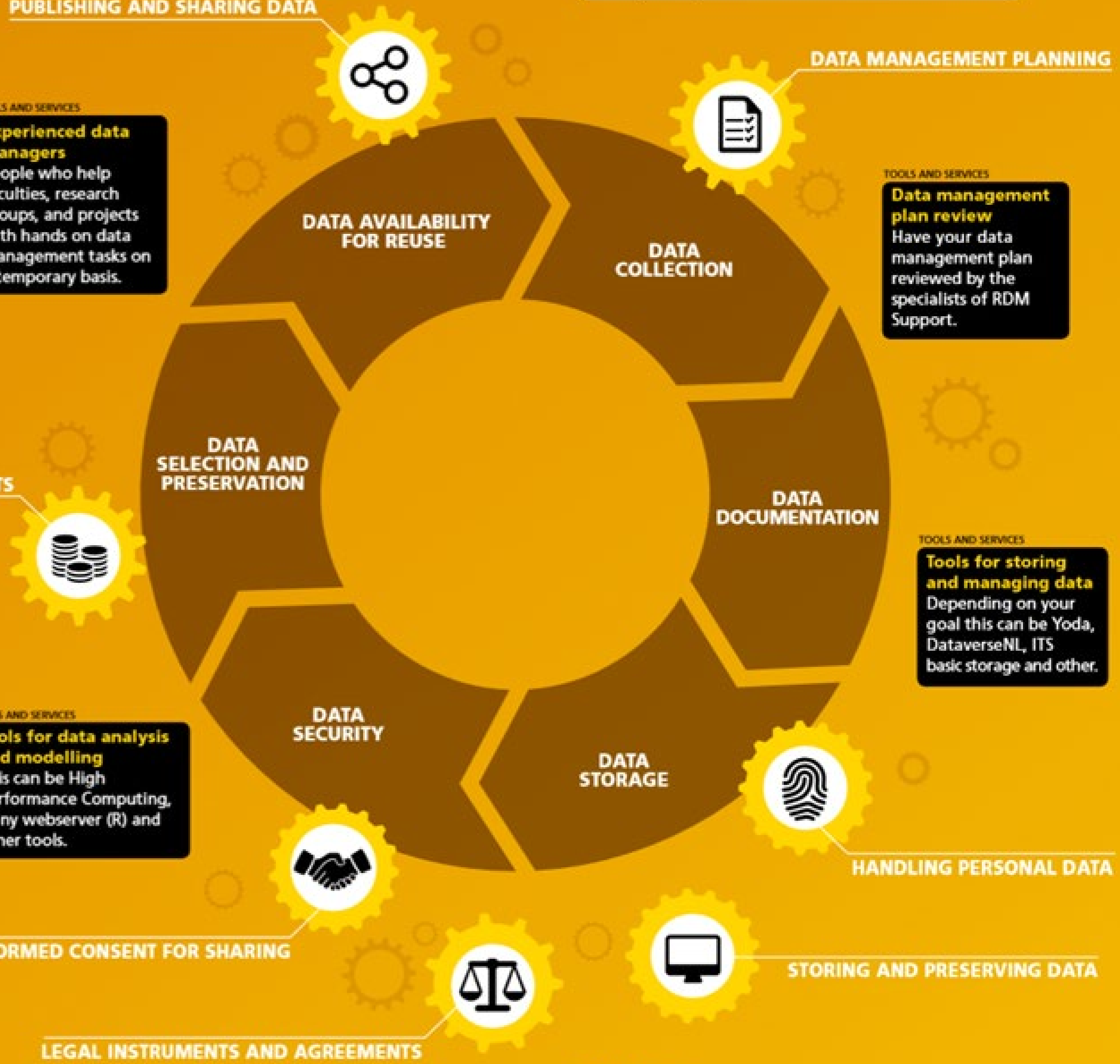
Web scraping

Sander Prins
Technical Support Assistant

Why web scraping?



DATA COLLECTION

DATA DOCUMENTATION

DATA STORAGE

DATA SECURITY

DATA SELECTION AND PRESERVATION

DATA AVAILABILITY FOR REUSE

DATA MANAGEMENT PLANNING

HANDLING PERSONAL DATA

STORING AND PRESERVING DATA

LEGAL INSTRUMENTS AND AGREEMENTS

INFORMED CONSENT FOR SHARING

COSTS

PUBLISHING AND SHARING DATA

TOOLS AND SERVICES

DMPonline to write your Data Management Plan
You can write, share and download your data management plan online.

TOOLS AND SERVICES

Data management plan review
Have your data management plan reviewed by the specialists of RDM Support.

TOOLS AND SERVICES

Tools for storing and managing data
Depending on your goal this can be Yoda, DataverseNL, ITS basic storage and other.

TOOLS AND SERVICES

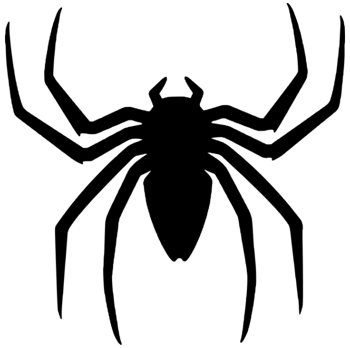
Training and workshops
RDM Support offers a learning programme with an online training ‘Learn to write your DMP’, basic, in-depth and tailor made workshops and more.

TOOLS AND SERVICES

Tools for data analysis and modelling
This can be High Performance Computing, shiny webserver (R) and other tools.

TOOLS AND SERVICES

Experienced data managers
People who help faculties, research groups, and projects with hands on data management tasks on a temporary basis.



Research data from the web

- Contact website (owner) for data request
- (Scientific) data publications
- Application Programming Interface (API)
- Web scraping
- ~~Data brokers~~

Social media

- Software suite created by Digital Methods Initiative (University of Amsterdam)
- GUI for interfacing with APIs and doing quick data analysis

What is web scraping?

Web scraping

- Extracting data from websites
  - Reverse engineering
- Parsing
  - HTML to preferred data format

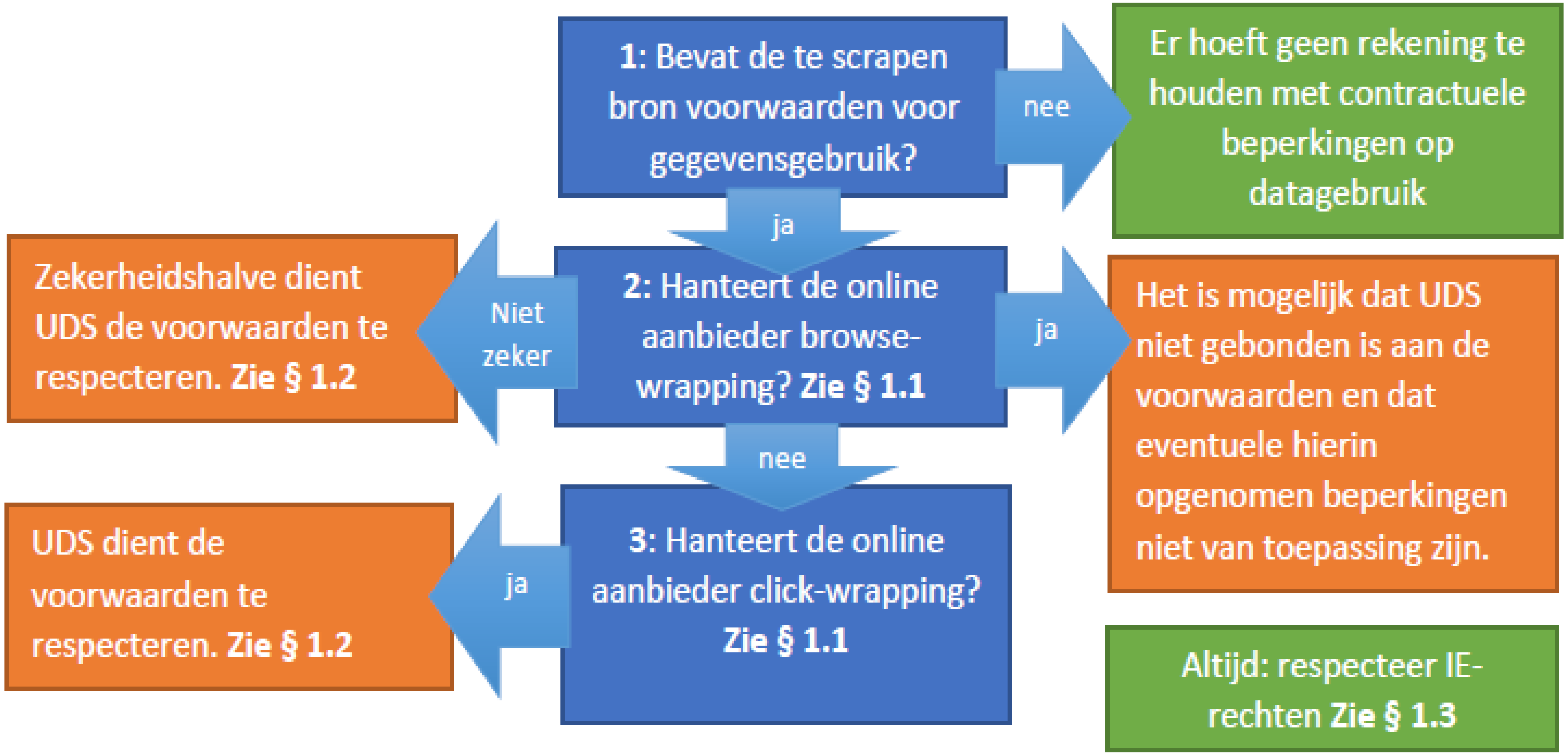
Ethical Considerations

Considerations

- Robots.txt: Understanding website terms of service.
- Platform policies and agreements
- Rate limiting: avoiding overwhelming website servers.

Auteur: Joost Gerritsen (Legal Beetle). Bestemd voor: Utrecht Data School (UDS)
Onderwerp: methodiek t.b.v. rechtmatige scrapingactiviteiten voor UDS-onderzoek (v. 1.0)

Beslisboom #1: Websitevoorwaarden



1: Bevat de te scrapen bron voorwaarden voor gegevensgebruik?

nee

Er hoeft geen rekening te houden met contractuele beperkingen op datagebruik

ja

2: Hanteert de online aanbieder browse-wrapping? **Zie § 1.1**

Niet zeker

Zekerheidshalve dient UDS de voorwaarden te respecteren. **Zie § 1.2**

ja

Het is mogelijk dat UDS niet gebonden is aan de voorwaarden en dat eventuele hierin opgenomen beperkingen niet van toepassing zijn.

nee

3: Hanteert de online aanbieder click-wrapping? **Zie § 1.1**

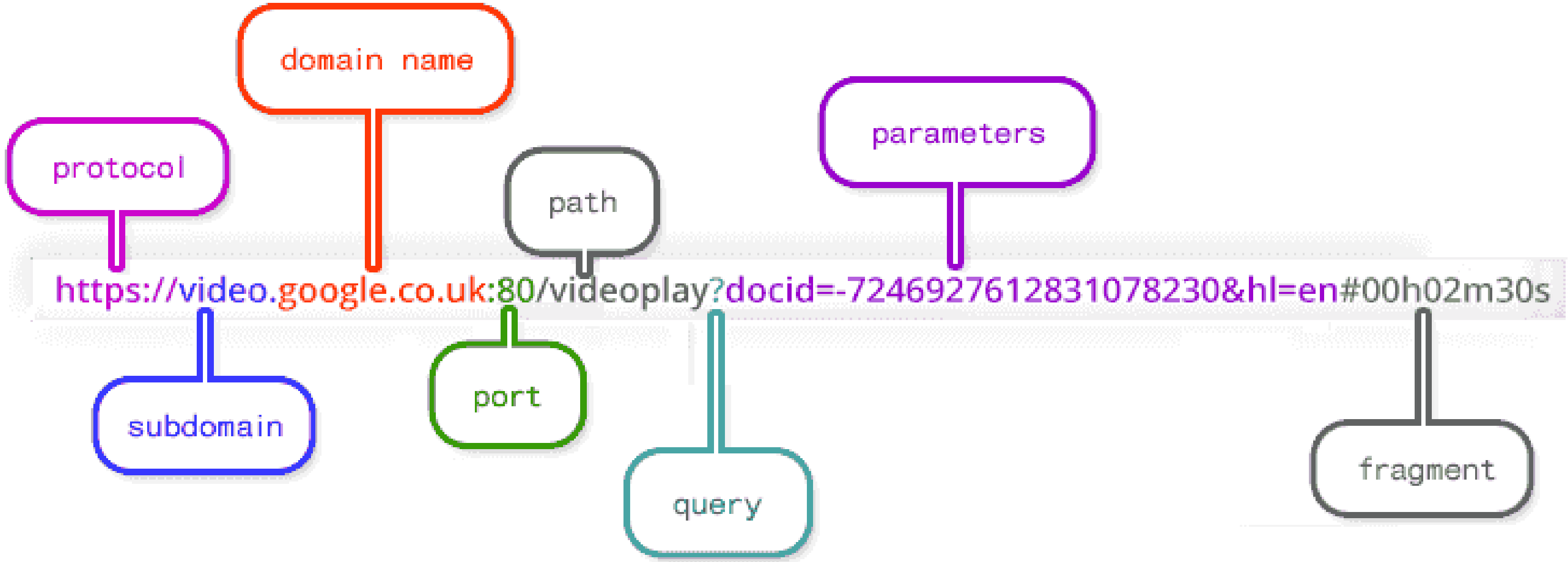
ja

UDS dient de voorwaarden te respecteren. **Zie § 1.2**

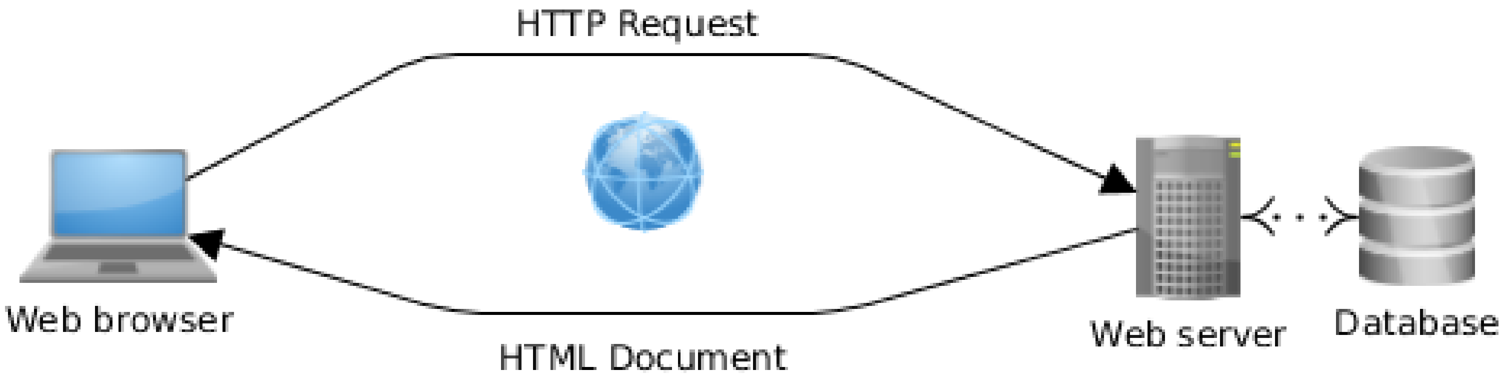
Altijd: respecteer IE-rechten **Zie § 1.3**

Reverse Engineering

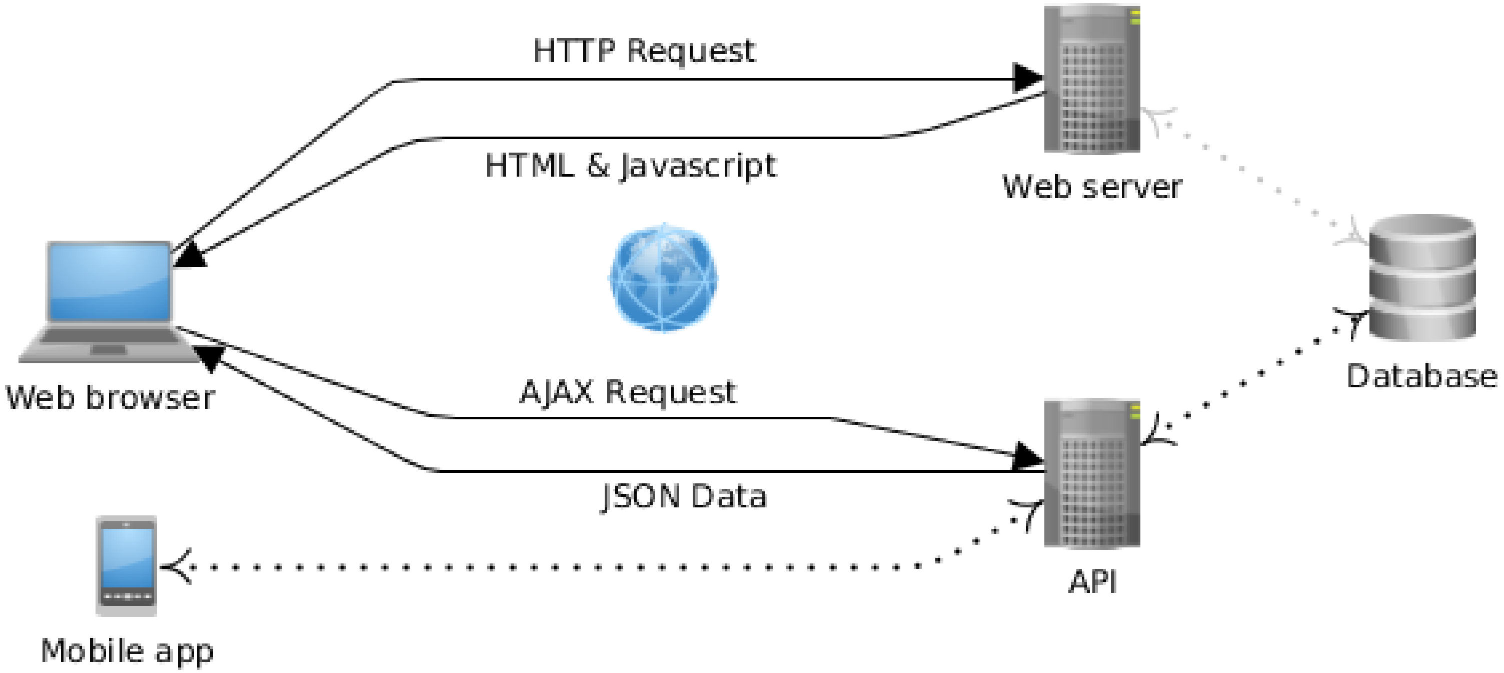
Uniform Resource Locator (URL)


domain name
protocol
path
parameters
https://video.google.co.uk:80/videoplay?docid=-7246927612831078230&hl=en#00h02m30s
subdomain
port
query
fragment

Traditional web applications



Modern web applications

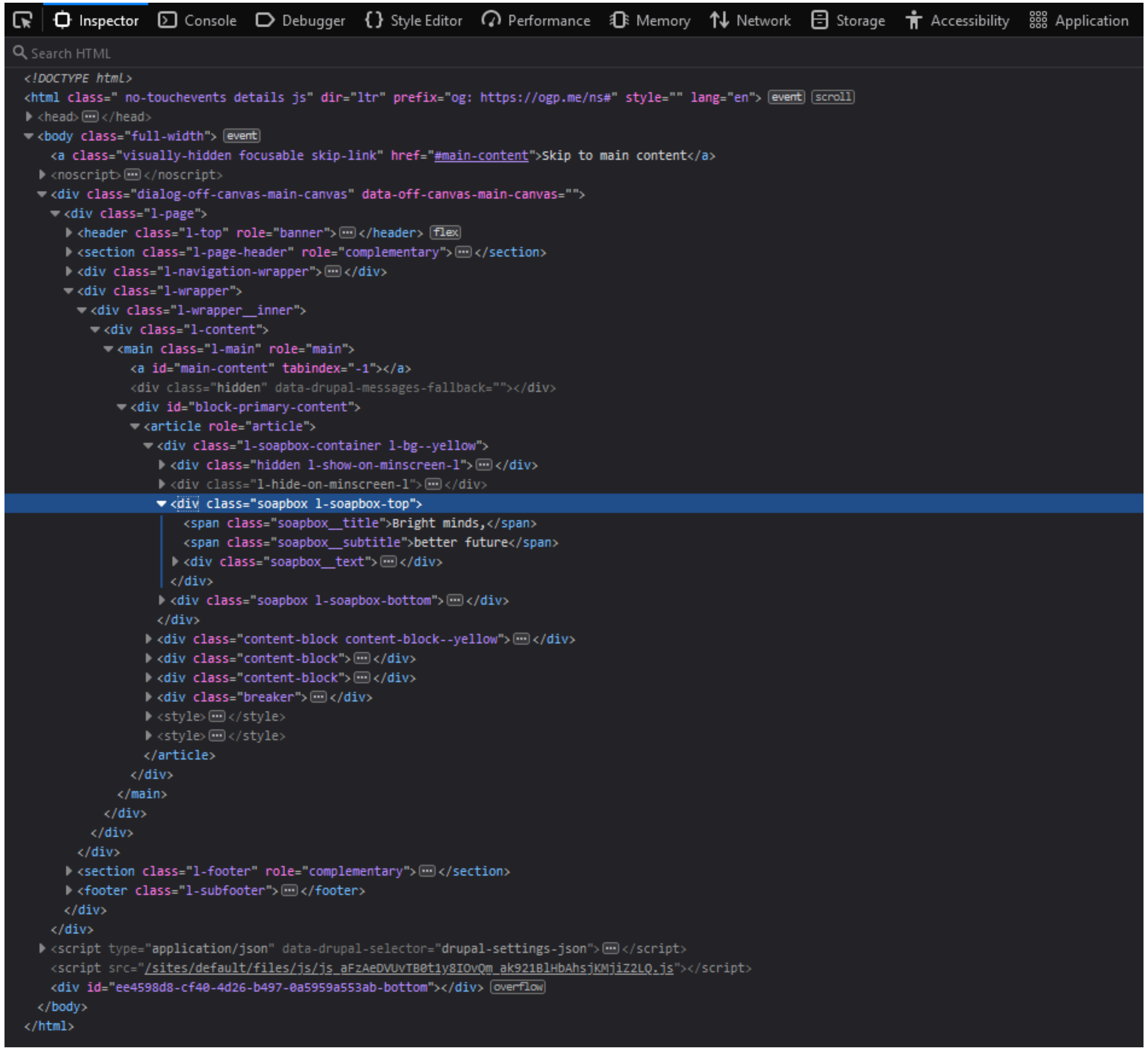


Developer Tools

Open *Inspector* by pressing F12

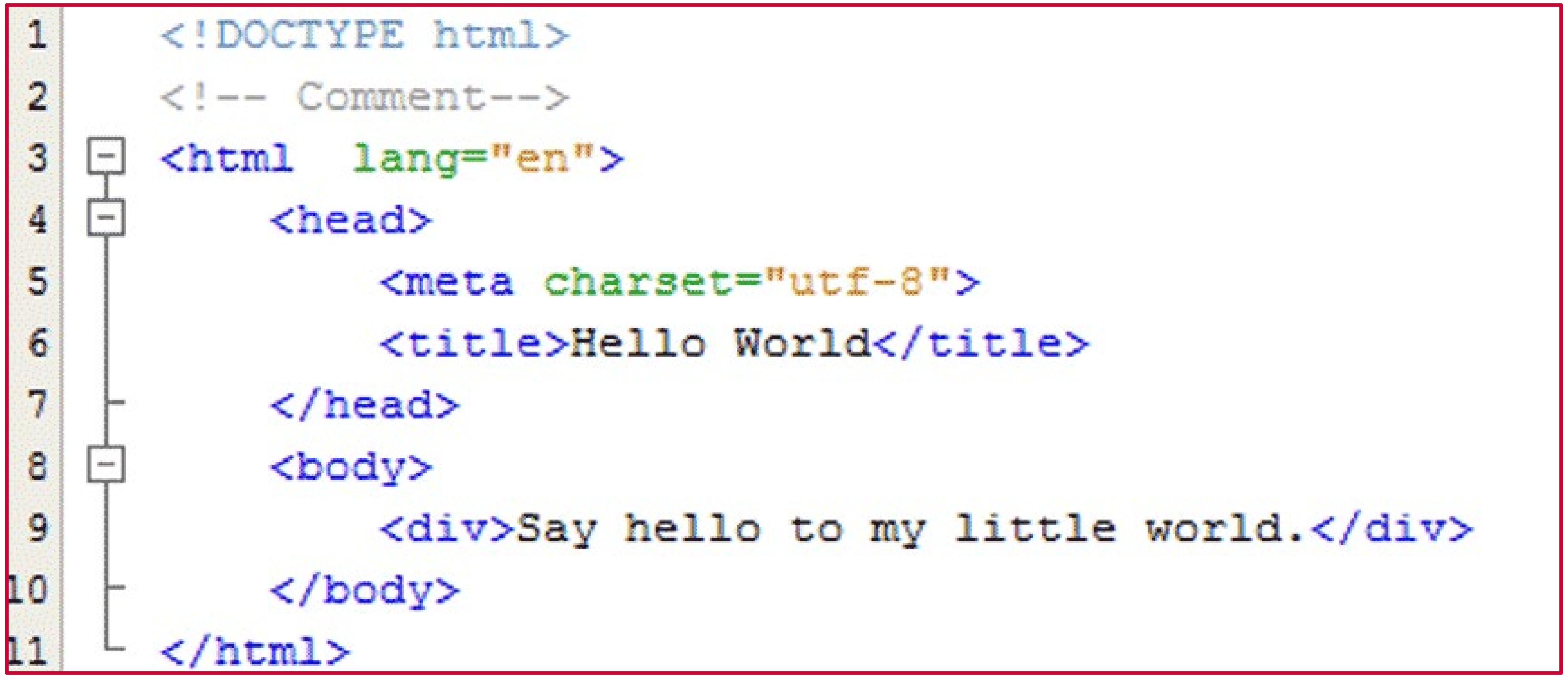
or

Right click on the element that you want to explore and click *Inspect element*



```
<!DOCTYPE html>
<html class=" no-touchevents details js" dir="ltr" prefix="og: https://ogp.me/ns#" style="" lang="en">
  <head>⋯</head>
  <body class="full-width">
    <a class="visually-hidden focusable skip-link" href="#main-content">Skip to main content</a>
    <noscript>⋯</noscript>
    <div class="dialog-off-canvas-main-canvas" data-off-canvas-main-canvas="">
      <div class="l-page">
        <header class="l-top" role="banner">⋯</header>
        <section class="l-page-header" role="complementary">⋯</section>
        <div class="l-navigation-wrapper">⋯</div>
        <div class="l-wrapper">
          <div class="l-wrapper__inner">
            <div class="l-content">
              <main class="l-main" role="main">
                <a id="main-content" tabindex="-1"></a>
                <div class="hidden" data-drupal-messages-fallback=""></div>
                <div id="block-primary-content">
                  <article role="article">
                    <div class="l-soapbox-container l-bg--yellow">
                      <div class="hidden l-show-on-minscreen-l">⋯</div>
                      <div class="l-hide-on-minscreen-l">⋯</div>
                      <div class="soapbox l-soapbox-top">
                        <span class="soapbox__title">Bright minds,</span>
                        <span class="soapbox__subtitle">better future</span>
                        <div class="soapbox__text">⋯</div>
                      </div>
                      <div class="soapbox l-soapbox-bottom">⋯</div>
                    </div>
                    <div class="content-block content-block--yellow">⋯</div>
                    <div class="content-block">⋯</div>
                    <div class="content-block">⋯</div>
                    <div class="breaker">⋯</div>
                    <style>⋯</style>
                    <style>⋯</style>
                  </article>
                </div>
              </main>
            </div>
          </div>
        </div>
        <section class="l-footer" role="complementary">⋯</section>
        <footer class="l-subfooter">⋯</footer>
      </div>
    </div>
    <script type="application/json" data-drupal-selector="drupal-settings-json">⋯</script>
    <script src="/sites/default/files/js/js_aFzAeDVUvTB0t1y8IOvQm_ak921BlHbAhsjKMjiZ2LQ.js"></script>
    <div id="ee4598d8-cf40-4d26-b497-0a5959a553ab-bottom"></div>
  </body>
</html>
```

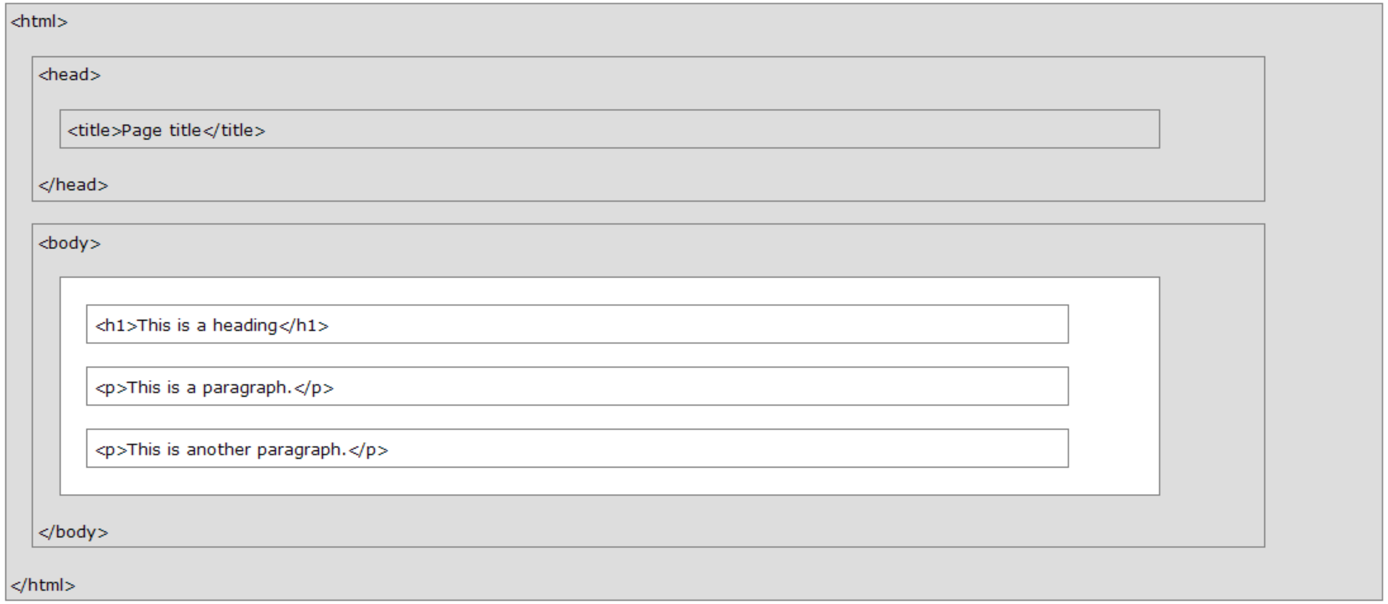
Hyper Text Markup Language (HTML)



```
<!DOCTYPE html>
<!-- Comment-->
<html  lang="en">
    <head>
        <meta charset="utf-8">
        <title>Hello World</title>
    </head>
    <body>
        <div>Say hello to my little world.</div>
    </body>
</html>
```

HTML Page Structure

`<tagname>Content goes here...</tagname>`

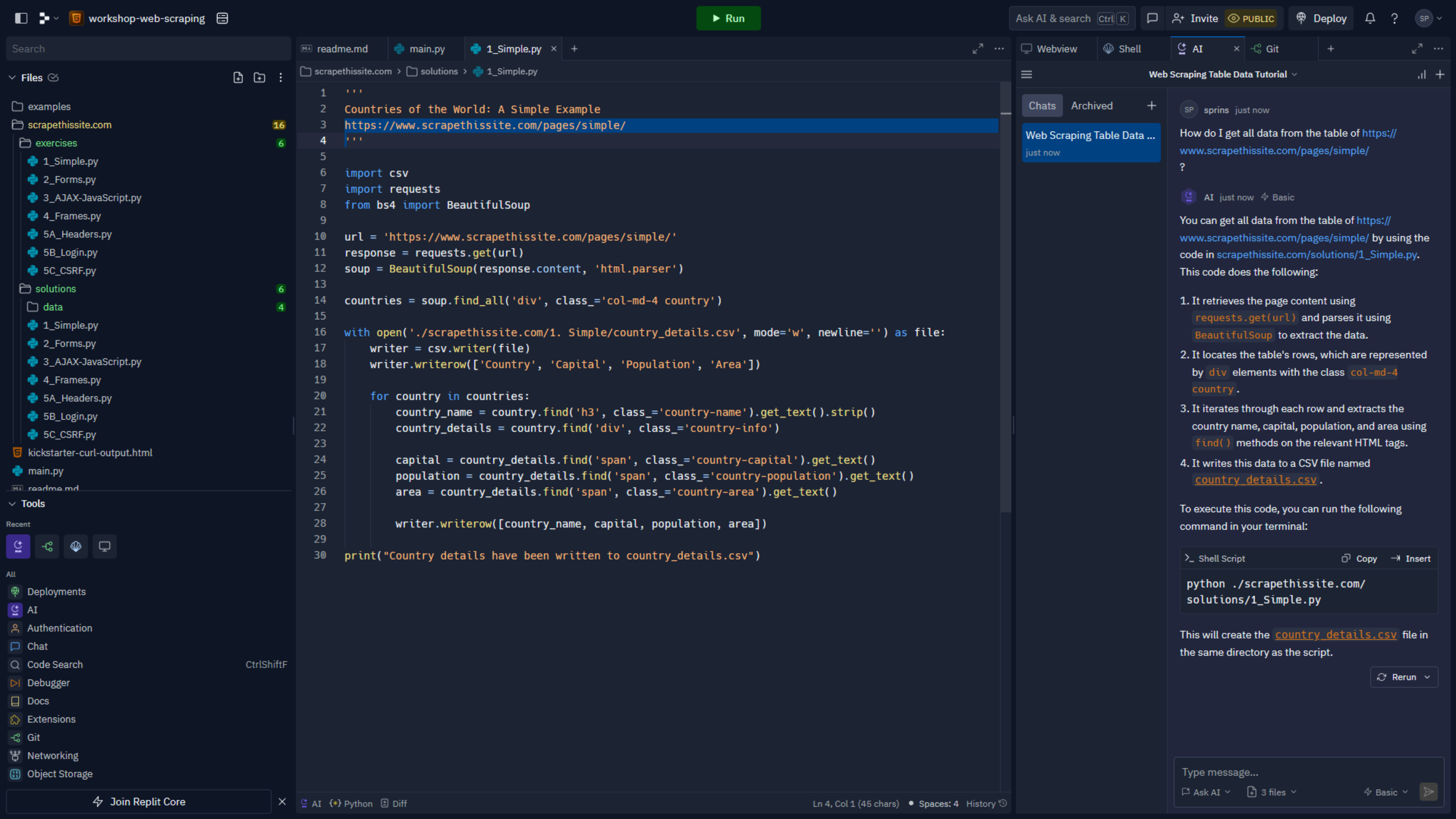


```
<html>
  <head>
    <title>Page title</title>
  </head>
  <body>
      <h1>This is a heading</h1>
      <p>This is a paragraph.</p>
      <p>This is another paragraph.</p>
  </body>
</html>
```

Python

- requests : Fetching the webpage
- Beautiful Soup : Parsing the HTML
- AI (Replit AI, GitHub CoPilot, ChatGPT)

Replit



```python
'''
Countries of the World: A Simple Example
https://www.scrapethissite.com/pages/simple/
'''

import csv
import requests
from bs4 import BeautifulSoup

url = 'https://www.scrapethissite.com/pages/simple/'
response = requests.get(url)
soup = BeautifulSoup(response.content, 'html.parser')

countries = soup.find_all('div', class_='col-md-4 country')

with open('./scrapethissite.com/1. Simple/country_details.csv', mode='w', newline='') as file:
    writer = csv.writer(file)
    writer.writerow(['Country', 'Capital', 'Population', 'Area'])

    for country in countries:
        country_name = country.find('h3', class_='country-name').get_text().strip()
        country_details = country.find('div', class_='country-info')

        capital = country_details.find('span', class_='country-capital').get_text()
        population = country_details.find('span', class_='country-population').get_text()
        area = country_details.find('span', class_='country-area').get_text()

        writer.writerow([country_name, capital, population, area])

print("Country details have been written to country_details.csv")
```

sprins just now

How do I get all data from the table of https://www.scrapethissite.com/pages/simple/ ?

AI just now Basic

You can get all data from the table of https://www.scrapethissite.com/pages/simple/ by using the code in scrapethissite.com/solutions/1_Simple.py. This code does the following:

1. It retrieves the page content using `requests.get(url)` and parses it using `BeautifulSoup` to extract the data.
2. It locates the table's rows, which are represented by `div` elements with the class `col-md-4 country`.
3. It iterates through each row and extracts the country name, capital, population, and area using `find()` methods on the relevant HTML tags.
4. It writes this data to a CSV file named `country_details.csv`.

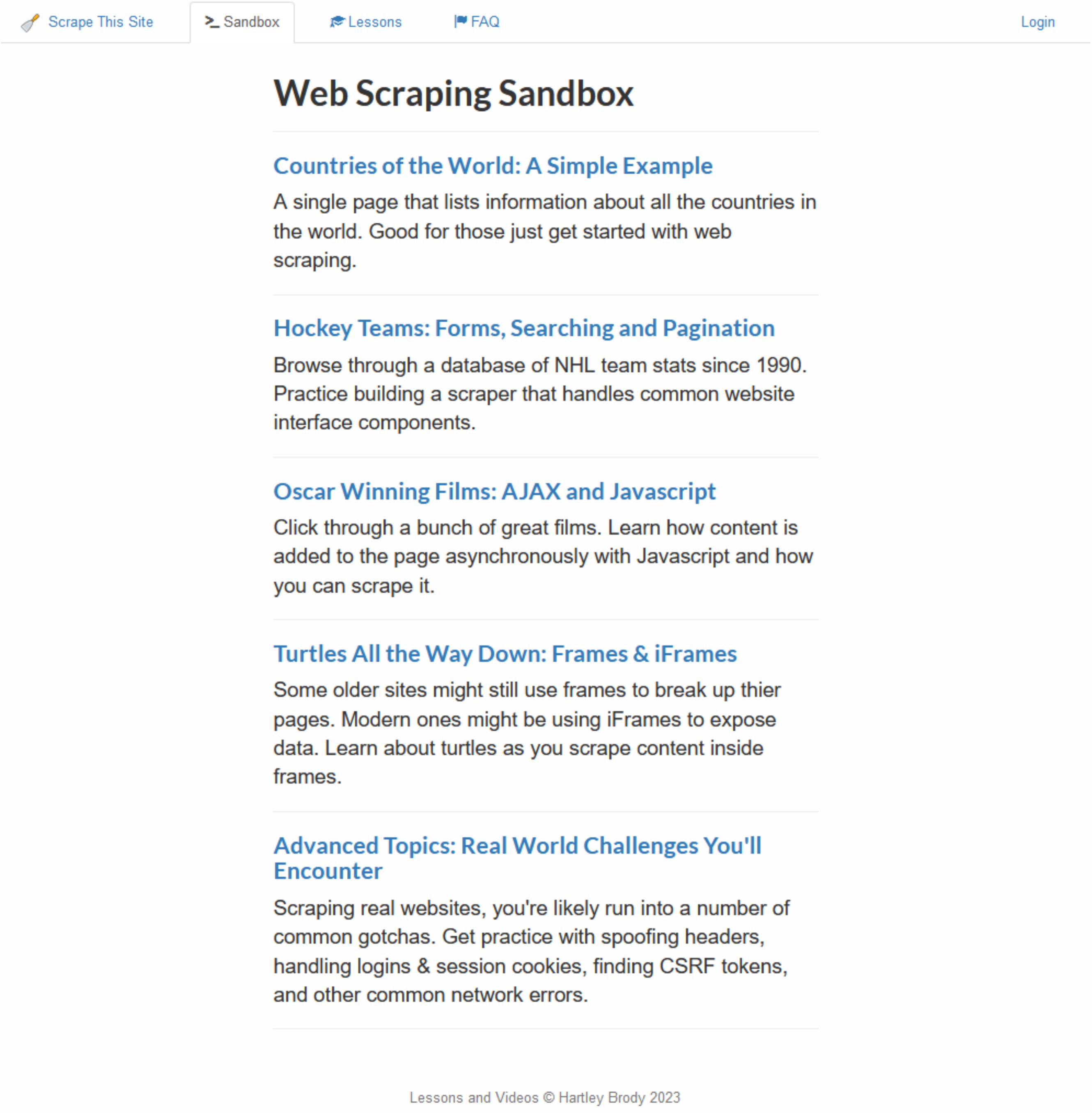
To execute this code, you can run the following command in your terminal:

Shell Script Copy Insert

```
python ./scrapethissite.com/
solutions/1_Simple.py
```

This will create the `country_details.csv` file in the same directory as the script.

Rerun



Web Scraping Sandbox

Countries of the World: A Simple Example

A single page that lists information about all the countries in the world. Good for those just get started with web scraping.

Hockey Teams: Forms, Searching and Pagination

Browse through a database of NHL team stats since 1990. Practice building a scraper that handles common website interface components.

Oscar Winning Films: AJAX and Javascript

Click through a bunch of great films. Learn how content is added to the page asynchronously with Javascript and how you can scrape it.

Turtles All the Way Down: Frames & iFrames

Some older sites might still use frames to break up thier pages. Modern ones might be using iFrames to expose data. Learn about turtles as you scrape content inside frames.

Advanced Topics: Real World Challenges You'll Encounter

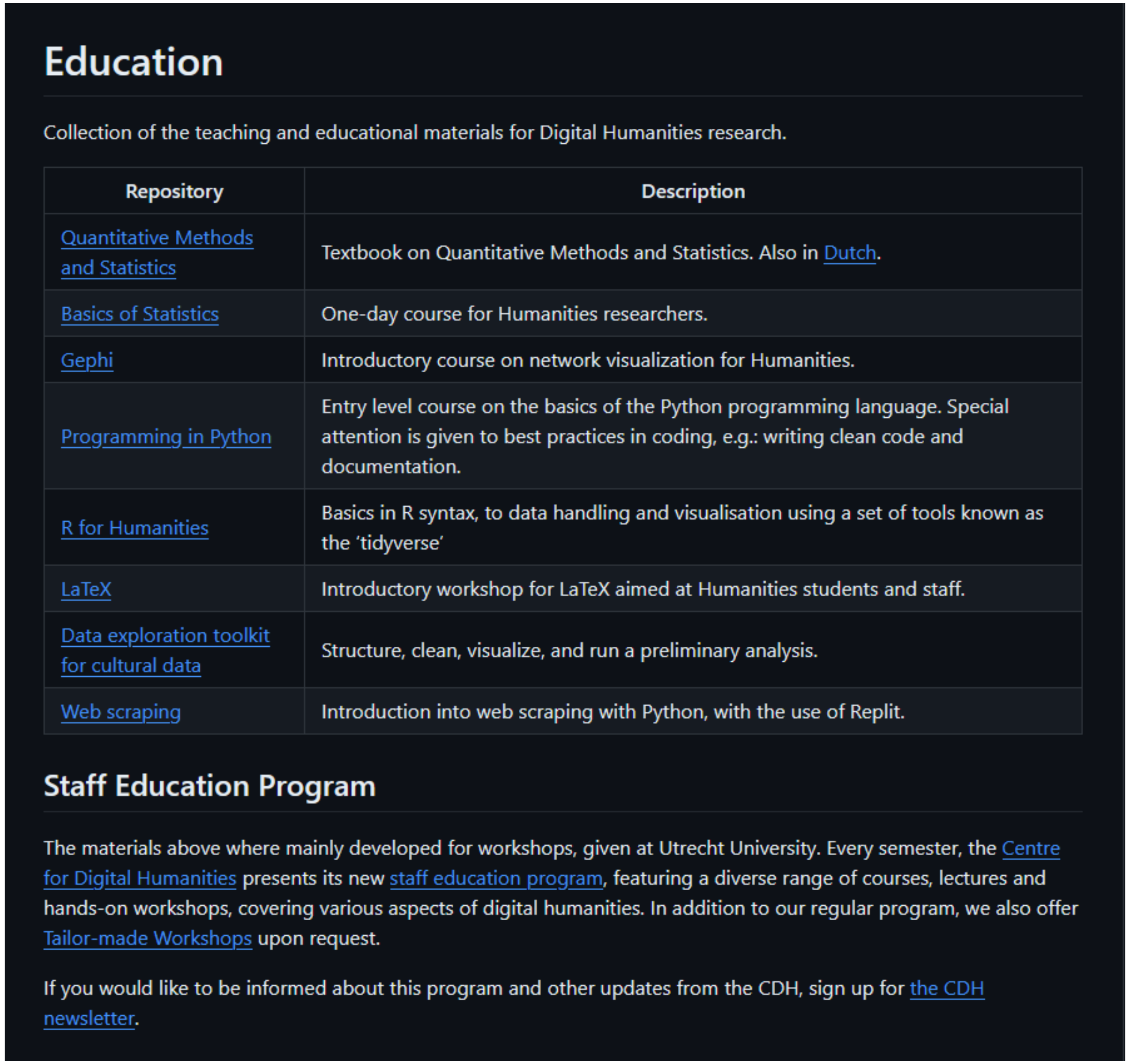
Scraping real websites, you're likely run into a number of common gotchas. Get practice with spoofing headers, handling logins & session cookies, finding CSRF tokens, and other common network errors.

Lessons and Videos © Hartley Brody 2023

Getting started

1. https://replit.com/
2. Open Shell
3. git clone https://github.com/CentreForDigitalHumanities/workshop-web-scraping

Materials



Education

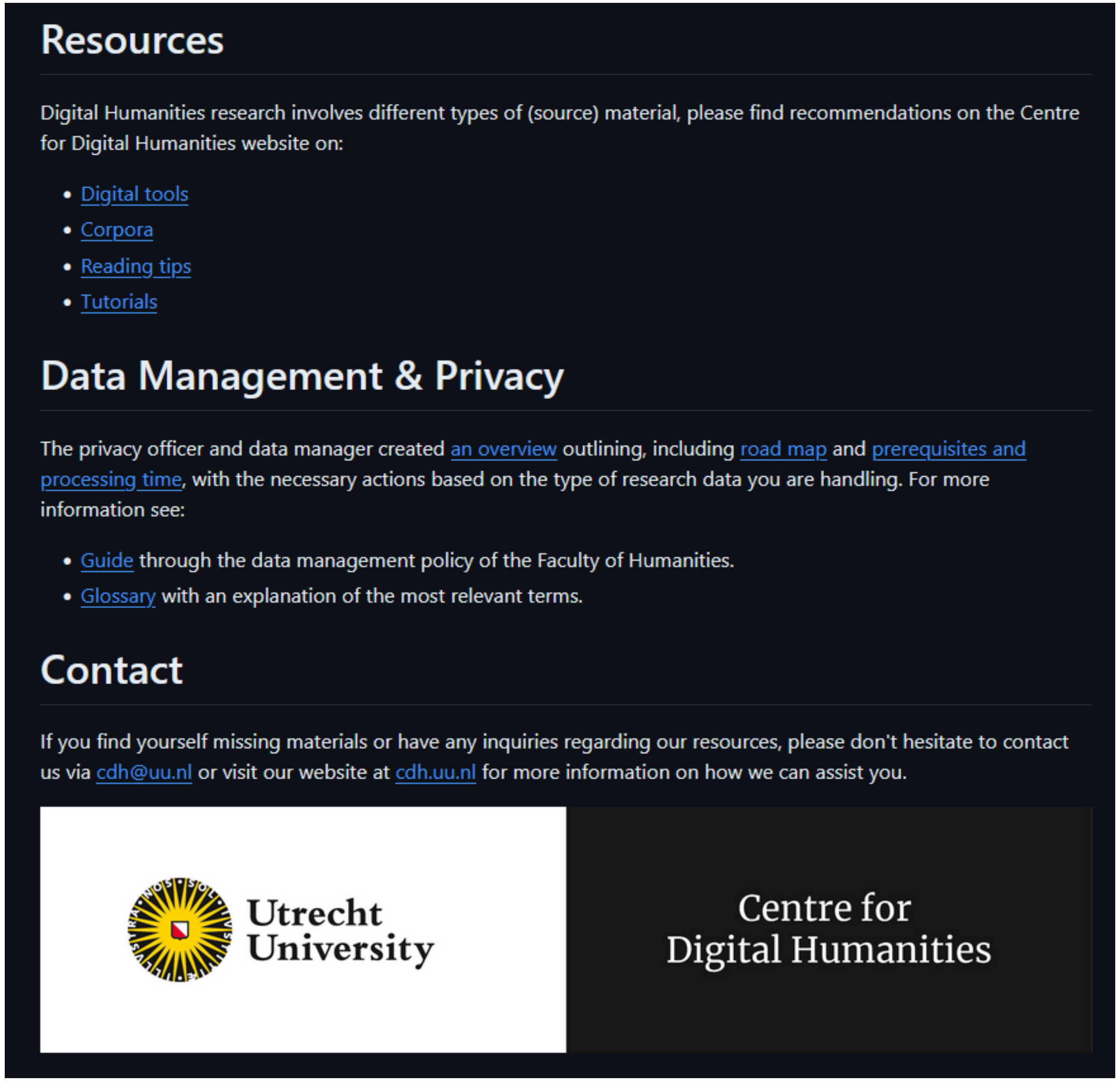
Collection of the teaching and educational materials for Digital Humanities research.

Repository	Description
Quantitative Methods and Statistics	Textbook on Quantitative Methods and Statistics. Also in Dutch.
Basics of Statistics	One-day course for Humanities researchers.
Gephi	Introductory course on network visualization for Humanities.
Programming in Python	Entry level course on the basics of the Python programming language. Special attention is given to best practices in coding, e.g.: writing clean code and documentation.
R for Humanities	Basics in R syntax, to data handling and visualisation using a set of tools known as the 'tidyverse'
LaTeX	Introductory workshop for LaTeX aimed at Humanities students and staff.
Data exploration toolkit for cultural data	Structure, clean, visualize, and run a preliminary analysis.
Web scraping	Introduction into web scraping with Python, with the use of Replit.

Staff Education Program

The materials above where mainly developed for workshops, given at Utrecht University. Every semester, the Centre for Digital Humanities presents its new staff education program, featuring a diverse range of courses, lectures and hands-on workshops, covering various aspects of digital humanities. In addition to our regular program, we also offer Tailor-made Workshops upon request.

If you would like to be informed about this program and other updates from the CDH, sign up for the CDH newsletter.



Resources

Digital Humanities research involves different types of (source) material, please find recommendations on the Centre for Digital Humanities website on:

- Digital tools
- Corpora
- Reading tips
- Tutorials

Data Management & Privacy

The privacy officer and data manager created an overview outlining, including road map and prerequisites and processing time, with the necessary actions based on the type of research data you are handling. For more information see:

- Guide through the data management policy of the Faculty of Humanities.
- Glossary with an explanation of the most relevant terms.

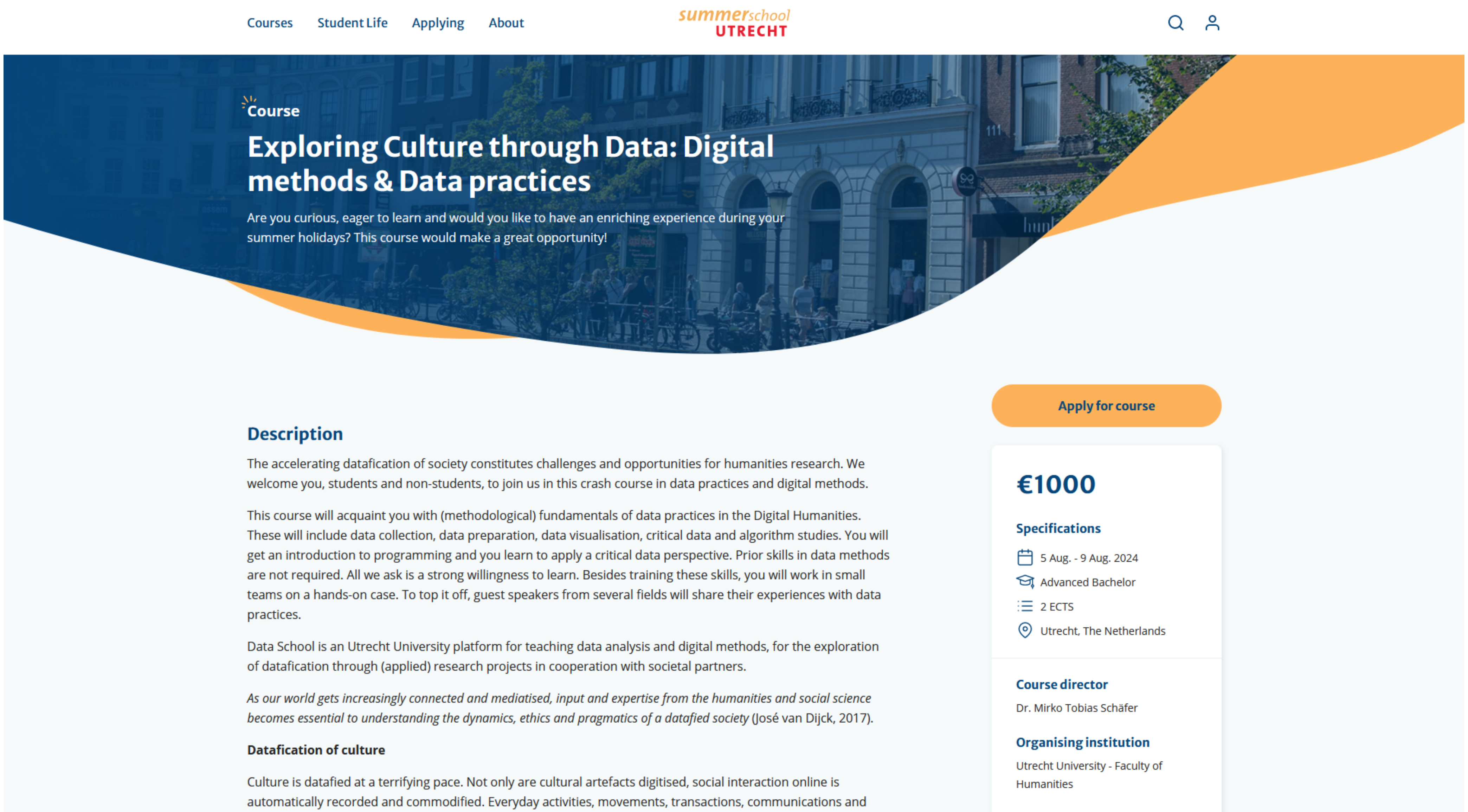
Contact

If you find yourself missing materials or have any inquiries regarding our resources, please don't hesitate to contact us via cdh@uu.nl or visit our website at cdh.uu.nl for more information on how we can assist you.

Utrecht University

Centre for Digital Humanities

https://github.com/CentreForDigitalHumanities/Education



Course

Exploring Culture through Data: Digital methods & Data practices

Are you curious, eager to learn and would you like to have an enriching experience during your summer holidays? This course would make a great opportunity!

Description

The accelerating datafication of society constitutes challenges and opportunities for humanities research. We welcome you, students and non-students, to join us in this crash course in data practices and digital methods.

This course will acquaint you with (methodological) fundamentals of data practices in the Digital Humanities. These will include data collection, data preparation, data visualisation, critical data and algorithm studies. You will get an introduction to programming and you learn to apply a critical data perspective. Prior skills in data methods are not required. All we ask is a strong willingness to learn. Besides training these skills, you will work in small teams on a hands-on case. To top it off, guest speakers from several fields will share their experiences with data practices.

Data School is an Utrecht University platform for teaching data analysis and digital methods, for the exploration of datafication through (applied) research projects in cooperation with societal partners.

As our world gets increasingly connected and mediatised, input and expertise from the humanities and social science becomes essential to understanding the dynamics, ethics and pragmatics of a datafied society (José van Dijck, 2017).

Datafication of culture

Culture is datafied at a terrifying pace. Not only are cultural artefacts digitised, social interaction online is automatically recorded and commodified. Everyday activities, movements, transactions, communications and

€1000

Specifications

5 Aug. - 9 Aug. 2024
Advanced Bachelor
2 ECTS
Utrecht, The Netherlands

Course director

Dr. Mirko Tobias Schäfer

Organising institution

Utrecht University - Faculty of Humanities

https://utrechtsummerschool.nl/courses/humanities/exploring-culture-through-data-digital-methods-data-practices

Questions?

Centre for Digital Humanities

cdh@uu.nl | cdh.uu.nl

Utrecht
University

Sharing science,
shaping tomorrow