

Introduction to Web Scraping with R

Selenium: Case Study



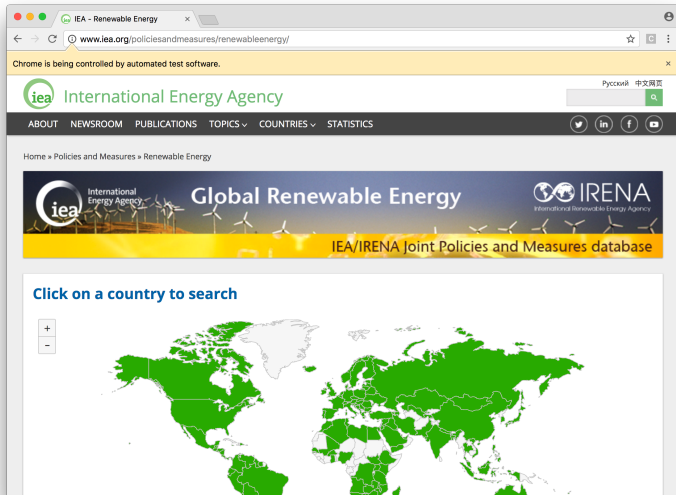
Simon Munzert | IPSDS

Selenium: Case Study

Example

Overview

- **goal:** scrape data from the IEA/RENA Joint Policies and Measures database
- URL: <http://www.iea.org/policiesandmeasures/renewableenergy/>
- the query form has multiple parameters
- the output comes in form of an HTML table, but content is injected into DOM



Example

Overview

- **goal:** scrape data from the IEA/RENA Joint Policies and Measures database
- URL: <http://www.iea.org/policiesandmeasures/renewableenergy/>
- the query form has multiple parameters
- the output comes in form of an HTML table, but content is injected into DOM

The screenshot shows a web browser window with the URL www.iea.org/policiesandmeasures/renewableenergy/. The page title is "IEA - Renewable Energy". A yellow banner at the top states "Chrome is being controlled by automated test software." The main content area is titled "Advanced search" and contains several filter sections:

- Countries:** + Regions, + Countries
- Policy Type:** + ☐ Economic Instruments, + ☐ Information and Education, + ☐ Policy Support, + ☐ Regulatory Instruments, + ☐ Research, Development and Deployment (RD&D), + ☐ Voluntary Approaches
- Renewable Energy Policy Target:** + ☐ Bioenergy, + ☐ Geothermal, ☐ Hydropower, + ☐ Multiple Renewable Energy Sources, + ☐ Ocean, + ☐ Solar, + ☐ Solar Thermal, + ☐ Wind
- Sector:** ☐ Electricity, ☐ Framework Policy, ☐ Heating and Cooling, ☐ Multi-sectoral Policy, ☐ Transport
- Effective between:** Select [dropdown] and [dropdown]
- Jurisdiction:** ☐ International, ☐ National, ☐ State/Regional, ☐ Municipal
- Policy Status:** ☐ Ended, ☐ In Force, ☐ Planned, ☐ Superseded, ☐ Under Review
- Size of Plant:** ☐ Large, ☐ Small
- Search by keyword(s):** [text input]
- ☐ Search only recently updated policies

At the bottom of the search section are two buttons: "RESET" and "SEARCH". Below the search section is a link for "Analytical resources".

Example

Overview

- **goal:** scrape data from the IEA/RENA Joint Policies and Measures database
- URL: <http://www.iea.org/policiesandmeasures/renewableenergy/>
- the query form has multiple parameters
- the output comes in form of an HTML table, but content is injected into DOM

IEA - Renewable Energy

www.iea.org/policiesandmeasures/renewableenergy/

Chrome is being controlled by automated test software.

Highlighted records constitute key elements of renewable energy policy framework

Found: 494 results. (Tip: sort columns by clicking on the column header)
[Perform another search](#)

Filter:

Show statistics timeline for Austria

Title	Country	Year	Policy Status	Policy Type	Policy Target
Energy Concept	Germany	2010	In Force	Regulatory Instruments, Policy Support>Strategic planning, Policy Support	Multiple RE Sources
National Renewable Energy Action Plan (NREAP)	France	2010	In Force	Policy Support>Strategic planning	Multiple RE Sources, Multiple RE Sources>All, Multiple RE Sources>Cooling, Multiple RE Sources>CHP, Multiple RE Sources>Heating, Multiple RE Sources>Power
Green innovation funding: the French programme of investments for the future	France	2010	In Force	Research, Development and Deployment (RD&D), Research, Development and Deployment (RD&D)>Demonstration project, Research, Development and Deployment (RD&D)>Research programme, Research, Development and Deployment (RD&D)>Research programme >Technology	Multiple RE Sources

Example

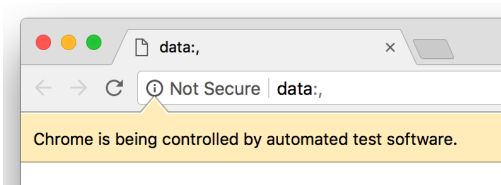
Setup

- load **RSelenium**
- check installed Java version (helpful for debugging if Selenium fails to launch)
- initiate SeleniumDriver and browser

R code

```
1 library(RSelenium)
2 system("java -version")
  java version "9"
  Java(TM) SE Runtime Environment (
  build 9+181)
  Java HotSpot(TM) 64-Bit Server VM (
  build 9+181, mixed mode)
3 rD <- rsDriver()
4 remDr <- rD[["client"]]
```

end



Example

R calls

- `navigate()` to URL
- page opens "automatically"

R code _____

```
url <- "http://www.iea.org/  
policiesandmeasures/renewableenergy/"  
remDr$navigate(url)
```

_____ end



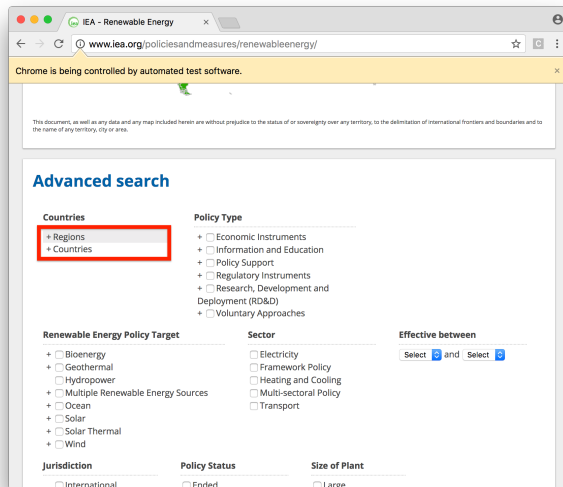
Example

R calls

- click on the "Regions" menu to unfold it
- extract XPath expression of element from Web Developer Tools
- pass XPath to `findElement()`, then click on it with `clickElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[1]/ul/li[1]/span'
regionsElem <- remDr$findElement(using = 'xpath', value = xpath)
openRegions <- regionsElem$clickElement()
_____ end
```



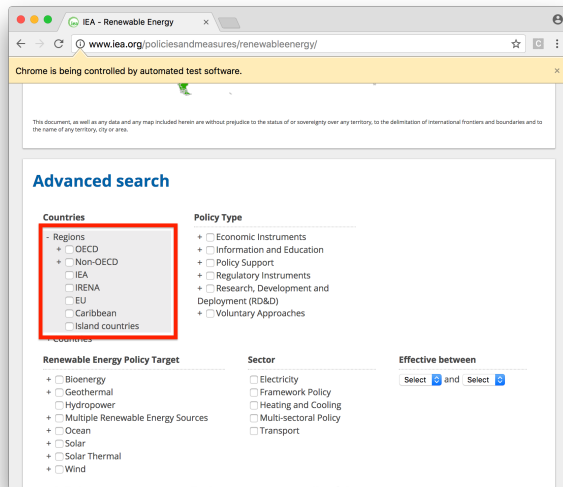
Example

R calls

- click on the "Regions" menu to unfold it
- extract XPath expression of element from Web Developer Tools
- pass XPath to `findElement()`, then click on it with `clickElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[1]/ul/li[1]/span'
regionsElem <- remDr$findElement(using = 'xpath', value = xpath)
openRegions <- regionsElem$clickElement()
end
```



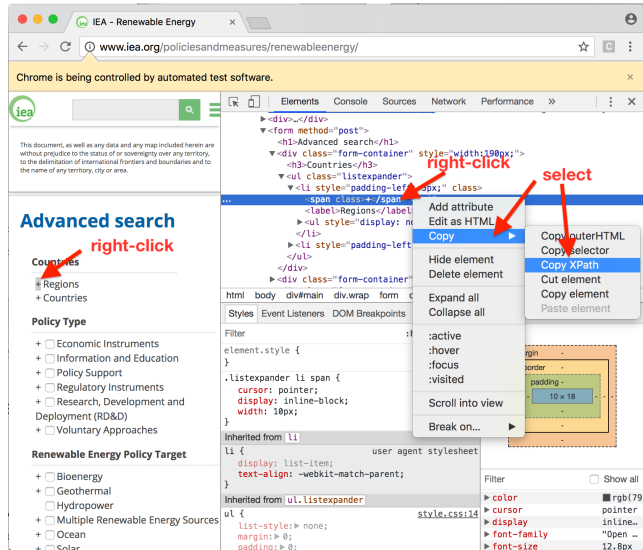
Example

R calls

- click on the "Regions" menu to unfold it
- extract XPath expression of element from Web Developer Tools
- pass XPath to `findElement()`, then click on it with `clickElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[1]/ul/li[1]/span'
regionsElem <- remDr$findElement(using = 'xpath', value = xpath)
openRegions <- regionsElem$clickElement()
_____ end
```



Example

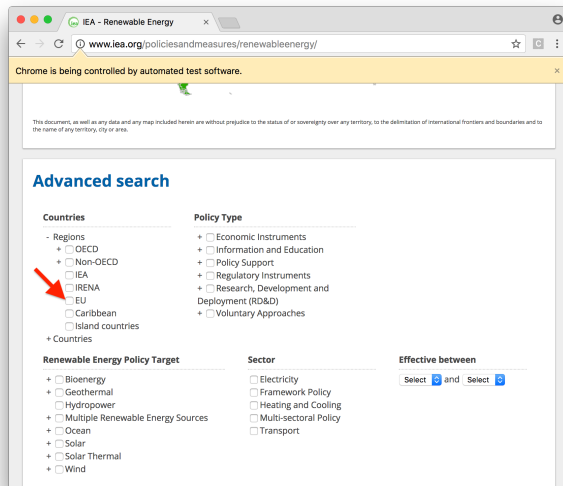
R calls

- select "EU" option
- extract XPath expression of element from Web Developer Tools
- pass XPath to `findElement()`, then click on it with `clickElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[1]/ul/li[1]/ul/li[5]/label/input'
euElem <- remDr$findElement(using = 'xpath', value = xpath)
selectEU <- euElem$clickElement()
```

end



Example

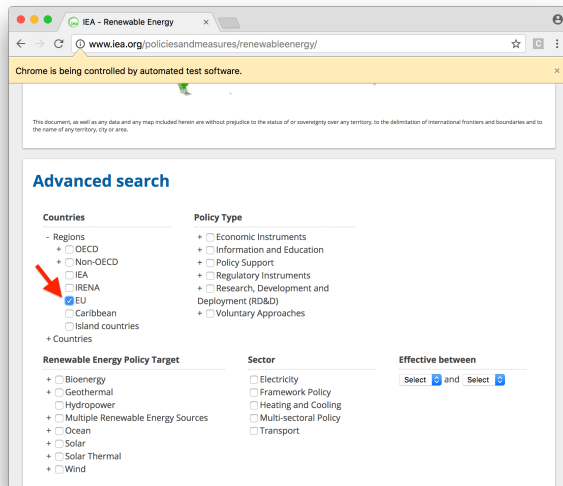
R calls

- select "EU" option
- extract XPath expression of element from Web Developer Tools
- pass XPath to `findElement()`, then click on it with `clickElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[1]/ul/li[1]/ul/li[5]/label/input'
euElem <- remDr$findElement(using = 'xpath', value = xpath)
selectEU <- euElem$clickElement()
```

end



Example

R calls

- set time frame
- pass XPath to `findElement()`, then click on it, then enter text with `sendKeysToElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[5]/select[1]'  
fromDrop <- remDr$findElement(using = 'xpath', value = xpath)  
clickFrom <- fromDrop$clickElement()  
writeFrom <- fromDrop$sendKeysToElement(list("2000"))
```

end

IEA - Renewable Energy

www.iea.org/policiesandmeasures/renewableenergy/

Chrome is being controlled by automated test software.

Advanced search

Countries

- + Regions
- + Countries

Policy Type

- + ☐ Economic Instruments
- + ☐ Information and Education
- + ☐ Policy Support
- + ☐ Regulatory Instruments
- + ☐ Research, Development and Deployment (RD&D)
- + ☐ Voluntary Approaches

Renewable Energy Policy Target

- + ☐ Bioenergy
- + ☐ Geothermal
- + ☐ Hydropower
- + ☐ Multiple Renewable Energy Sources
- + ☐ Ocean
- + ☐ Solar
- + ☐ Solar Thermal
- + ☐ Wind

Sector

- ☐ Electricity
- ☐ Framework Policy
- ☐ Heating and Cooling
- ☐ Multi-sectoral Policy
- ☐ Transport

Effective between

Select and Select

1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996

Jurisdiction

- ☐ International
- ☐ National
- ☐ State/Regional
- ☐ Municipal

Policy Status

- ☐ Ended
- ☐ In Force
- ☐ Planned
- ☐ Superseded
- ☐ Under Review

Size of Plant

- ☐ Large
- ☐ Small

☐ Search only recently updated

RESET SEARCH

Analytical resource

Find recent IEA publications

able energy here:

Example

R calls

- set time frame
- pass XPath to `findElement()`, then click on it, then enter text with `sendKeysToElement()`

R code

```
xpath <- '//*[@id="main"]/div/form/div[5]/select[2]'  
fromDrop <- remDr$findElement(using = 'xpath', value = xpath)  
clickFrom <- fromDrop$clickElement()  
writeFrom <- fromDrop$sendKeysToElement(list("2010"))
```

end

IEA - Renewable Energy

www.iea.org/policiesandmeasures/renewableenergy/

Chrome is being controlled by automated test software.

Advanced search

Countries

- + Regions
- + Countries

Policy Type

- + ☐ Economic Instruments
- + ☐ Information and Education
- + ☐ Policy Support
- + ☐ Regulatory Instruments
- + ☐ Research, Development and Deployment (RD&D)
- + ☐ Voluntary Approaches

Renewable Energy Policy Target

- + ☐ Bioenergy
- + ☐ Geothermal
- + ☐ Hydropower
- + ☐ Multiple Renewable Energy Sources
- + ☐ Ocean
- + ☐ Solar
- + ☐ Solar Thermal
- + ☐ Wind

Sector

- ☐ Electricity
- ☐ Framework Policy
- ☐ Heating and Cooling
- ☐ Multi-sectoral Policy
- ☐ Transport

Effective between

Select Select

- 1974
- 1975
- 1976
- 1977
- 1978
- 1979
- 1980
- 1981
- 1982
- 1983
- 1984
- 1985
- 1986
- 1987
- 1988
- 1989
- 1990
- 1991
- 1992
- 1993
- 1994
- 1995
- 1996

Jurisdiction

- ☐ International
- ☐ National
- ☐ State/Regional
- ☐ Municipal

Policy Status

- ☐ Ended
- ☐ In Force
- ☐ Planned
- ☐ Superseded
- ☐ Under Review

Size of Plant

- ☐ Large
- ☐ Small

Search by keyword(s)

☐ Search only recently updated policies

Analytical resources

Find recent IEA publications on renewable energy here:

Example

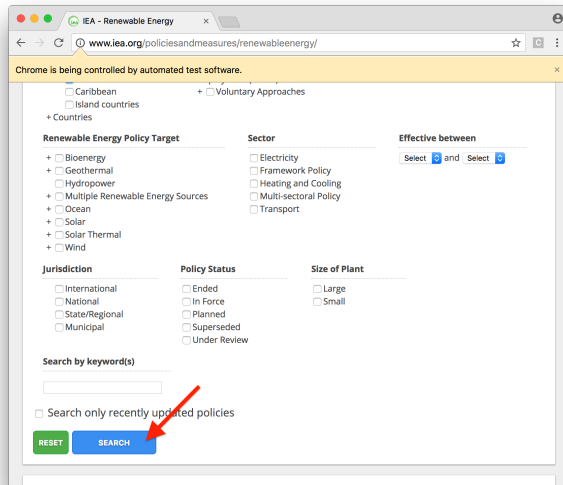
R calls

- click on search button
- extract XPath expression of element from Web Developer Tools
- pass XPath to `findElement()`, then click on it

R code

```
xpath <- '//*[@id="main"]/div/form/button[2]'  
searchElem <- remDr$findElement(using = 'xpath', value = xpath)  
resultsPage <- searchElem$clickElement() #  
click on button
```

end



Example

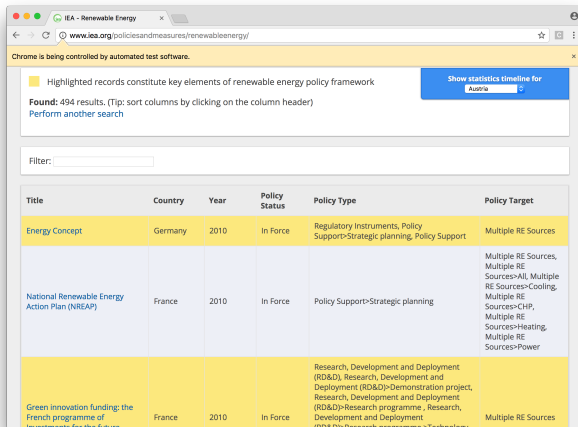
R calls

- finally, we're there!
- now take snapshot of the DOM with `getPageSource()` and store it as HTML with `write()`

R code _____

```
output <- remDr$getPageSource(header =  
TRUE)  
write(output[[1]], file = "iea-renewables.  
html")
```

_____ end



The screenshot shows a web browser window with the URL www.iea.org/policiesandmeasures/renewableenergy/. The page displays a table of renewable energy policy records. The table has columns for Title, Country, Year, Policy Status, Policy Type, and Policy Target. The first three rows are highlighted in yellow.

Title	Country	Year	Policy Status	Policy Type	Policy Target
Energy Concept	Germany	2010	In Force	Regulatory Instruments, Policy Support>Strategic planning, Policy Support	Multiple RE Sources
National Renewable Energy Action Plan (NREAP)	France	2010	In Force	Policy Support>Strategic planning	Multiple RE Sources, Multiple RE Sources>All, Multiple RE Sources>Cooling, Multiple RE Sources>CHP, Multiple RE Sources>Heating, Multiple RE Sources>Power
Green innovation funding: the French programme of investments for the future	France	2010	In Force	Research, Development and Deployment (RD&D), Research, Development and Deployment (RD&D)>Demonstration project, Research, Development and Deployment (RD&D)>Research programme, Research, Development and Deployment (RD&D)>Research programme>Technology	Multiple RE Sources

Example

Parsing HTML into data frame

- close the connection to Selenium server with `closeServer()`
- proceed with business as usual (`rvest` package)

R code

```
35 remDr$closeServer()
36 content <- read_html("iea-renewables.html", encoding = "utf8")
37 tabs <- html_table(content, fill = TRUE)
38 tab <- tabs[[1]]
39 "target")
40 tab[1,]
```

```
      title country year  status
      type
1 Energy Concept Germany 2010 In Force Regulatory Instruments, Policy Support>Strategic
planning, Policy Support
      target
1 Multiple RE Sources
```

Summary

Summary

- Selenium is an excellent tool to scrape data from dynamic, JavaScript-enriched webpages where ordinary scraping methods fail
- once the setup is established (which can be troublesome, as many software components have to work together), its use is pretty simple
- I do not recommend it as a substitute for every scraping task because it is too slow and unreliable for that purpose



By W. Oelen (CC BY-SA 3.0), <https://commons.wikimedia.org/w/index.php?curid=15369617>