

# Webscrapping with RSelenium

Automate your browser actions

Etienne Bacher

LISER

2022-08-04

# Introduction

Scraping can be divided in two steps:

1. getting the HTML that contains the information
2. cleaning the HTML to extract the information we want

These 2 steps don't require the same tools, and *shouldn't be made at the same time*.

Here, we will focus on step 1: *how to get the HTML we need with dynamic webpages?*

# Static vs dynamic

**Static webpage:** all the information is loaded with the page.

Example: Wikipedia.

**Dynamic webpage:** the website uses JavaScript to fetch data from their servers and *dynamically* update the page.

Example: see later.

**(R)Selenium**

# Idea

Idea: control the browser from the command line.

*I wish I could click on this button to open a modal*

```
remote_driver$  
  findElement(using = "css", value = ".my-button")$  
  clickElement()
```

*I wish I could fill these inputs to automatically connect*

```
remote_driver$  
  findElement(using = "id", value = "password")$  
  sendKeysToElement(list("my_super_secret_password"))
```

Almost everything you can do "by hand" in a browser, you can reproduce with Selenium:

- open a browser
  - click on something
  - enter values
  - go to previous/next page
  - refresh the page
  - get all the HTML that is currently displayed
- `open()` / `navigate()`
  - `clickElement()`
  - `sendKeysToElement()`
  - `goBack()` / `goForward()`
  - `refresh()`
  - `getPageSource()`

...

**Get started**

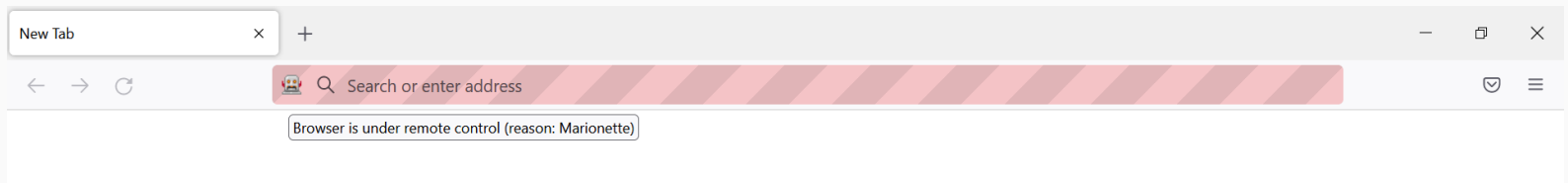


In the beginning there was ~~light~~ `rsDriver()`:

```
# if not already installed
# install.packages("RSelenium")
library(RSelenium)

driver <- rsDriver(browser = "firefox") # can also be chrome
remote_driver <- driver[["client"]]
```

This will print a bunch of messages and open a "marionette browser".



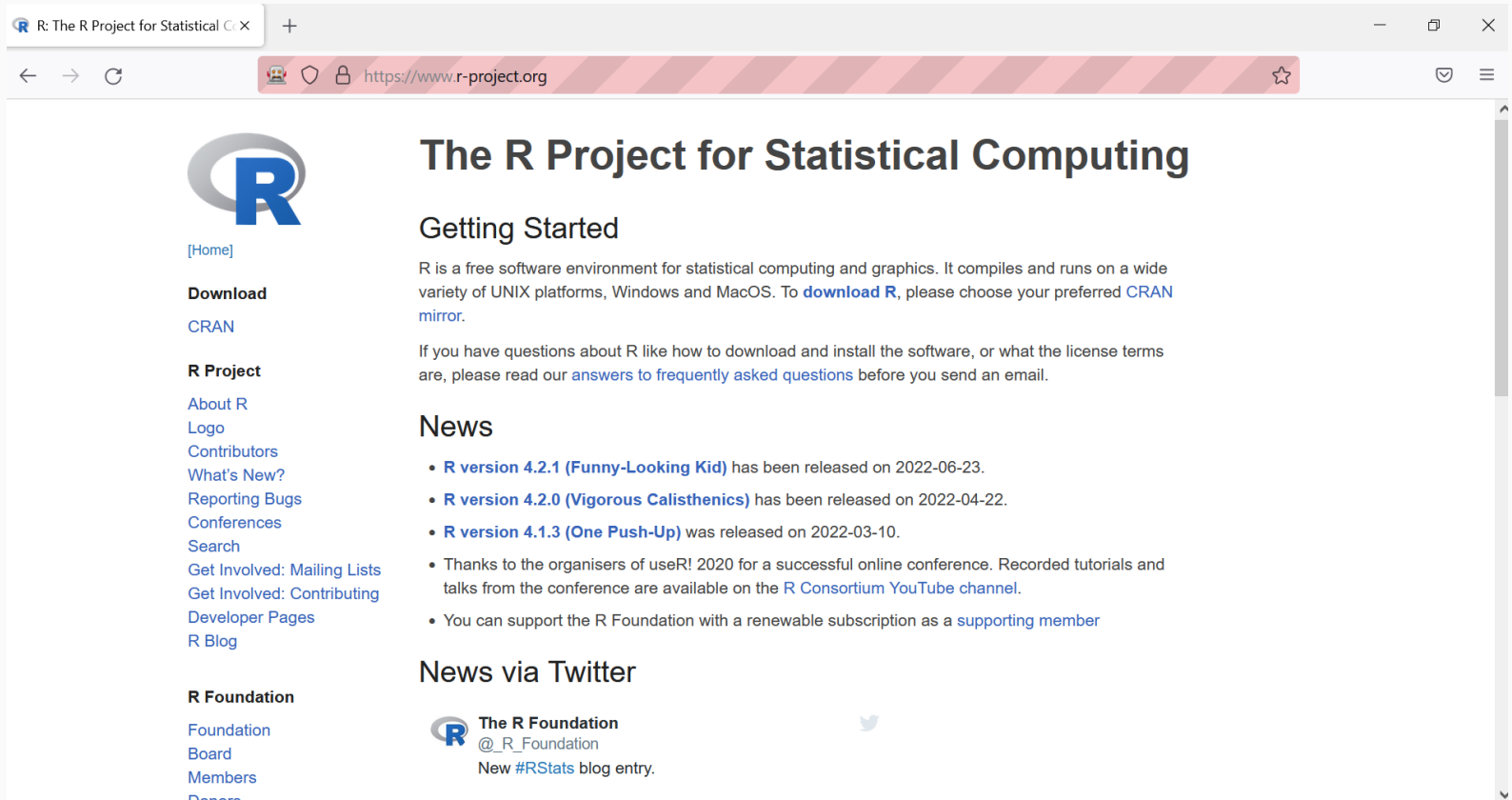
From now on, everything we do is calling `<function>()` starting with `remote_driver$`<sup>1</sup>.

**Objective:** get the list of core contributors to R located [here](#).

<sup>1</sup>: Or whatever you called it in the previous step

# Navigate

```
remote_driver$navigate("https://r-project.org")
```



The screenshot shows a web browser window displaying the official website of the R Project for Statistical Computing. The browser's address bar shows the URL <https://www.r-project.org>. The website features a large blue 'R' logo on the left, with a navigation menu including links to [Home], Download, CRAN, R Project, About R, Logo, Contributors, What's New?, Reporting Bugs, Conferences, Search, Get Involved: Mailing Lists, Get Involved: Contributing, Developer Pages, R Blog, R Foundation, Foundation, Board, Members, and Donors. The main content area is titled 'The R Project for Statistical Computing' and 'Getting Started'. It describes R as a free software environment for statistical computing and graphics, available on various platforms. It provides instructions on how to download R and choose a preferred CRAN mirror. A section titled 'News' lists recent releases: R version 4.2.1 (Funny-Looking Kid) released on 2022-06-23, R version 4.2.0 (Vigorous Calisthenics) released on 2022-04-22, and R version 4.1.3 (One Push-Up) released on 2022-03-10. It also mentions a successful online conference in 2020 and a supporting member option. A 'News via Twitter' section shows a tweet from The R Foundation (@\_R\_Foundation) about a new #RStats blog entry.

R: The R Project for Statistical Computing

← → ↻ <https://www.r-project.org> ☆

## The R Project for Statistical Computing

### Getting Started


R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

### News

- [R version 4.2.1 \(Funny-Looking Kid\)](#) has been released on 2022-06-23.
- [R version 4.2.0 \(Vigorous Calisthenics\)](#) has been released on 2022-04-22.
- [R version 4.1.3 \(One Push-Up\)](#) was released on 2022-03-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

### News via Twitter

 **The R Foundation**  
[@\\_R\\_Foundation](#)  
New [#RStats](#) blog entry.

# Click on "Contributors"

This requires two things:

1. find the element
2. click on it

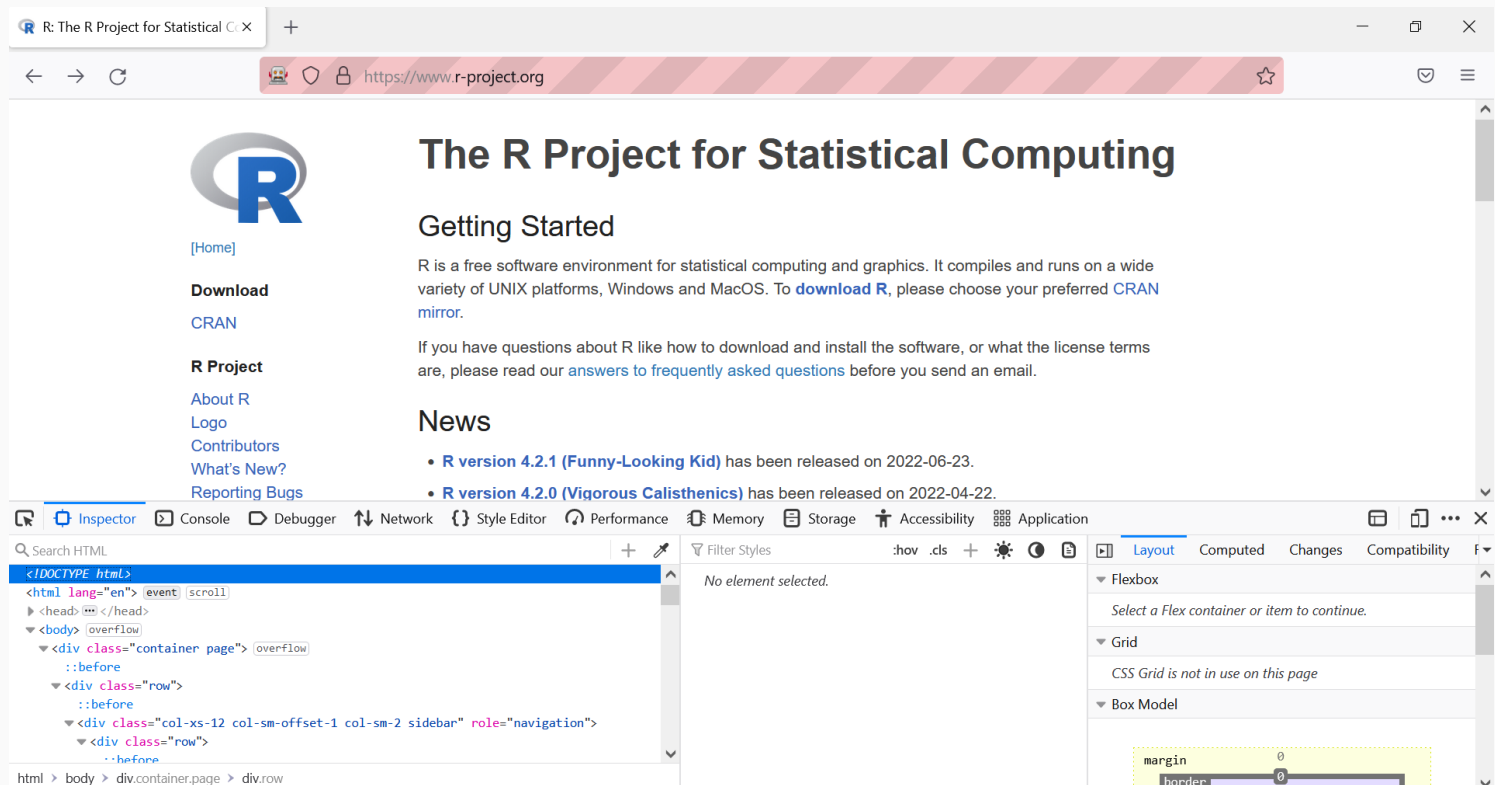
Humans -> eyes

Computers -> HTML/CSS

To find the element, we need to open the console to see the structure of the page.

Several ways to do it:

- right-click -> "Inspect"
- `Ctrl` + `Shift` + `C`



Then, hover the element we're interested in: the link "Contributors".



How can we find this with `RSelenium`?

```
?RSelenium::remoteDriver
```

-> `findElement`

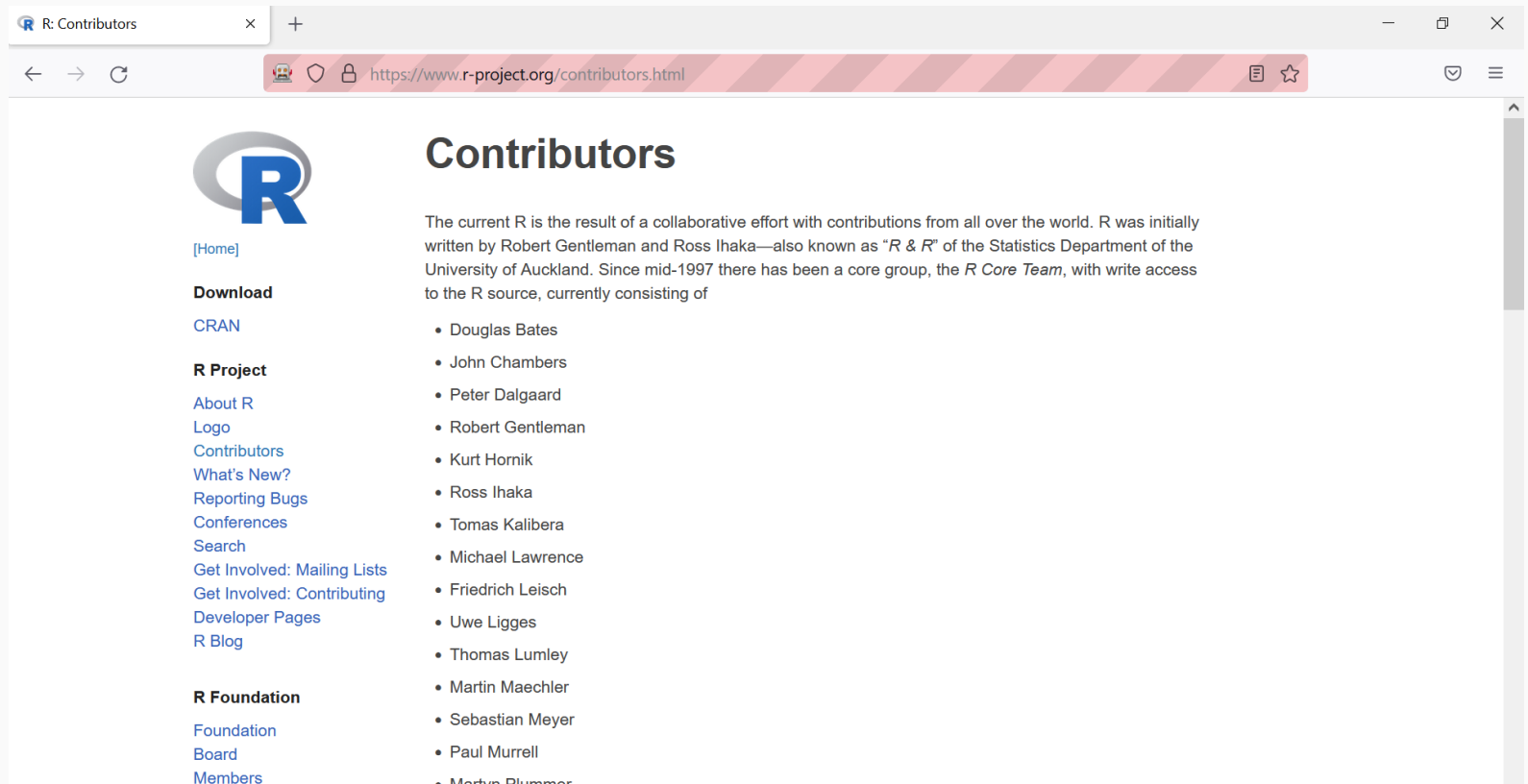
- class name ✗
- id ✗
- name ✗
- tag name ✗
- css selector ✓
- link text ✓
- partial link text ✓
- xpath ✓

All of these work:

```
remote_driver$  
  findElement("link text", "Contributors")$  
  clickElement()  
  
remote_driver$  
  findElement("partial link text", "Contributors")$  
  clickElement()  
  
remote_driver$  
  findElement("xpath", "/html/body/div/div[1]/div[1]/div/div[1]/ul/li[3]")$  
  clickElement()  
  
remote_driver$  
  findElement("css selector", "div.col-xs-6:nth-child(1) > ul:nth-child(1)")$  
  clickElement()
```



We are now on the right page!



The screenshot shows a web browser window with the title "R: Contributors". The address bar displays the URL "https://www.r-project.org/contributors.html". The page content includes the R logo, a sidebar with navigation links, and a main section titled "Contributors" with a paragraph of text and a list of names.

**R: Contributors**

[\[Home\]](#)

**Download**

[CRAN](#)

**R Project**

[About R](#)  
[Logo](#)  
[Contributors](#)  
[What's New?](#)  
[Reporting Bugs](#)  
[Conferences](#)  
[Search](#)  
[Get Involved: Mailing Lists](#)  
[Get Involved: Contributing](#)  
[Developer Pages](#)  
[R Blog](#)

**R Foundation**

[Foundation](#)  
[Board](#)  
[Members](#)

## Contributors

The current R is the result of a collaborative effort with contributions from all over the world. R was initially written by Robert Gentleman and Ross Ihaka—also known as “*R & R*” of the Statistics Department of the University of Auckland. Since mid-1997 there has been a core group, the *R Core Team*, with write access to the R source, currently consisting of

- Douglas Bates
- John Chambers
- Peter Dalgaard
- Robert Gentleman
- Kurt Hornik
- Ross Ihaka
- Tomas Kalibera
- Michael Lawrence
- Friedrich Leisch
- Uwe Ligges
- Thomas Lumley
- Martin Maechler
- Sebastian Meyer
- Paul Murrell
- Martin Plummer

Last step: obtain the HTML of the page.

```
remote_driver$getPageSource()
```

To read it with `rvest`:

```
x <- remote_driver$getPageSource()[[1]]  
rvest::read_html(x)
```

Do we read the HTML and extract the information in the same script?

**No!**

Rather, we save the HTML in an external file, and we will be able to access it in another script (and offline) to manipulate it as we want<sup>1</sup>.

```
write(x, file = "contributors.html")  
# Later and in another script  
rvest::read_html("contributors.html")
```

Click [here](#) to see the results.

<sup>1</sup>: Although, in this case, it wouldn't cost too much to treat it directly in the same script.

**A harder & real-life example**

The previous example was not a *dynamic* page: we could have used the link to the page and apply webscraping methods for static webpages.

Let's now dive into a more complex example, where RSelenium is the only way to scrape.

# Appendix

For reference, here's the code to extract the list of contributors:

```
library(rvest)

html <- read_html("contributors.html")

bullet_points <- html %>%
  html_elements(css = "div.col-xs-12 > ul > li") %>%
  html_text()

blockquote <- html %>%
  html_elements(css = "div.col-xs-12.col-sm-7 > blockquote") %>%
  html_text() %>%
  strsplit(., split = ", ")

blockquote <- blockquote[[1]] %>%
  gsub("\\r|\\n|\\.|and", "", .)

others <- html %>%
  html_elements(xpath = "/html/body/div/div[1]/div[2]/p[5]") %>%
  html_text() %>%
  strsplit(., split = ", ")

others <- others[[1]] %>%
  gsub("\\r|\\n|\\.|and", "", .)

all_contributors <- c(bullet_points, blockquote, others)
```

# Appendix

##	[1]	"Douglas Bates"	"John Chambers"	"Peter Dalgaard"	"Robert Gentleman"
##	[6]	"Ross Ihaka"	"Tomas Kalibera"	"Michael Lawrence"	"Friedrich Leisch"
##	[11]	"Thomas Lumley"	"Martin Maechler"	"Sebastian Meyer"	"Paul Murrell"
##	[16]	"Brian Ripley"	"Deepayan Sarkar"	"Duncan Temple Lang"	"Luke Tierney"
##	[21]	"Valerio Aimala"	"Suharto Anggono"	"Thomas Baier"	"Gabe Becker"
##	[26]	"Roger Biv"	"Ben Bolker"	"David Brahm"	"Göran Broström"
##	[31]	"Vince Carey"	"Saikat DebRoy"	"Matt Dowle"	"Brian D'Urso"
##	[36]	"Dirk Eddelbuettel"	"Claus Ekstrom"	"Sebastian Fischmeister"	"John Fox"
##	[41]	"Yu Gong"	"Gabor Grothendieck"	"Frank E Harrell Jr"	"Peter M Haverty"
##	[46]	"Robert King"	"Kjetil Kjernsmo"	"Roger Koenker"	"Philippe Lambert"
##	[51]	"Jim Lindsey"	"Patrick Lindsey"	"Catherine Loader"	"Gordon Maclean"
##	[56]	"John Maindonald"	"David Meyer"	"Ei-ji Nakama"	"Jens Oehlschägel"
##	[61]	"Richard O'Keefe"	"Hubert Palme"	"Roger D Peng"	"José C Pinheiro"
##	[66]	"Anthony Rossini"	"Jonathan Rougier"	"Petr Savicky"	"Günther Sawitzki"
##	[71]	"Arun Srinivasan"	"Detlef Steuer"	"Bill Simpson"	"Gordon Smyth"
##	[76]	"Terry Therneau"	"Rolf Turner"	"Bill Venables"	"Gregory R Warnes"
##	[81]	"Morten Welinder"	"James Wettenhall"	"Simon Wood"	"Achim Zeileis"
##	[86]	"David J Best"	"Richard Brent"	"Kevin Buhr"	"Michael A Covington"
##	[91]	"Robert Clevel,"	"G W Cran"	"C G Ding"	"Ulrich Drepper"
##	[96]	"J O Evans"	"David M Gay"	"H Frick"	"G W Hill"
##	[101]	"Eric Grosse"	"Shelby Haberman"	"Bruno Haible"	"John Hartigan"
##	[106]	"Trevor Hastie"	"Min Long Lam"	"George Marsaglia"	"K J Martin"
##	[111]	"C R Mckenzie"	"Jean McRae"	"Cyrus Mehta"	"Fionn Murtagh"
##	[116]	"Finbarr O'Sullivan"	"R E Odeh"	"William Patefield"	"Nitin Patel"
##	[121]	"D E Roberts"	"Patrick Royston"	"Russell Lenth"	"Ming-Jen Shyu"
##	[126]	"S G Springer"	"Supoj Sutanthavibul"	"Irma Terpenning"	"G E Thomas"
##	[131]	"Wai Wan Tsang"	"Berwin Turlach"	"Gary V Vaughan"	"Michael Wichura"
##	[136]	"M A Wong"			

[Back](#)