



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

LEGAL & ETHICAL ISSUES IN WEB SCRAPING

14 OCT 2024

Dr Jessica Witte



**SUPPORT FOR DATA-LED
AND APPLIED DIGITAL
RESEARCH ACROSS THE
ARTS, HUMANITIES AND
SOCIAL SCIENCES.**





ABOUT ME

- CDCS Digital Research Analyst
- Digital methods: web scraping, text analysis, natural language processing, machine learning, data visualisation, generative AI
- Python, R, HTML, CSS
- 'Self-taught'





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

OVERVIEW

14:00-14:50

- Introduction & housekeeping
- Discussion
- Overview of UoE ethics policies
- Web scraping basics and Terms of Service
- APIs and the 'post-API era' in research
- Group activity & discussion

Comfort break

15:00-15:50

- Data privacy, data ethics, and the law
- Group activity & discussion
- Challenges and 'grey areas'
- Conclusion



www.cdcs.ed.ac.uk



What brought you to the course today?

Have you ever faced (potential) ethical or legal challenges in acquiring data? If so, what happened?

DISCUSSION





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

UNIVERSITY RESEARCH ETHICS POLICY



- ‘Safeguard the interests and well-being’ of all involved with or impacted by a research project
- Ethical research principles:
 - Beneficence and non-maleficence
 - Integrity, openness and transparency
 - Dignity and respect
 - Responsibility and accountability
 - Equality, diversity and inclusion
- Must be integrated into all stages of a project’s design, including its impact



www.cdcs.ed.ac.uk



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



UNIVERSITY RESEARCH MISCONDUCT POLICY

Research misconduct includes:

- Misuse of personal data
- Lack of informed consent from participants
- Breach of confidentiality
- Noncompliance with legal and ethical requirements
- Breach of duty of care
 - Disclosing participants' identity
 - Exposing personal or sensitive data
 - Improper conduct



www.cdcs.ed.ac.uk

WHY SCRAPE THE WEB?

- To collect data—social media, public records, government data
- To expand, update, or complete datasets
- To examine public discourse about a topic
- To analyse the relationship between online and offline behaviour



KEY TERMS

- **Web crawling:** automated process of systematically browsing the Internet using a **spider**, a programme that ‘crawls’ from webpage to webpage
 - Indexing web pages for search engines (Google search results)
 - Saving web pages in real time (the [Wayback Machine](#))
- **Web scraping:** extracting specific data from webpages through HTML parsing using a script called a **scraper**
- **Browser automation:** a script that interacts with web content, such as clicking on links or downloading content
- **Rate limit:** a cap set on certain interactions with web content designed to manage web traffic; some sites have these limits listed in a ‘robots.txt’ file

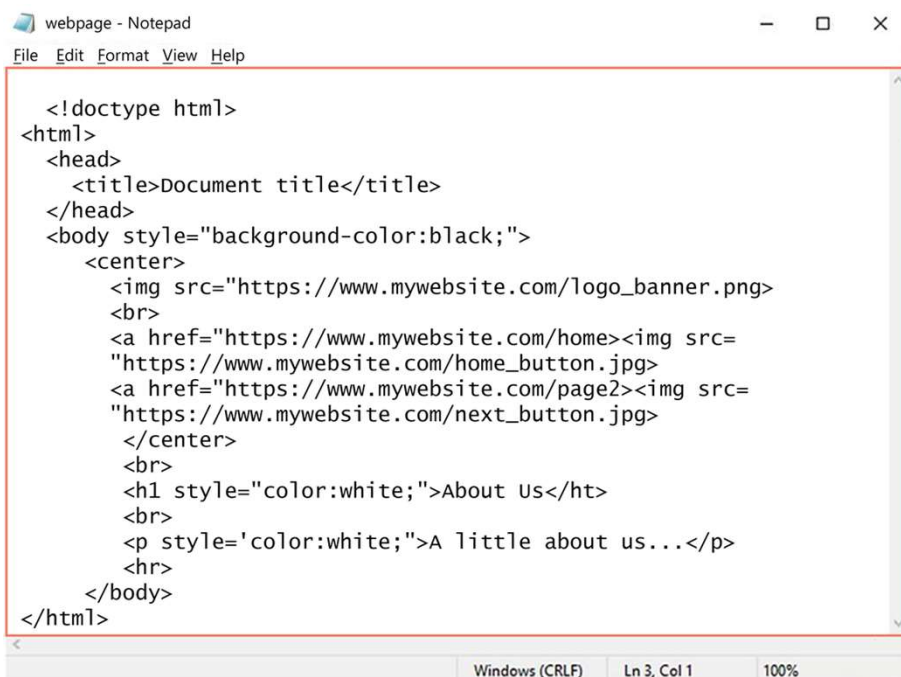


STATIC & DYNAMIC WEBPAGES

- Static pages:
 - Typically small, displaying a limited amount of content
 - Look the same for all users
 - Built in scripting languages (e.g. HTML, CSS, JavaScript)
 - Blogs, GitHub Pages sites, CV page on personal website
- Dynamic pages:
 - Can change based on users' data, device, and behaviour
 - Contain interactive content
 - Built in content management systems using a combination of scripting and server-side languages
 - Vulnerable to more security risks
 - Amazon, Netflix, BBC News



HTML



```
<!doctype html>
<html>
<head>
<title>Document title</title>
</head>
<body style="background-color:black;">
<center>

<br>
<a href="https://www.mywebsite.com/home"><img src=
"https://www.mywebsite.com/home_button.jpg">
<a href="https://www.mywebsite.com/page2"><img src=
"https://www.mywebsite.com/next_button.jpg">
</center>
<br>
<h1 style="color:white;">About Us</ht>
<br>
<p style='color:white;'>A little about us...</p>
<hr>
</body>
</html>
```

[Image credit](#)

- Standard markup language creating the structure and content of a static webpage
- HTML **documents** are built of **paired HTML tags**
`<html>`
`</html>`
- **Elements** are objects or features such as images and text
- HTML **attributes** modify the appearance or behaviour of elements





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

METHODS FOR WEB SCRAPING

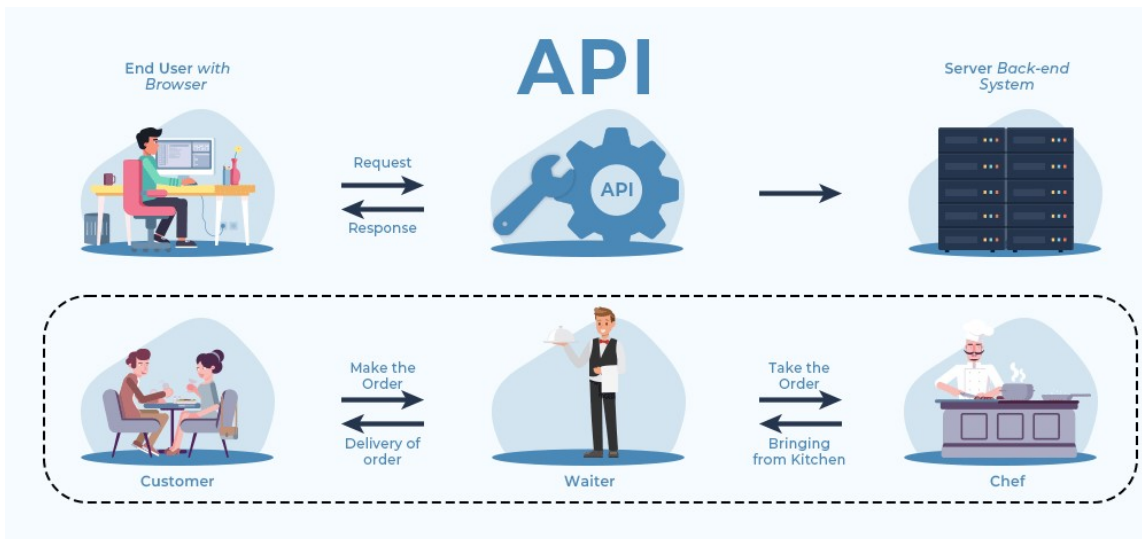
- **Scraping and crawling HTML/XML**—generate a list of URLs from which to extract information, selecting relevant sections by HTML, and downloading content
- **APIs**—request data from sites on their own terms. Standard approach for social media data pre-2023, but has become more complicated
- **Browser automation**—using a script to click through a website like a human user and download content page by page



www.cdcs.ed.ac.uk

APPLICATION PROGRAMMING INTERFACE (API)

- Connects a computer to another computer, server, or database
- Often purpose-built or platform-specific
- Can be free or paid; often rate-limited
- Allow for user-friendly batch data collection in compliance with platform terms of service (ToS)



[Image credit](#)

THE 'POST-API ERA'

- Extracting social media data using platforms' APIs has become standard practice in many fields
- After the 2018 Cambridge Analytica scandal, Facebook closed its API
- In 2023, Twitter and Reddit removed free API access for researchers
- A consensus has yet to be reached on how to ethically acquire data from these platforms
 - Paying for data
 - Finding a new source of data
 - Collecting data through web scraping



2023 Reddit API controversy

Article Talk

From Wikipedia, the free encyclopedia

In April 2023, the [discussion](#) and [news aggregation](#) website [Reddit](#) announced its intentions to charge for its [application programming interface](#) (API), a feature which had been free since 2008, causing a dispute. The move forced multiple third-party applications to shut down and threatened accessibility applications and moderation tools.

TERMS OF SERVICE (TOS)

- Also called 'Terms of Use'
- The '**fine print**' users agree to when accessing a particular platform or site
- Most people accept the ToS without reading them
- **Violating ToS can lead to a range of consequences**
 - rate limiting
 - account deletion
 - possible legal action
 - for researchers, a breach of ethics





ACTIVITY : MAKING SENSE OF TOS

Table 1: WhatsApp

Table 2: GoodReads

Table 3: Amazon

Table 4: Twitter/X

Table 5: Wikipedia

**ToS are notoriously difficult to read. Give it a go on your own, but if you're really stuck, the 'Terms of Service; Didn't Read' project is a good resource: <https://tosdr.org/>*

With your table, you'll be investigating a popular platform's Terms of Service/Terms of Use agreement. First, find the site's Terms of Use/Terms of Service page. Then, find the answers to the following questions in the ToS.*

1. **Data extraction**—What are the platform's policies on web scraping? Does the platform have an API? If not, how can users acquire its data?
2. **Data collection**—What information can the platform collect from users, and what can they do with it? Who else can see users' data?
3. **Data storage & deletion**—How long does the platform keep personal data? If users delete their account, what happens?



IS WEB SCRAPING LEGAL?

- **Existing policies have grey areas** and are incomplete for some use cases
- Best practices differ by country, discipline, sector, and interpretation of laws and policies
- **Illegal web scraping:**
 - Collects sensitive or private data
 - Violates copyright or intellectual property law
 - Extracts personally identifying information
 - Sells data collected through web scraping is illegal
- **Legal web scraping:**
 - Collects public or openly available information following a site's policies
 - Where possible, uses a platform's approved infrastructure (e.g. an API)
 - Takes care to remove potentially identifying data
 - Has prior approval from the School's/University's ethics review



PUBLIC & PRIVATE DATA

- Ethics guidelines often **distinguish between *public* and *private* sources of data** when it comes to determining if web scraping is appropriate
- This often depends on **users' expectations about who will see their content**
- However, there are still grey areas:
 - A public group could go private (or vice-versa)
 - Researchers could receive informed consent to collect data from individuals or communities
 - A user could delete their content—or their account
 - A public platform could institute a paywall
 - A platform could change its ToS to prohibit/permit data collection
 - A platform could close or restrict its API



PUBLIC & PRIVATE DATA

Decide whether the following data sources are public or private:

1. Gov.uk news
2. Posts in a Twitter thread about mental health from those experiencing depression
3. A private Facebook group for first-year students attending the University of Edinburgh
4. A WhatsApp group for Spanish speakers in London
5. A Teams group for postgraduate students in CAHSS



STORING & WORKING WITH DATA

- Use a secure cloud environment within the University's IT infrastructure
- Gather data only as required for a specific purpose
- Keep data only as long as is necessary for completing a project
- Collect data without violating legal frameworks or ethical policies
- When working with participant data:
 - Receive informed consent from all participants before collecting, storing, or analysing data
 - Anonymise data and remove personally identifying information



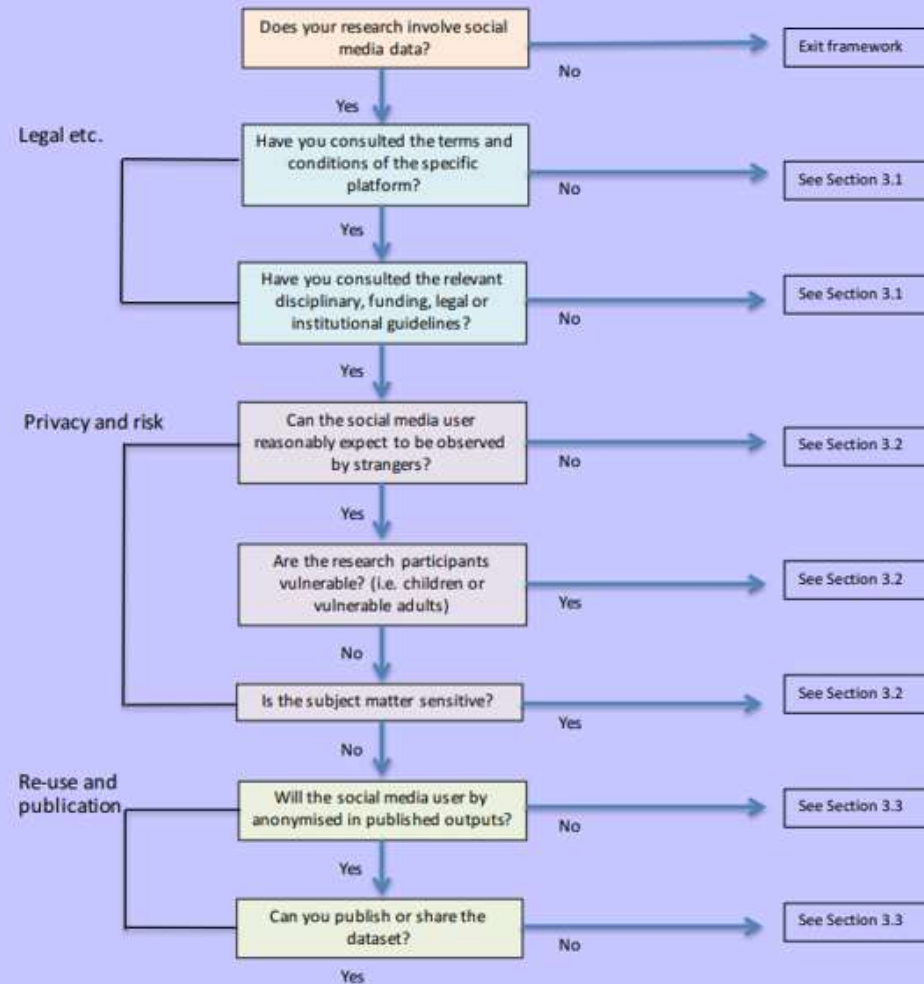
ANONYMISING DATA

- Datasets acquired through platforms' APIs do not violate ToS, but they do not necessarily comply with ethics guidelines
- Remove personally identifying information
 - Name, age, gender
 - Username and bio
- In outputs such as presentations and papers, discuss data at the aggregate level
- Consider the potential for reverse searching when quoting
- Consider whether users would expect their data to be public or private



3 Framework for ethical research with social media data

Social Media Ethics Framework:





OPEN SCIENCE & FAIR DATA

- Transparency, reproducibility and sharing in research practice and dissemination
- Open-source data, software, resources, hardware, publications
- Engaging the public through crowdsourcing, citizen science, crowdfunding, and collaboration
- Inclusion of groups and communities that have been historically marginalised

Data should be:

- **F**indable
- **A**ccessible
- **I**nteroperable, or compatible with other data, systems, and technologies
- **R**eusable for other purposes



THE DANISH OKCUPID STUDY

- In 2016, Danish researchers analysing the dating site OkCupid shared data on the Open Science Framework
- Their dataset included users' personal information such as their names, ages, gender, and responses to questions
- Open science values include transparency and sharing to ensure reproducibility in research design
- However, open research is not always ethical (and vice-versa)

Researchers Caused an
Uproar By Publishing Data
From 70,000 OkCupid
Users

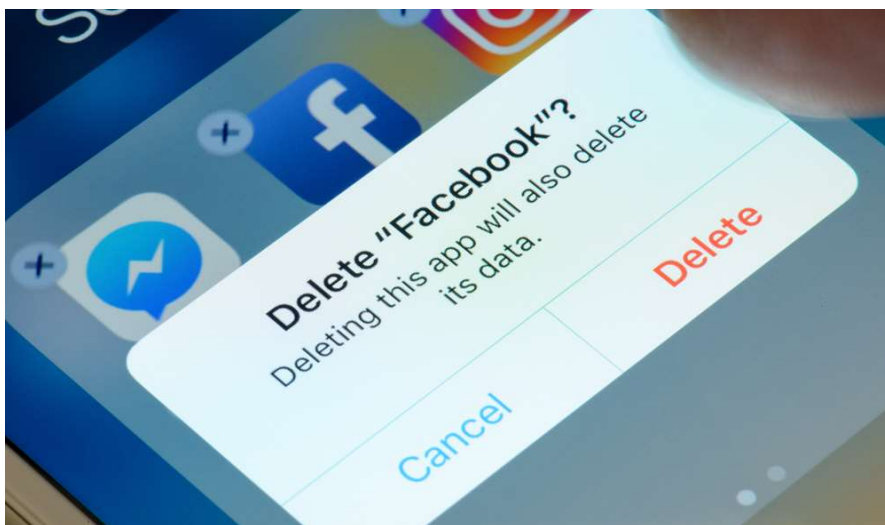


OkCupid Study Reveals the Perils of Big-Data Science

The data of 70,000 OKCupid users is now searchable in a database. Ethicist Michael T Zimmer explains why it doesn't matter that it was "already public."



THE CAMBRIDGE ANALYTICA SCANDAL



- In the 2010s, the consulting firm Cambridge Analytica collected personal data from 87 million Facebook profiles without users' knowledge or consent
- Created psychological profiles of users based on aspects of their personal data such as their location, gender, interests, and age
- Political campaigns in the US used the data to show targeted ads based on these profiles
- In 2018, a whistleblower reported the data breach
- Investigation led to Facebook restricting its API and third-party data access



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

IS WEB SCRAPING ETHICAL?

- It depends
- Ensure that the methods for data acquisition are legal and comply with relevant School/funder policies
 - ToS compliance
 - Public/private data
 - Informed consent
- What are the potential consequences of acquiring the data?
- What are the potential consequences of *not* acquiring the data?



www.cdcs.ed.ac.uk



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



WHAT TO CONSIDER BEFORE YOU WEB SCRAPE

- Data
 - What, where, why, and how much?
 - Does it contain personal or sensitive information?
 - Who made it? Who owns it?
- Platform
 - ToS and other regulations
 - Have other researchers scraped its data? How?
- National and local laws
- Other policies—your School's and University's; funder's (if applicable)



www.cdcs.ed.ac.uk



ACTIVITY 2: MAKING DATA ETHICS DECISIONS

You and your table are reviewing data collection methods for projects awaiting ethics approval. Read each example and discuss whether the proposed method(s) comply with the principles of research & data ethics. If not, how might the methodology be improved?



1. A professor is examining the role of emotion in how readers form opinions about books. She plans to analyse book review data scraped from Goodreads using the platform's API.
2. A team of researchers are mapping the popularity of cycling in different areas of Edinburgh. They have received permission from the moderator of a private Facebook group for Edinburgh residents to extract all posts and comments contributed since 2019.
3. A lecturer has received UKRI funding to analyse Twitter posts about climate change. He has budgeted £25k for a six-month subscription to Twitter's API, which he will use to collect the data.
4. A postdoc is tracking health outcomes in people diagnosed with COVID-19. 1,000 patients have agreed to donate their health records, which the postdoc will anonymise and analyse at the aggregate level.





THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



**THANKS FOR YOUR
ATTENTION**

Q & A



www.cdcs.ed.ac.uk