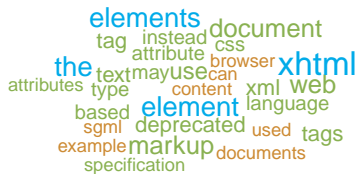


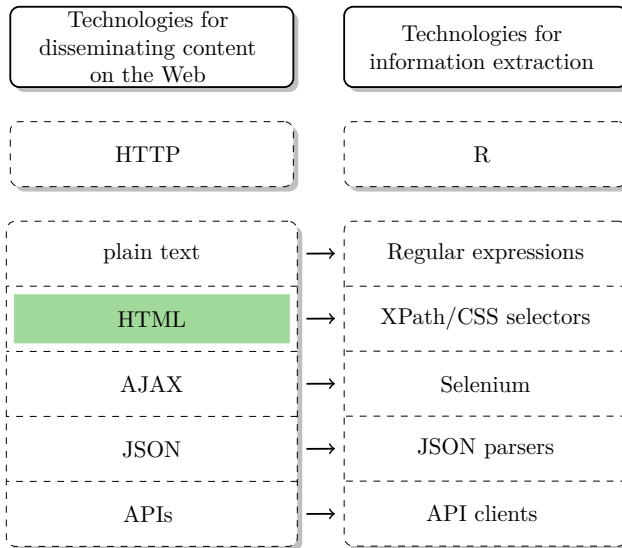
Introduction to Web Scraping with R

HTML



Simon Munzert | IPSDS

Technologies of the World Wide Web



HTML basics

HTML – a quick primer

What's HTML?

- **H**yper**T**ext **M**arkup **L**anguage
- markup language = plain text + markups
- W3C standard for the construction of websites
- lies underneath of what you see in your browser

HTML – a quick primer

What's HTML?

- **H**yper**T**ext **M**arkup **L**anguage
- markup language = plain text + markups
- W3C standard for the construction of websites
- lies underneath of what you see in your browser

Why is this important to us?

- it determines where and how information is stored
- a basic understanding of HTML helps us locate the information we want to retrieve
- relax. A passive understanding of HTML is sufficient

HTML in the wild

W Berlin - Wikipedia

Simon

← → ↻ 🏠

Sicher https://en.wikipedia.org/wiki/Berlin

☆ 🔴 🟢 🟠 🟡

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

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



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

[Interaction](#)
[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Wikidata item](#)
[Cite this page](#)

[Print/export](#)

Berlin

From Wikipedia, the free encyclopedia

This article is about the capital of Germany. For other uses, see [Berlin \(disambiguation\)](#).

Berlin (/bərˈlɪn/, German: [ˈbɛʁˈlɪn] (listen)) is the **capital** and the largest city of **Germany** as well as one of its constituent 16 **states**. With a population of approximately 3.5 million people,^[4] Berlin is the second **most populous city proper** and the seventh **most populous urban area** in the **European Union**.^[5] Located in northeastern Germany on the banks of rivers **Spree** and **Havel**, it is the centre of the **Berlin-Brandenburg Metropolitan Region**, which has about 6 million residents from more than 180 nations.^{[6][7][8][9]} Due to its location in the **European Plain**, Berlin is influenced by a **temperate** seasonal climate. Around one-third of the city's area is composed of forests, parks, gardens, rivers and lakes.^[10]

First documented in the 13th century and situated at the crossing of two important historic **trade routes**,^[11] Berlin became the capital of the **Margraviate of Brandenburg** (1417–1701), the **Kingdom of Prussia** (1701–1918), the **German Empire** (1871–1918), the **Weimar Republic** (1919–1933) and the **Third Reich** (1933–1945).^[12] Berlin in the 1920s was the third largest municipality in the world.^[13] After **World War II** and its consequent occupation by the victorious countries, the city was divided; **East Berlin** became the capital of **East Germany** while **West Berlin** became a **de facto West German exclave**, surrounded by the **Berlin Wall** (1961–1989).^[14] Following **German reunification** in

Berlin

State of Germany



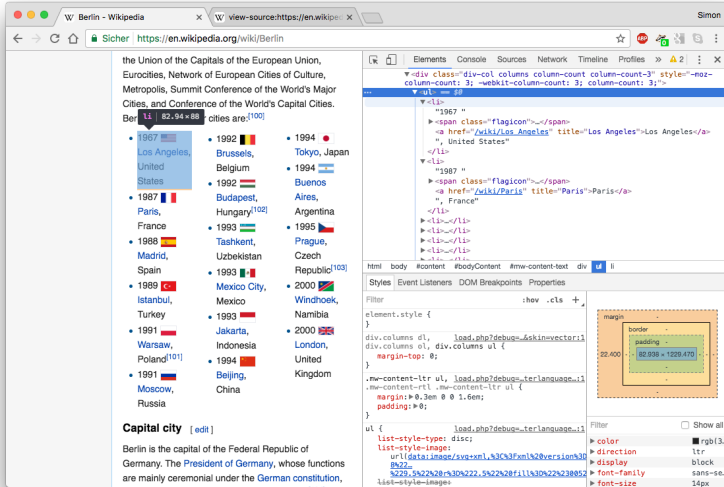
https://en.wikipedia.org/wiki/File:Alte_Nationalgalerie_Berlin,_2011.jpg

ory.^[14] Following **German reunification** in

HTML in the wild

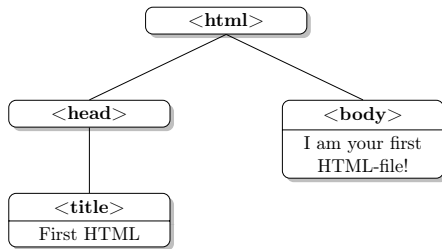


HTML in the wild



Tree structure

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <title id=1>First HTML</title>
5   </head>
6   <body>
7     I am your first HTML file!
8   </body>
9 </html>
```



Elements and attributes

Elements and attributes

Elements

Elements are a combination of *start tags*, content, and *end tags*.

Example:

1

```
<title>First HTML</title>
```

Syntax

element title	<code>title</code>
start tag	<code><title></code>
end tag	<code></title></code>
value	<code>First HTML</code>

Elements and attributes

Attributes

Attributes describe elements and are stored in the start tag. In HTML, there are specific attributes for specific elements.

Example:

```
1 <a href="http://www.r-datacollection.com/">Link to Homepage</a>
```

Syntax

- name-value pairs: `name="value"`
- simple and double quotation marks possible
- several attributes per element possible

Important tags and attributes

Why tags and attributes are important

- tags structure HTML documents
- everything that structures a document can be used to extract information
- in the following, we get to know some important tags which are useful when scraping information from the Web

Important tags and attributes

Anchor tag `<a>`

- links to other pages or resources
- classical links are always formatted with an anchor tag
- the `href` attribute determines the target location
- the value is the name of the link

Link to another resource:

```
1 <a href="en.wikipedia.org/wiki/List_of_lists_of_lists">Link with absolute path</a>
```

Reference in a document:

```
1 <a id="top">Reference Point</a>
```

Link to a reference:

```
1 <a href="#top">Link to Reference Point</a>
```

Important tags and attributes

Heading tags `<h1>`, `<h2>`, ..., and paragraph tag `<p>`

- structure text and paragraphs
- heading tags range from level 1 to 6
- paragraph tag induces line break

Examples:

```
1 <p>This text is going to be a paragraph one day and separated from other  
2 text by line breaks.</p>
```

```
1 <h1>heading of level 1 -- this will be BIG</h1>  
2 ...  
3 <h6>heading of level 6 -- the smallest heading</h6>
```

Important tags and attributes

Listing tags ``, `` and `<dl>`

- the `` tag creates a numeric list, `` an unnumbered list, `<dl>` a definition list
- list elements are indicated with the `` tag

Example:

```
1 <ul>
2   <li>Dogs</li>
3   <li>Cats</li>
4   <li>Fish</li>
5 </ul>
```


Important tags and attributes

Organizational tags `<div>` and ``

- grouping of content over lines (`<div>`) or within lines (``)
- do not change the layout themselves but work together with CSS

Example of CSS definition

```
1  div.happy { color:pink;  
2              font-family:"Comic Sans MS";  
3              font-size:120% }  
4  span.happy { color:pink;  
5              font-family:"Comic Sans MS";  
6              font-size:120% }
```

In the HTML document

```
1  <div class="happy"><p>I am a happy styled paragraph</p></div>  
2  non-happy text with <span class="happy">some happiness</span>
```

Important tags and attributes

Form tag `<form>`

- allows to incorporate HTML forms
- client can send information to the HTTP server via forms
- whenever you type something into a field or click on radio buttons in your browser, you are interacting with forms

Example:

```
1 <form name="submitPW" action="Passed.html" method="get">
2   password:
3   <input name="pw" type="text" value="">
4   <input type="submit" value="SubmitButtonText">
5 </form>
```

Important tags and attributes

Table tags `<table>`, `<tr>`, `<td>`, and `<th>`

- standard HTML tables always follow a standard architecture
- the different tags allow to define the table as a whole, individual rows (including the heading), and cells
- if the data is hidden in tables, scraping will be straightforward

Example:

```
1 <table>
2   <tr> <th>Rank</th> <th>Nominal GDP</th> <th>Name</th> </tr>
3   <tr> <th></th> <th>(per capita, USD)</th> <th></th> </tr>
4   <tr> <td>1</td> <td>170,373</td> <td>Lichtenstein</td> </tr>
5   <tr> <td>2</td> <td>167,021</td> <td>Monaco</td> </tr>
6   <tr> <td>3</td> <td>115,377</td> <td>Luxembourg</td> </tr>
7   <tr> <td>4</td> <td>98,565</td> <td>Norway</td> </tr>
8   <tr> <td>5</td> <td>92,682</td> <td>Qatar</td> </tr>
9 </table>
```

Summary

Summary

- HTML is the *lingua franca* on the web
- content on webpages is structured by HTML tags that are nested in a tree structure
- to break open information, we will have to locate it in the HTML tree
- for web scraping purposes, a mostly passive knowledge of HTML is sufficient

`<DIV>Q: HOW DO YOU ANNOY A WEB DEVELOPER?`

Source: <https://xkcd.com/1144/> (Randall Munroe)