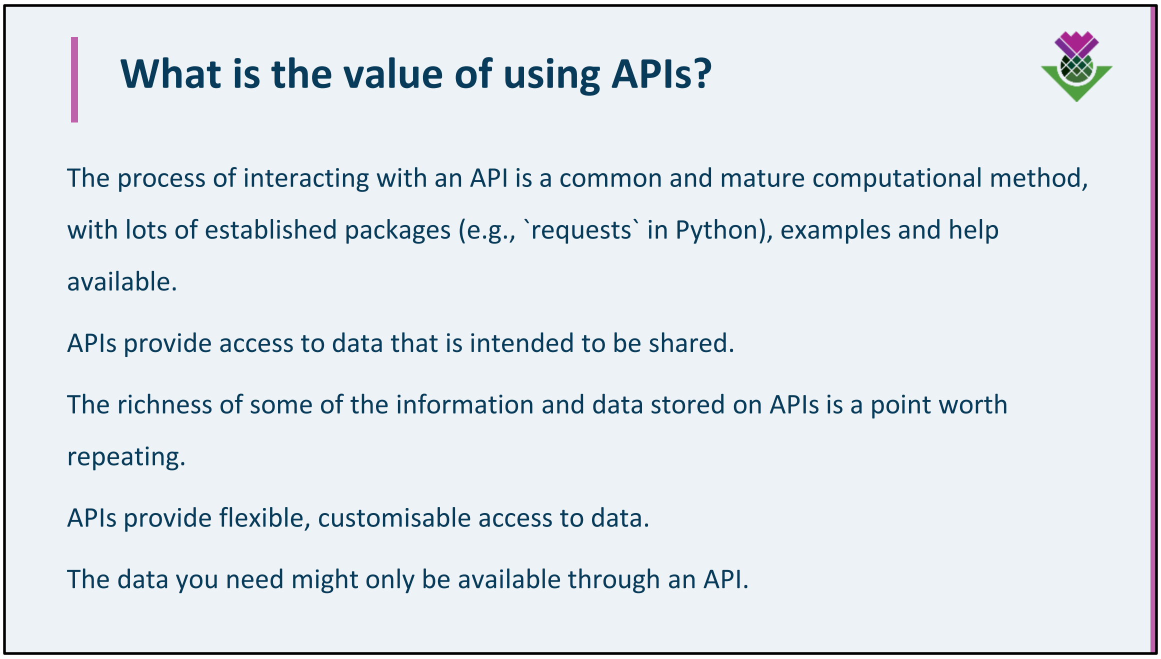
For example, data on street-level crime from the UK Police API is available at: https://data.police.uk/api/crimes-street. The location of the data is known as its endpoint.



What is the logic of using an API?

Then we need to **do** the following:

4. Register your use of the API (if required).
5. Request data from the endpoint of interest, supplying authentication if required. This process is known as making a call to the API.
6. Write this data to a file for future use.



What is the value of using APIs?

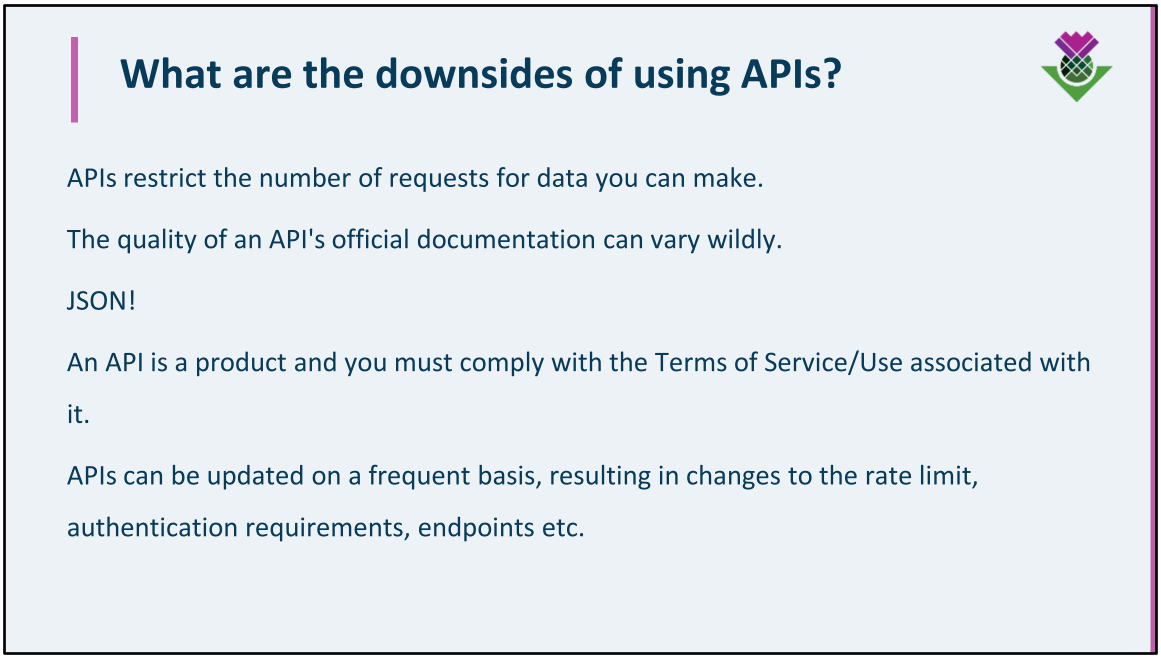
The process of interacting with an API is a common and mature computational method, with lots of established packages (e.g., `requests` in Python), examples and help available.

APIs provide access to data that is intended to be shared.

The richness of some of the information and data stored on APIs is a point worth repeating.

APIs provide flexible, customisable access to data.

The data you need might only be available through an API.



What are the downsides of using APIs?

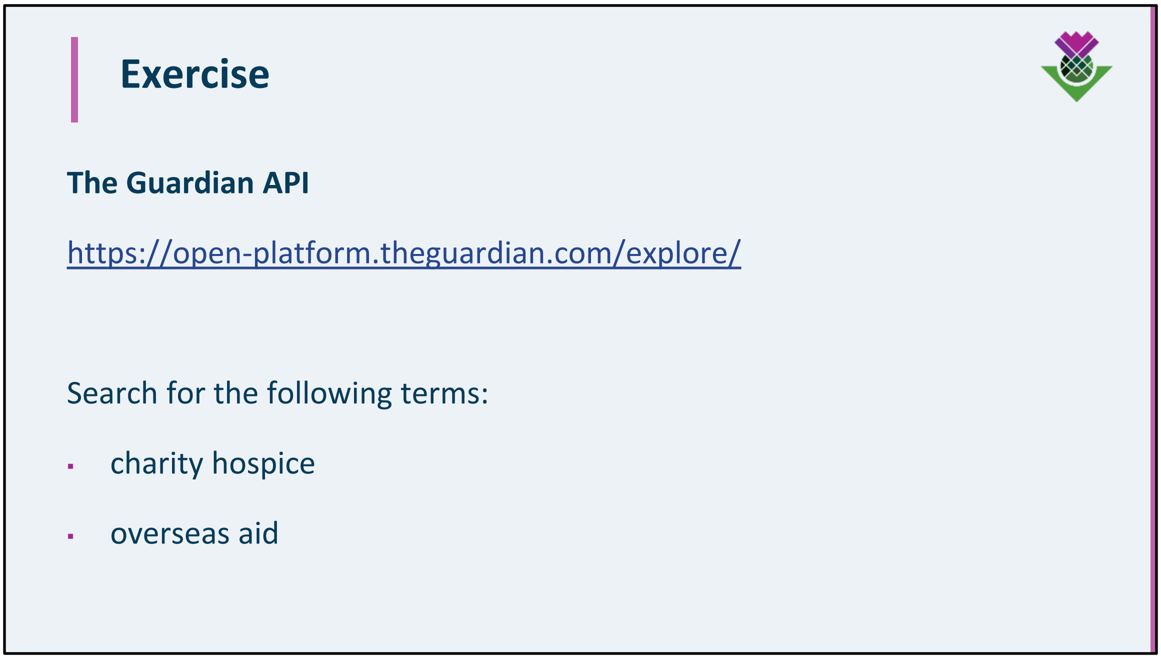
APIs restrict the number of requests for data you can make.

The quality of an API's official documentation can vary wildly.

JSON!

An API is a product and you must comply with the Terms of Service/Use associated with it.

APIs can be updated on a frequent basis, resulting in changes to the rate limit, authentication requirements, endpoints etc.



Exercise

The Guardian API

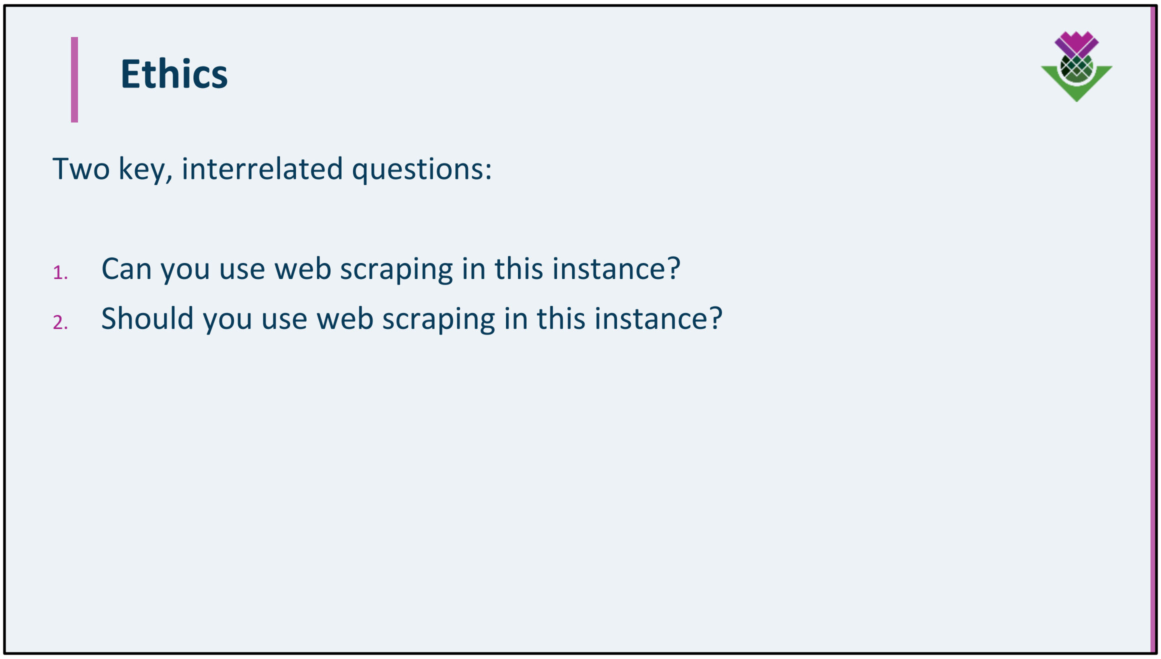
https://open-platform.theguardian.com/explore/

Search for the following terms:

- charity hospice
- overseas aid

Here is the API of the Guardian Newspaper, in particular the exploration tool that anybody can use to make requests to its API.

Get some practice of making requests and seeing the responses that come back. In particular note the structure of the data that are returned, as it may be unfamiliar.



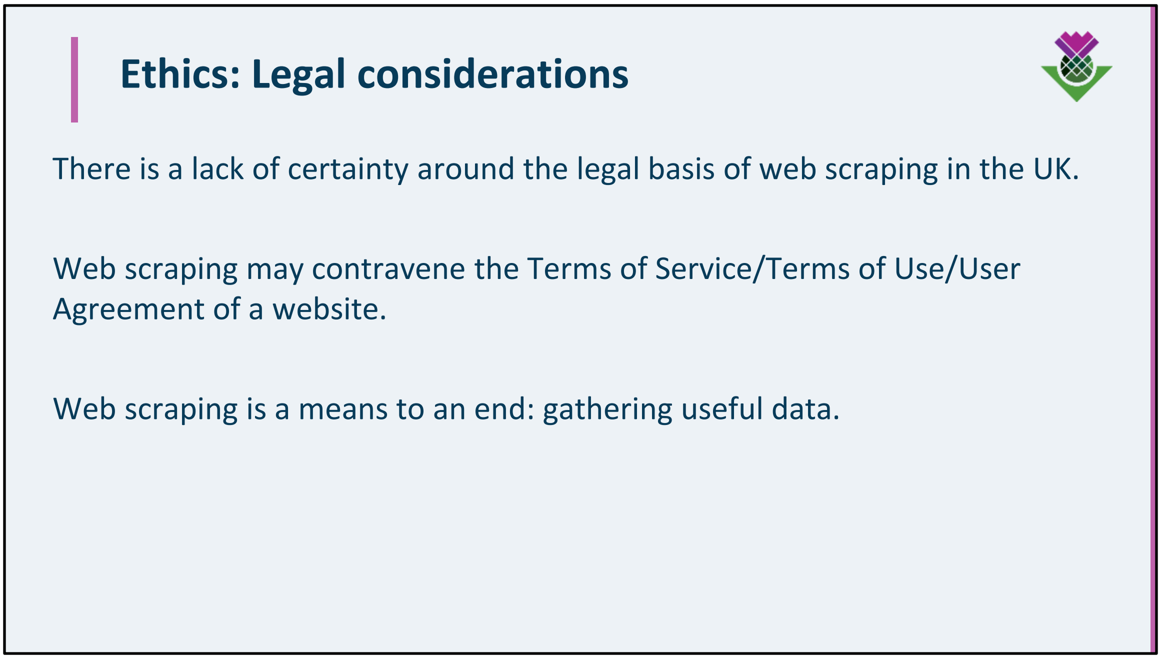
Ethics

Two key, interrelated questions:

1. Can you use web scraping in this instance?
2. Should you use web scraping in this instance?

The legal and ethical dimensions are distinct but interrelated e.g., the terms of service/use of a website may not expressly prohibit web scraping but the information you want may be personal, confidential or sensitive (GDPR) – therefore can you make an argument for collecting AND using the data on the grounds of public or legitimate interest?

We will mainly focus on web-scraping as APIs are designed with sharing and use of the data in mind, and therefore will often clearly state the terms/license under which the data can be accessed and analysed.



Ethics: Legal considerations

There is a lack of certainty around the legal basis of web scraping in the UK.

Web scraping may contravene the Terms of Service/Terms of Use/User Agreement of a website.

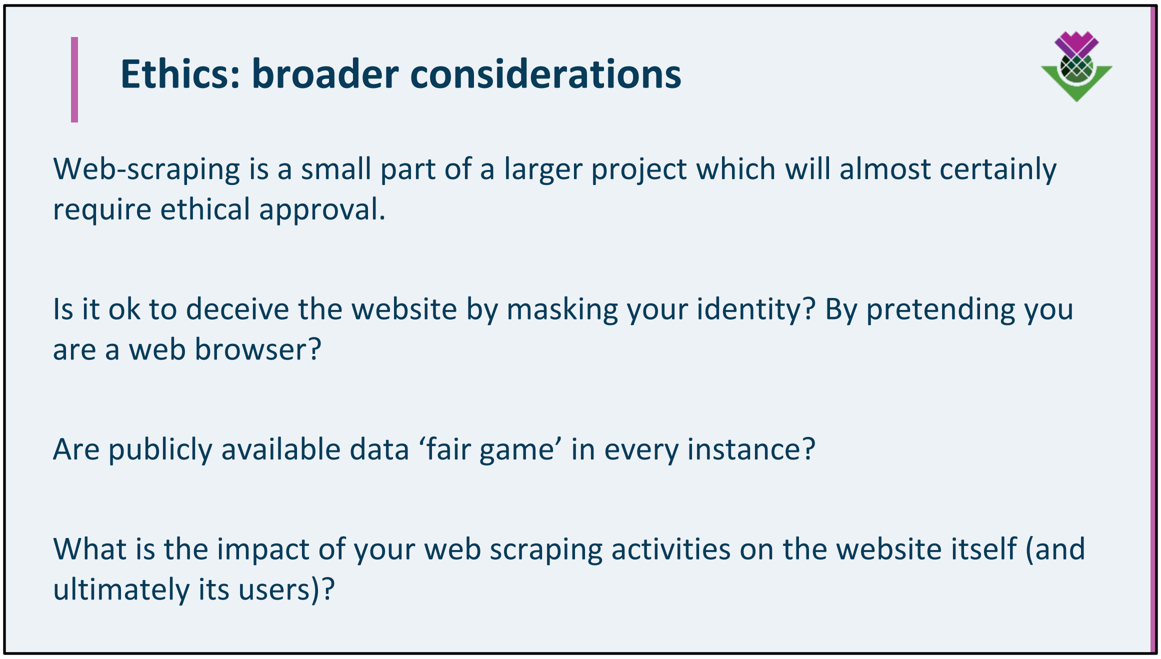
Web scraping is a means to an end: gathering useful data.

That is, can you use web scraping?

Web scraping is a tool, not a criminal act in-and-of itself e.g., nothing illegal about scraping your own website! No specific legislation covering the use of web scraping: legal decisions are often made based on other legislation e.g., data protection, copyright, intellectual or commercial property.

It may come as an unpleasant surprise but many websites have terms and conditions that govern the way you must/should interact with a website. These may not always be enforceable – usually you are banned or blocked - but there is a much stronger case against you if you are required to provide authentication before accessing parts of a website e.g., signing in to LinkedIn and then scraping user profiles.

Even if you have a strong legal basis for engaging in web scraping – no specific legislation and ToS allow it - , you might still have legal obligations with regards to the data you are collecting. Most information on websites is public/open but some may refer to individuals, commercial operations etc. Also, are you using the data for personal, academic or commercial purposes?



Ethics: broader considerations

Web-scraping is a small part of a larger project which will almost certainly require ethical approval.

Is it ok to deceive the website by masking your identity? By pretending you are a web browser?

Are publicly available data ‘fair game’ in every instance?

What is the impact of your web scraping activities on the website itself (and ultimately its users)?

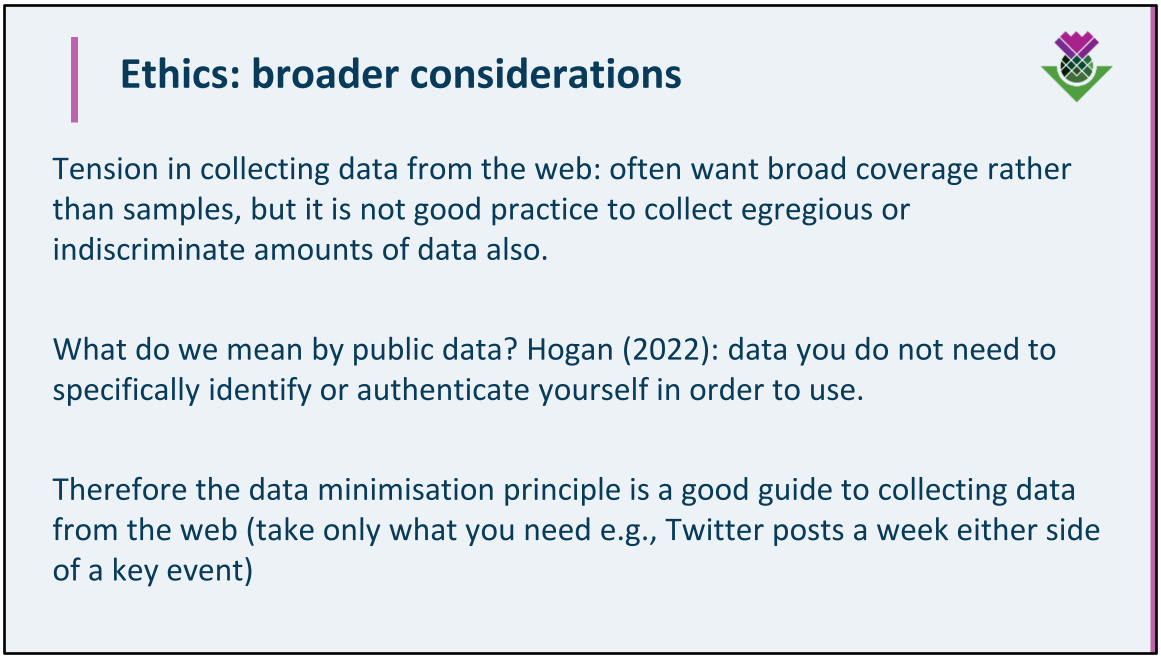
That is, should you use web scraping?

Deception may be legally unproblematic but of some concern to your own ethical code (or even the codes of academic publishers, thesis committees).

Each request you make to a website consumes computational resources, on your end and theirs;

Overload a server by making too many requests, causing the website to crash;

Individuals and organisations may rely on a website for vital and timely information, and causing a website to crash could carry significant real-world implications.

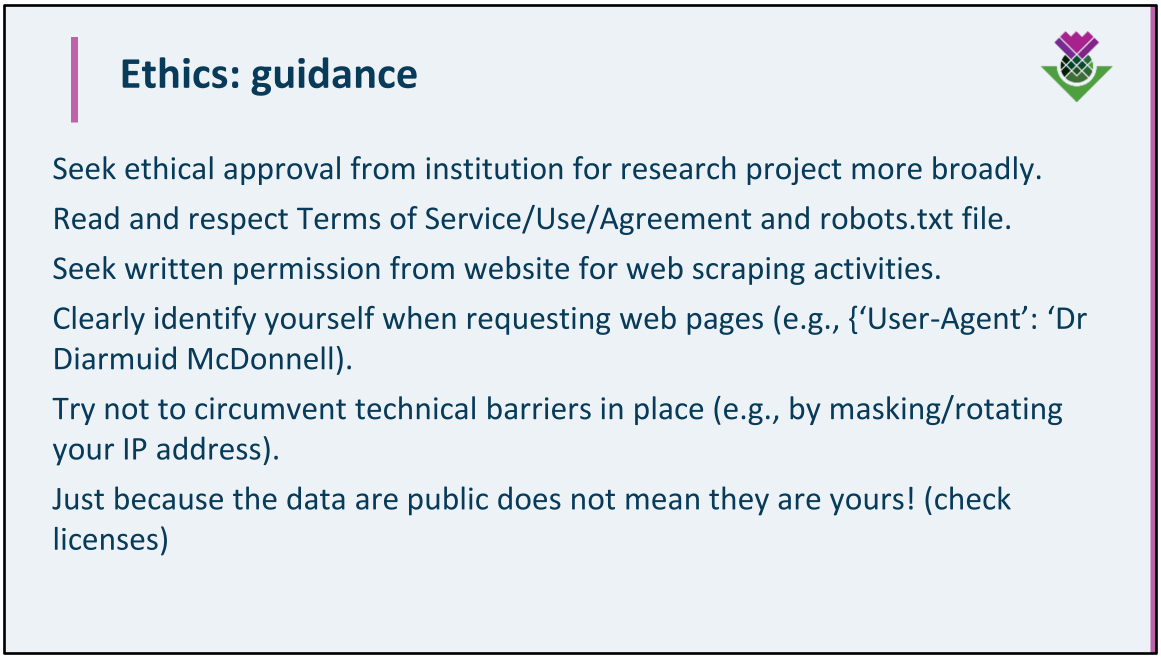


Ethics: broader considerations

Tension in collecting data from the web: often want broad coverage rather than samples, but it is not good practice to collect egregious or indiscriminate amounts of data also.

What do we mean by public data? Hogan (2022): data you do not need to specifically identify or authenticate yourself in order to use.

Therefore the data minimisation principle is a good guide to collecting data from the web (take only what you need e.g., Twitter posts a week either side of a key event)



Ethics: guidance

Seek ethical approval from institution for research project more broadly.

Read and respect Terms of Service/Use/Agreement and robots.txt file.

Seek written permission from website for web scraping activities.

Clearly identify yourself when requesting web pages (e.g., {‘User-Agent’: ‘Dr Diarmuid McDonnell).

Try not to circumvent technical barriers in place (e.g., by masking/rotating your IP address).

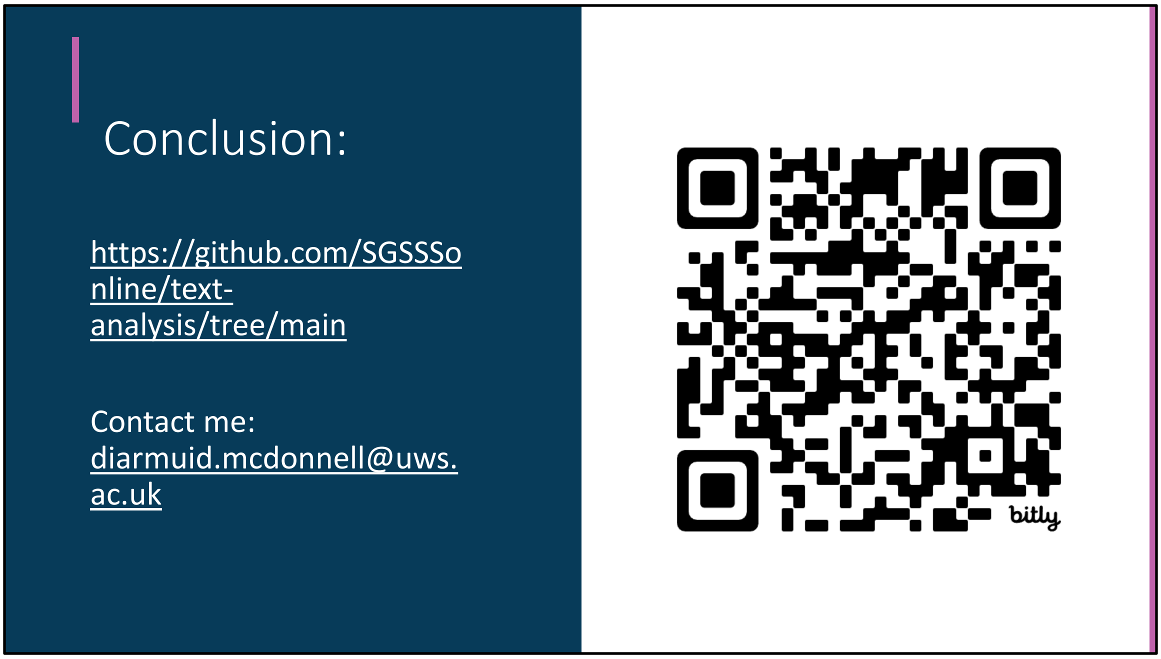
Just because the data are public does not mean they are yours! (check licenses)

Exercise

You are interested in studying the discursive (language and arguments) elements of fundraising campaigns. Your colleague suggested collecting data from JustGiving, a well-known fundraising platform.

https://www.justgiving.com/crowdfunding

1. Select a fundraising topic
2. View a sample of fundraising pages
3. Devise a plan for collecting fundraising campaign details from these pages:
   - Which details (story, donors, £)?
   - Are there ethical issues? Can you and/or should you collect data?



Conclusion:

https://github.com/SGSSSonline/text-analysis/tree/main

Contact me: diarmuid.mcdonnell@uws.ac.uk