

Introduction to Web Scraping with R

An Introductory Case Study



Simon Munzert | IPSDS

Data on the web



The screenshot shows a web browser window with the URL <https://en.wikipedia.org/wiki/Berlin>. The page title is "Berlin - Wikipedia". The browser's address bar shows the URL and a "Sicher" (Secure) icon. The page content includes the Wikipedia logo, a sidebar with navigation links, and the main article text. The article text starts with "Berlin (/bəˈlɪn/, German: [bɛʁˈliːn] ⓘ 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). Following German reunification in 1990, Berlin became the capital of the unified Germany.

The sidebar on the left includes links to the 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, and Print/export.

The main article text is titled "Berlin" and includes a sub-header "From Wikipedia, the free encyclopedia". It also includes a note: "This article is about the capital of Germany. For other uses, see Berlin (disambiguation).".

Below the text is a gallery of images related to Berlin, including the Brandenburg Gate, the Berlin Wall, and the Reichstag building.

Data on the web

boroughs mayors make up the council of mayors (*rat der burgermeister*), which is led by the city's Governing Mayor and advises the Senate. The neighborhoods have no local government bodies.

Twin towns – sister cities [\[edit \]](#)

See also: [List of twin towns and sister cities in Germany](#)

Berlin maintains official partnerships with 17 cities.^[100] **Town twinning** between Berlin and other cities began with its sister city Los Angeles in 1967. East Berlin's partnerships were canceled at the time of German reunification but later partially reestablished. West Berlin's partnerships had previously been restricted to the borough level. During the Cold War era, the partnerships had reflected the different power blocs, with West Berlin partnering with capitals in the Western World, and East Berlin mostly partnering with cities from the [Warsaw Pact](#) and its allies.


There are several joint projects with many other cities, such as [Beirut](#), [Belgrade](#), [São Paulo](#), [Copenhagen](#), [Helsinki](#), [Johannesburg](#), [Mumbai](#), [Oslo](#), [Shanghai](#), [Seoul](#), [Sofia](#), [Sydney](#), [New York City](#) and [Vienna](#). Berlin participates in international city associations such as the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities. Berlin's official sister cities are:^[100]

• 1967  Los Angeles , United States	• 1992  Brussels , Belgium	• 1994  Tokyo , Japan
• 1987  Paris , France	• 1992  Budapest , Hungary ^[102]	• 1994  Buenos Aires , Argentina
• 1988  Madrid , Spain	• 1993  Tashkent , Uzbekistan	• 1995  Prague , Czech Republic ^[103]
• 1989  Istanbul , Turkey	• 1993  Mexico City , Mexico	• 2000  Windhoek , Namibia
• 1991  Warsaw , Poland ^[101]	• 1993  Jakarta , Indonesia	• 2000  London , United Kingdom
• 1991  Moscow , Russia	• 1994  Beijing , China	

Capital city [\[edit \]](#)

Berlin is the capital of the Federal Republic of Germany. The [President of Germany](#), whose functions are mainly ceremonial under the [German constitution](#), has his official residence in [Schloss Bellevue](#).^[104] Berlin is the seat of the [German executive](#), housed in the [Chancellery](#), the *Bundeskanzleramt*. Facing the Chancellery is the [Bundestag](#), the German Parliament, housed in the renovated [Reichstag building](#) since the government relocated to Berlin in 1998. The [Bundesrat](#) ("federal council", performing the function of an upper house) is the representation of the Federal States (*Bundesländer*) of Germany and has its

Data on the web






The screenshot shows a web browser window with the URL <https://en.wikipedia.org/wiki/Berlin>. The page content includes a paragraph about the council of mayors, a section on twin towns, and a section on the capital city. A red rectangular box highlights the list of twin towns, which is organized into three columns. Each entry includes a year, a flag, and the city name with its country.

Twin towns – sister cities [edit]

See also: *List of twin towns and sister cities in Germany*

Berlin maintains official partnerships with 17 cities.^[100] **Town twinning** between Berlin and other cities began with its sister city Los Angeles in 1967. East Berlin's partnerships were canceled at the time of German reunification but later partially reestablished. West Berlin's partnerships had previously been restricted to the borough level. During the Cold War era, the partnerships had reflected the different power blocs, with West Berlin partnering with capitals in the Western World, and East Berlin mostly partnering with cities from the **Warsaw Pact** and its allies.

There are several joint projects with many other cities, such as Beirut, Belgrade, São Paulo, Copenhagen, Helsinki, Johannesburg, Mumbai, Oslo, Shanghai, Seoul, Sofia, Sydney, New York City and Vienna. Berlin participates in international city associations such as the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities. Berlin's official sister cities are:^[100]

• 1967  Los Angeles , United States	• 1992  Brussels , Belgium	• 1994  Tokyo , Japan
• 1987  Paris , France	• 1992  Budapest , Hungary ^[102]	• 1994  Buenos Aires , Argentina
• 1988  Madrid , Spain	• 1993  Tashkent , Uzbekistan	• 1995  Prague , Czech Republic ^[103]
• 1989  Istanbul , Turkey	• 1993  Mexico City , Mexico	• 2000  Windhoek , Namibia
• 1991  Warsaw , Poland ^[101]	• 1993  Jakarta , Indonesia	• 2000  London , United Kingdom
• 1991  Moscow , Russia	• 1994  Beijing , China	

Capital city [edit]

Berlin is the capital of the Federal Republic of Germany. The **President of Germany**, whose functions are mainly ceremonial under the **German constitution**, has his official residence in **Schloss Bellevue**.^[104] Berlin is the seat of the **German executive**, housed in the **Chancellery**, the *Bundeskanzleramt*. Facing the Chancellery is the **Bundestag**, the German Parliament, housed in the renovated **Reichstag building** since the government relocated to Berlin in 1998. The **Bundesrat** ("federal council", performing the function of an upper house) is the representation of the Federal States (*Bundesländer*) of Germany and has its

Let's grab these data!

Step 1: Load packages

R code

```
1 library(rvest)
2 library(stringr)
```

end

Let's grab these data!

Step 2: Parse page source

R code

```
3 library(rvest)
4 library(stringr)
5 parsed_url <- read_html("https://en.wikipedia.org/wiki/Berlin")
```

end

Let's grab these data!

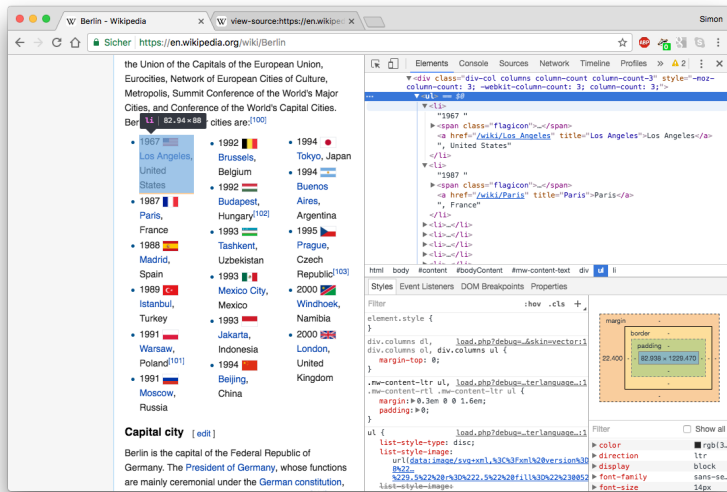
Step 3: Extract information

R code

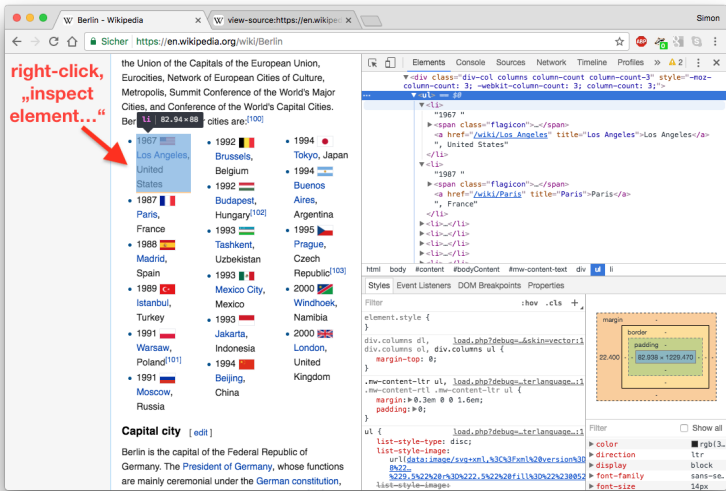
```
6 library(rvest)
7 library(stringr)
8 parsed_url <- read_html("https://en.wikipedia.org/wiki/Berlin")
9 parsed_nodes <- html_nodes(parsed_url, xpath = "//div[contains(@class, 'column-count-3')]/li"
)
10 cities <- html_text(parsed_nodes)
11 cities[1:10]
[1] "1967 Los Angeles, United States" "1987 Paris, France"
[3] "1988 Madrid, Spain"              "1989 Istanbul, Turkey"
[5] "1991 Warsaw, Poland[103]"        "1991 Moscow, Russia"
[7] "1992 Brussels, Belgium"          "1992 Budapest, Hungary[104]"
[9] "1993 Tashkent, Uzbekistan"        "1993 Mexico City, Mexico"
```

end

Why was this so easy?



Why was this so easy?



Why was this so easy?

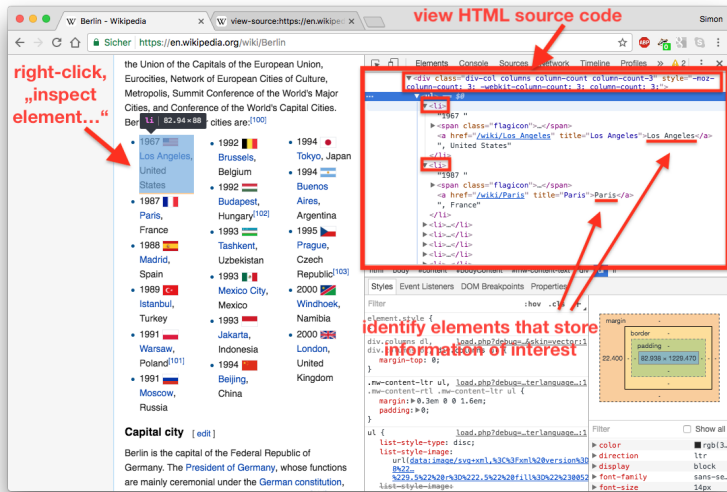
The screenshot shows a web browser window with the Wikipedia page for Berlin. The page title is "Berlin - Wikipedia" and the URL is "https://en.wikipedia.org/wiki/Berlin". The page content includes a list of cities in the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities. The list of cities is organized by year, with each entry including a flag and the city name. The list is as follows:

Year	City	Year	City	Year	City
1967	Los Angeles, United States	1992	Brussels, Belgium	1994	Tokyo, Japan
1987	Paris, France	1992	Buenos Aires, Argentina	1994	Buenos Aires, Argentina
1988	Madrid, Spain	1993	Tashkent, Uzbekistan	1995	Prague, Czech Republic
1989	Istanbul, Turkey	1993	Mexico City, Mexico	2000	Windhoek, Namibia
1991	Warsaw, Poland	1993	Jakarta, Indonesia	2000	London, United Kingdom
1991	Moscow, Russia	1994	Beijing, China	2000	Kingdom

The "Capital city" section is also visible, stating that Berlin is the capital of the Federal Republic of Germany. The HTML source code is open, showing the structure of the page. A red arrow points to the "view HTML source code" button in the top right corner. Another red arrow points to the "right-click, inspect element..." action on the list of cities. The source code shows the following structure:

```
<div class="div-col columns column-count column-count-3" style="moz-column-count: 3; webkit-column-count: 3; column-count: 3;">
  <ul>
    <li>
      "1967 "
      <span class="flagicon"></span>
      <a href="/wiki/Los_Angeles" title="Los Angeles">Los Angeles</a>
      ", United States"
    </li>
    <li>
      "1987 "
      <span class="flagicon"></span>
      <a href="/wiki/Paris" title="Paris">Paris</a>
      ", France"
    </li>
    </ul>
  </div>
```

Why was this so easy?



Let's clean up these data!

R code

```
12 cities[1:10]
```

```
[1] "1967 Los Angeles, United States" "1987 Paris, France"  
[3] "1988 Madrid, Spain"             "1989 Istanbul, Turkey"  
[5] "1991 Warsaw, Poland[103]"       "1991 Moscow, Russia"  
[7] "1992 Brussels, Belgium"         "1992 Budapest, Hungary[104]"  
[9] "1993 Tashkent, Uzbekistan"       "1993 Mexico City, Mexico"
```

end

Let's clean up these data!

R code

```
15 cities[1:10]
```

```
[1] "1967 Los Angeles, United States" "1987 Paris, France"  
[3] "1988 Madrid, Spain"              "1989 Istanbul, Turkey"  
[5] "1991 Warsaw, Poland[103]"        "1991 Moscow, Russia"  
[7] "1992 Brussels, Belgium"          "1992 Budapest, Hungary[104]"  
[9] "1993 Tashkent, Uzbekistan"        "1993 Mexico City, Mexico"
```

end

Step 1: Remove footnotes with a regular expression

R code

```
16 cities <- str_replace(cities, "\\[\\d+\\]", "")
```

```
17 cities[1:10]
```

```
[1] "1967 Los Angeles, United States" "1987 Paris, France"  
[3] "1988 Madrid, Spain"              "1989 Istanbul, Turkey"  
[5] "1991 Warsaw, Poland"            "1991 Moscow, Russia"  
[7] "1992 Brussels, Belgium"          "1992 Budapest, Hungary"  
[9] "1993 Tashkent, Uzbekistan"        "1993 Mexico City, Mexico"
```

end

Let's clean up these data!

Step 2: Extract data with regular expressions

R code

```
18 year <- str_extract(cities, "\\d{4}")
19 city <- str_extract(cities, "[[:alpha:]]+") %>% str_trim
20 country <- str_extract(cities, "[[:alpha:]]+$") %>% str_trim
21 year[1:10]
  [1] "1967" "1987" "1988" "1989" "1991" "1991" "1992" "1992" "1993" "1993"
22 city[1:10]
  [1] "Los Angeles" "Paris"          "Madrid"          "Istanbul"        "Warsaw"
  [6] "Moscow"      "Brussels"       "Budapest"       "Tashkent"       "Mexico City"
23 country[1:10]
  [1] "United States" "France"          "Spain"          "Turkey"
  [5] "Poland"        "Russia"         "Belgium"       "Hungary"
  [9] "Uzbekistan"   "Mexico"
```

end

Let's clean up these data!

Step 3: Put everything into data frame

R code

```
24 cities_df <- data.frame(year, city, country)
25 head(cities_df)
```

	year	city	country
1	1967	Los Angeles	United States
2	1987	Paris	France
3	1988	Madrid	Spain
4	1989	Istanbul	Turkey
5	1991	Warsaw	Poland
6	1991	Moscow	Russia

end

Let's map these data!

Step 1: Load necessary packages

R code

```
26 library(ggmap)
```

```
27 library(maps)
```

end

Let's map these data!

Step 2: Geocode cities with the Google Maps API

R code

```
28 library(ggmap)
29 library(maps)
30 cities_coords <- geocode(paste0(cities_df$city, ", ", cities_df$country))
31 cities_df$lon <- cities_coords$lon
32 cities_df$lat <- cities_coords$lat
33 cities_df$lon[1:10]
  [1] -118.243685    2.352222   -3.703790    28.978359         NA
  [6]   37.617300         NA   19.040235         NA  -99.133208
34 cities_df$lat[1:10]
  [1] 34.05223 48.85661 40.41678 41.00824         NA 55.75583         NA
  [8] 47.49791         NA 19.43261
```

end

Let's map these data!

Step 3: Plot world map, add coordinates

R code

```
35 map_world <- borders("world", colour = "gray50", fill = "white")
36 ggplot() + map_world + geom_point(aes(x = cities_df$lon, y = cities_df$lat), color = "red",
  size = 1) + theme_void()
```

end

Let's map these data!

Step 3: Plot world map, add coordinates

R code

```
37 map_world <- borders("world", colour = "gray50", fill = "white")
38 ggplot() + map_world + geom_point(aes(x = cities_df$lon, y = cities_df$lat), color = "red",
size = 1) + theme_void()
```

end

