

# Week 5: HTML, CSS, and Scraping Static Websites

LSE MY472: Data for Data Scientists

<https://lse-my472.github.io/>

Autumn Term 2024

**Ryan Hübert**

# Plan for today

- Introduction
- Some key features of the internet
- HTML and CSS
- Fundamentals of web scraping
- Coding

# Introduction

# Examples

An increasing amount of data is available on the web

- Speeches, biographical information ...
- Social media data, articles, press releases ...
- Geographic information, conflict data ...

These datasets are often provided in an **unstructured format**

**Web scraping** is the process of extracting this information automatically and transforming it into a **structured dataset**

# Why automate?

Copy & pasting is time-consuming, boring, prone to errors, and impractical or infeasible

## **In contrast, automated web scraping**

1. Scales well for large datasets
2. Allows for dynamic data collection
3. Is (mostly) reproducible
4. Involves adaptable techniques
5. Facilitates detecting and fixing errors

## **When to scrape?**

1. Trade-off between your time today and your time in the future.  
Invest in your future self!
2. Computer time is often cheap; human time more expensive

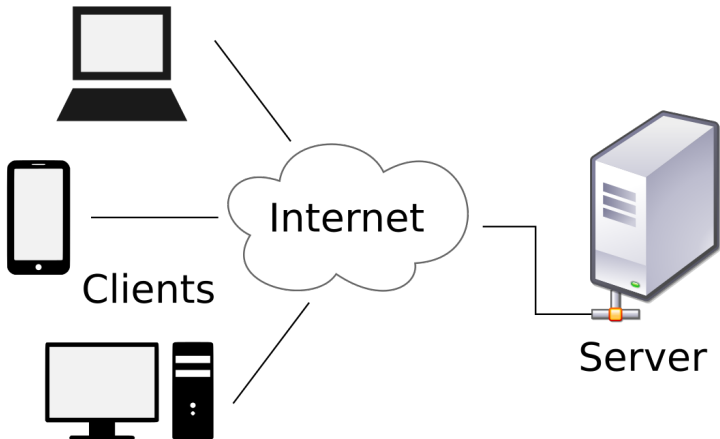
# Obtaining data from the web: Two approaches

## Two different approaches

- **Screen scraping** Extract data from source code of website, with html parser and/or regular expressions
  - `rvest` (this week) and `RSelenium` packages (week 7) in R
- **Web APIs** (week 8): A set of structured http requests that return JSON or XML data
  - `httr` package to construct API requests
  - Packages specific to each API: For example `WDI`, `Rfacebook`,
    - Check CRAN Task View on [Web Technologies and Services](#) for examples

Some key features of the internet

## Client-server model





# Client-server model

- Client: User computer, tablet, phone, software application, etc.
- Server: Web server, mail server, file server, Jupyter server, etc.

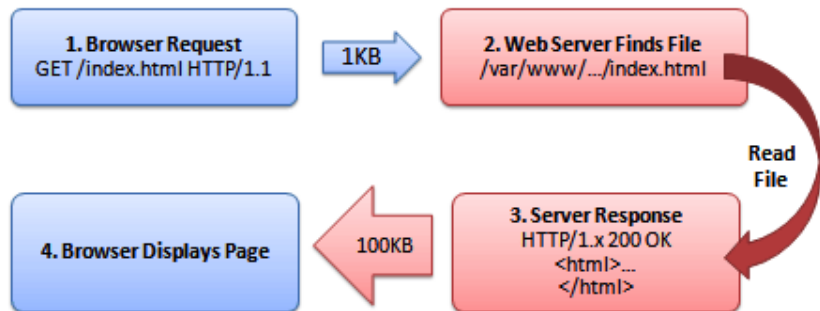
## 1. Client makes request to the server

- Depending on what you want to get, the request might be
  - HTTP: Hypertext Transfer Protocol
  - HTTPS: Hypertext Transfer Protocol Secure
  - SMTP: Simple Mail Transfer Protocol
  - FTP: File Transfer Protocol

## 2. Server returns response

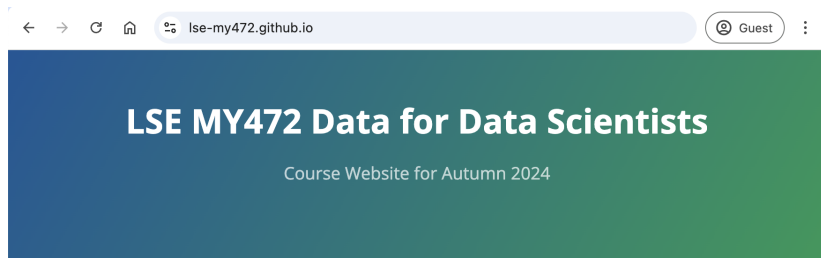
# Request and response in the case of HTTP

From [StackOverflow](#)



# Simple example: MY472 website

Let's see a very simple example of <https://lse-my472.github.io>



**Important note:** *The information on this page is provisional until the first lecture.*

## Course format and scheduling

**Lectures:** There is a two-hour lecture each week during the term on **Wednesdays from 13:00 to 15:00 in CLM.2.02.**

Simple example: MY472 website

Simple example: Request headers

Simple example: Response headers

Simple example: Reponse content

# HTML and CSS

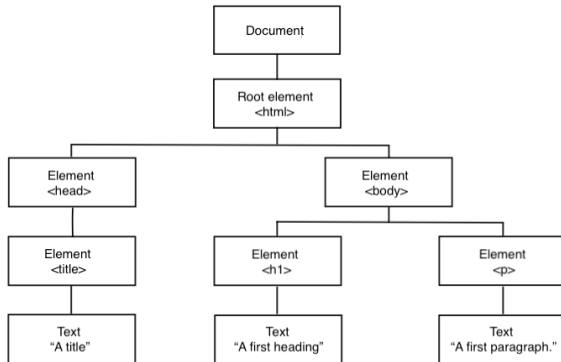


# HTML

## HTML: Hypertext Markup Language

- HTML displays mostly **static** content
- Many contents of dynamic webpages cannot be found in HTML
  - Example: Google Maps
- Understanding what is static and dynamic in a webpage is a crucial first step for web scraping

# HTML tree structure



# A very simple HTML file

```
<!DOCTYPE html>
<html>
  <head>
    <title>A title</title>
  </head>
  <body>
    <h1>A first heading</h1>
    <p>A first paragraph.</p>
  </body>
</html>
```

From: [https:](https://www.w3schools.com/html/tryit.asp?filename=tryhtml_intro)

[//www.w3schools.com/html/tryit.asp?filename=tryhtml\\_intro](https://www.w3schools.com/html/tryit.asp?filename=tryhtml_intro)

## Slightly more features

```
<!DOCTYPE html>
<html>
  <head>
    <title>A title</title>
  </head>
  <body>
    <h1>A first heading</h1>
    <p>A first paragraph.</p>
    <p>A second paragraph with some <b>formatted</b> text.</p>
    <p>A third paragraph with a <a href="http://www.lse.ac.uk">hyper
  </body>
</html>
```

# With some content divisions

```
<!DOCTYPE html>
<html>
  <head>
    <title>A title</title>
  </head>
  <body>
    <div>
      <h1>Heading of the first division</h1>
      <p>A first paragraph.</p>
      <p>A second paragraph with some <b>formatted</b> text.</p>
      <p>A third paragraph with a <a href="http://www.lse.ac.uk">
    </div>
    <div>
      <h1>Heading of the second division</h1>
      <p>Another paragraph with some text.</p>
    </div>
  </body>
</html>
```

# Beyond plain HTML

1. **Cascading Style Sheets (CSS)** Style sheet language which describes formatting of HTML components, useful for us because of selectors
2. **Javascript**: Adds functionalities to the websites, e.g. change content/structure after website has been loaded

## Adding some simple CSS (1/2)

```
<!DOCTYPE html>
<html>
  <head>

    <!-- CSS start -->
    <style>
    p {
      color: green;
    }
    </style>
    <!-- CSS end -->

    <title>A title</title>
  </head>
  <body>
    <div>
      <h1>Heading of the first division</h1>
      <p>A first paragraph.</p>
      <p>A second paragraph with some <b>formatted</b> text.</p>
      <p>A third paragraph with a <a href="http://www.lse.ac.uk">
    </div>
    <div>
      <h1>Heading of the second division</h1>
```

## Adding some simple CSS (2/2)

```
<!DOCTYPE html>
<html>
  <head>

    <!-- CSS start -->
    <style>
      .text-about-web-scraping {
        color: orange;
      }
      .division-two h1 {
        color: green;
      }
    </style>
    <!-- CSS end -->

    <title>A title</title>
  </head>
  <body>
    <div>
      <h1>Heading of the first division</h1>
      <p>A first paragraph.</p>
      <p>A second paragraph with some <b>formatted</b> text.</p>
      <p class="text-about-web-scraping">A third paragraph now co
```



# Fundamentals of web scraping

# Scenario 1: Data in table format



Article [Talk](#)

Read [Edit](#) [View history](#)

[Not logged in](#) [Talk](#) [Contributions](#) [Create account](#) [Log in](#)


## International court

From Wikipedia, the free encyclopedia

### List of international courts [\[ edit \]](#)

| Name  | Scope  | Years active | Subject matter                                    |
|---|--------|--------------|---|
| <a href="#">International Court of Justice</a>                | Global | 1945–present | General disputes                                  |
| <a href="#">International Criminal Court</a>                  | Global | 2002–present | Criminal prosecutions                             |
| <a href="#">Permanent Court of International Justice</a>      | Global | 1922–1946    | General disputes                                  |
| <a href="#">Appellate Body</a>                                | Global | 1995–present | Trade disputes within the <a href="#">WTO</a>     |
| <a href="#">International Tribunal for the Law of the Sea</a> | Global | 1994–present | Maritime disputes                                 |
| <a href="#">African Court of Justice</a>                      | Africa | 2009–present | Interpretation of <a href="#">AU</a> treaties     |
| <a href="#">African Court on Human and Peoples' Rights</a>    | Africa | 2006–present | Human rights                                      |
| <a href="#">COMESA Court of Justice</a>                       | Africa | 1998–present | Trade disputes within <a href="#">COMESA</a>      |
| <a href="#">ECOWAS Community Court of Justice</a>             | Africa | 1996–present | Interpretation of <a href="#">ECOWAS</a> treaties |
| <a href="#">East African Court of Justice</a>                 | Africa | 2001–present | Interpretation of <a href="#">EAC</a> treaties    |
| <a href="#">SADC Tribunal</a>                                 | Africa | 2005–2012    | Interpretation of <a href="#">SADC</a> treaties   |

# Scenario 2: Data in unstructured format



India English Register for updates Search 11,072,800 Visitors

I PAID A BRIBE

I DID NOT PAY A BRIBE

I MET AN HONEST OFFICER

BRIBE HOTLINE

ALL REPORTS

NEWS

REPORT A BRIBE

All Reports > I Paid A Bribe

ALL / I PAID A BRIBE / BRIBE FIGHTER / HONEST OFFICER / BRIBE HOTLINE

I PAID A BRIBE

1 day ago

76 views

**POLICE NILO GHUSS (bribe)**  
**Passport** | **Police Verification for Passport** | Paid INR 5,000  
Reported on **January 17, 2016** from **Bankura, West Bengal** | Report #89544  
  
What will happen to this country.. police mamu's govt income: 30,000 per month. Per day GHUSS income 5000 (per passport verification). Imagine they t...[Read more](#)  
  
**How to Get a Passport Verified in Ghaziabad**

I PAID A BRIBE

1 day ago

104 views

**Corruption due to vague rules**  
**Police** | **Traffic Violations** | Paid INR 500  
Reported on **January 16, 2016** from **Mumbai, Maharashtra** | Report #89509  
  
At Chembur near Eastern Expressway traffic cop stopped me and started checking docs..all was fine buy puc expired..then he pointed out film.. He took...[Read more](#)  
  
**Things to Know on Traffic Offences and Respective Penalties**

I PAID A BRIBE

2 days ago

105 views

**Bribe collected by Staff of Enrollment agency**  
**Municipal Services** | **Aadhaar or UID Related** | Paid INR 120  
Reported on **January 16, 2016** from **Mysore, Karnataka** | Report #89467  
  
UIDAI has to take a stand on fees to be paid to enrolment agencies for processing Aadhaar

**FILTER REPORTS**  
**Which city?**  
All cities  
**Department**  
All departments  
**Bribe Amount**  
All Amount  
**SUBMIT**

**INSPIRE OTHERS WITH YOUR STORY**  
Manik Taneja, a sports enthusiast, wrote against a custom official on [ipaidabribe.com](#), for cough up a hefty bribe by a Customs official at Bengaluru airport.  
**SEE HIS STORY**


Ever Paid A Bribe?


Report your Bribe Story!


See action taken.


<https://www.ipaidabribe.com/reports/paid>


# Scenario 3: Hidden behind web forms


 MONITOR  
LEGISLATIVO


 INICIO


 PERFIL IDEAL


 NOTICIAS

 CANDIDATOS

 ASAMBLEA NACIONAL

 ABUSOS

 CONTÁCTENOS



Seleccione ▾

Partido ▾

BUSCAR

## DIPUTADOS ENCONTRADOS



Julio Ygarza  
Estado: Amazonas



Mauligmer Baloa  
Estado: Amazonas



Nirma Guarulla  
Estado: Amazonas



José Brito  
Estado: Anzoátegui



Chaim Bucarán  
Estado: Anzoátegui



Richard Arteaga  
Estado: Anzoátegui



# Three main scenarios

## 1. Data in *table* format

- Automatic extraction with **rvest** or select specific table with *inspect element* in browser

## 2. Data in *unstructured* format

- Element identification key in this case
  - *Inspect element* in browser
- Identify the target e.g. with *CSS* (this week) or *XPath* selector (week 7)
- Automatic extraction with **rvest**

## 3. Data hidden *behind web forms* (week 7)

- Element identification to find text boxes, buttons, results, etc.
- Automation of web browser with **RSelenium**

# Identifying elements via CSS selector (1/2)

## → Selecting by tag-name

→ Example html code: `<h3>This is the main item</h3>`

→ Selector: `h3`

## → Selecting by class

→ Example html code: `<div class = 'itemdisplay'>This is the main item</div>`

→ Selector: `.itemdisplay`

## → Selecting by id

→ Example html code: `<div id = 'maintitle'>my main title</div>`

→ Selector: `#maintitle`

## Identifying elements via CSS selector (2/2)

### → Selecting by tag structure

→ Example html code (hyperlink tag a inside div tag): `<div><a href = 'https://www.google.com'>Google Link</a></div>`

→ Selector: `div a`

### → Selecting by nth child of a parent element

→ Example html code: `<body><p>First paragraph</p><p>Second paragraph.</p></body>`

→ Selector of second paragraph: `body > p:nth-child(2)`

You don't have to figure these out yourself: inspect!

Reference and further examples:

[https://www.w3schools.com/cssref/css\\_selectors.asp](https://www.w3schools.com/cssref/css_selectors.asp)

# The rules of the game

## 1. Respect the hosting site's wishes

- Check if an API exists or if data are available for download
- Respect copyright and ethics; what are you allowed to do?
- Keep in mind where data comes from and give credit
- Some websites disallow scrapers via `robots.txt` file

## 2. Limit your bandwidth use

- Wait some time after each hit
- Scrape only what you need, and just once

## 3. When using APIs, read documentation

- Is there a batch download option?
- Are there any rate limits?
- Can you share the data?



Coding

## Markdown files this week

- 01-selecting-elements.Rmd
- 02-scraping-tables.Rmd