Scottish Graduate School of Social Science

Sgoil Cheumnaichean Saidheans Sòisealta na h-Alba

Collecting Digital Data

The Role of Web-scraping and APIs

Dr Diarmuid McDonnell

Braw Data Ltd · Gradel Institute of Charity, University of Oxford

24 February 2026

About the Instructor

Dr Diarmuid McDonnell

- Director, **Braw Data Ltd**
- Visiting Fellow, **Gradel Institute of Charity**, University of Oxford
- Research: geographic distribution of civil society activity
- Background in quantitative social science and data science

https://www.brawdata.co.uk
https://www.gradelinstituteofcharity.co.uk/diarmuid-mcdonnell

Course Outline

Time	Session	Format
10:00–10:30	Welcome & How the Web Works	Lecture
10:30–11:15	Practical 1: Web Scraping	Colab notebook
11:15–11:30	Break	
11:30–11:45	What Are APIs?	Lecture
11:45–12:45	Practical 2: UK Police API	Colab notebook
12:45–13:30	Lunch	
13:30–13:45	API Landscape Survey	Lecture
13:45–14:45	Practical 3: API Challenge	Colab notebook
14:45–15:00	Break	
15:00–15:15	LLMs as Coding Assistants	Lecture
15:15–15:50	Practical 4: LLM Showdown	Colab notebook
15:50–16:00	Wrap-up & Q&A	

Getting Set Up

What you need

- A **Google account** (for Google Colab)
- A web browser (Chrome recommended)

How to access the notebooks

1. Go to the course **README** (link shared on Teams/email)
2. Click the **Open in Colab** badge for your chosen language
3. Sign in with your Google account
4. You're ready to go!

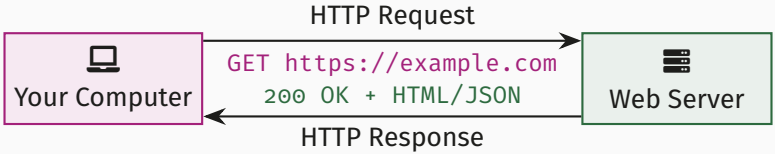
ℹ No software installation required — everything runs in the browser.

Why Collect Digital Data?

- **Government open data** — policy evaluation, freedom of information, public spending
- **Social media** — opinion analysis, political discourse, misinformation
- **Organisational websites** — third-sector research, corporate governance, labour markets
- **Administrative records** — health, education, justice system data via APIs

The web is a vast, continuously updated source of social science data — but accessing it requires the right tools and techniques.

How the Web Works

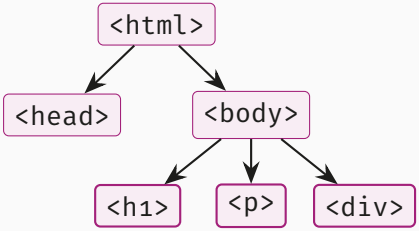


Common HTTP status codes:

200 *OK* · 301 *Moved Permanently* · 302 *Found (Redirect)* · 400 *Bad Request* · 401 *Unauthorised* · 403 *Forbidden* · 404 *Not Found* · 429 *Too Many Requests* · 500 *Server Error*

HTML Basics

HTML tree structure



Tags define structure and meaning

Example HTML

```
<html>
  <head>
    <title>My Page</title>
  </head>
  <body>
    <h1>Welcome</h1>
    <p>Hello, world!</p>
  </body>
</html>
```

Structured vs Unstructured Data

Unstructured (HTML)

```
<div class="person">
  <h2>Jane Smith</h2>
  <p>Age: 34</p>
  <p>City: Edinburgh</p>
</div>
```

- Designed for **human** reading
- Structure can change without notice
- Requires parsing to extract data

Structured (JSON)

```
{
  "name": "Jane Smith",
  "age": 34,
  "city": "Edinburgh"
}
```

- Designed for **machine** reading
- Consistent, documented format
- Easy to convert to tables

Ethics and Legality

- **Check robots.txt** — websites specify what can be crawled
- **Read Terms of Service** — some sites explicitly prohibit scraping
- **Respect rate limits** — don't overwhelm servers with requests
- **Consider personal data** — GDPR applies to identifiable information
- **Prefer APIs when available** — structured, legal, reliable
- **Web scraping as last resort** — use only when no API exists

Note

We will discuss ethical and legal considerations throughout the day as they arise in each practical.

Next Up

Practical 1: Web Scraping

- Extract text from a simple web page
- Scrape data across multiple pages
- Save structured data to a file

Open the **Practical 1** notebook in Google Colab
(Python or R — your choice!)