Prof. Dr. Simon Munzert
Hertie School
munzert@hertie-school.org
https://simonmunzert.github.io/

A Primer to Web Scraping with R

Online Webinar
October 15–16, 2020

Outline

The rapid growth of the World Wide Web over the past two decades tremendously changed the way we share, collect and publish data. Firms, public institutions and private users provide every imaginable type of information and new channels of communication generate vast amounts of data on human behavior. What was once a fundamental problem for the social sciences—the scarcity and inaccessibility of observations—is quickly turning into an abundance of data. But how to efficiently collect data from the Internet with statistical software? In this two-day webinar, you will learn how to scrape content from static and dynamic web pages, connect to APIs from popular web services to read out and process user data, and set up automatically working scraper programs.

Schedule

Time	Topic
October 15, 2-2.50pm	Scraping static webpages
October 15, 3-3.50pm	Scraping dynamic webpages
October 16, 12-12.50pm	Tapping web APIs
October 16, 1-1.50pm	Scraping ethics and workflow

Workshop Flow

The workshop consists of two parts: The first is you, on your own, studying video and coding materials. The second is all of us, together, live and online on October 15 and 16. Here's how the individual parts work:

- One week before the first live session, you will be given access to pre-recorded video material and accompanying R scripts. You are expected to have watched the videos that accompany the sessions in advance. Furthermore, you are strongly encouraged to work through the scripts to come prepared (and ideally with questions) to the live sessions.
- The live sessions will provide the opportunity for interactive Q&A and collaborative code development. They will fully take place online and will only be streamed, not recorded.

Accompanying the sessions, you will be provided additional training material that you can use to use to practice and deepen the techniques learned. This part is optional and I cannot provide feedback on your solutions due to time constraints.

Prerequisites

Although no special knowledge of web technologies or programming languages is required, participants are expected to have applied knowledge of R. **If you consider yourself a beginner in R, this course might be not particularly useful and is probably frustrating for you.** I assume basic command in

- handling data structures (lists, data frames, vectors) with base R
- data manipulation with dplyr
- iterative programming using for loops and the apply() family (for old-school R people) or the purrr package

A topic that you might not know much about yet but that I expect you to become familiar with **before the workshop** is **regular expressions** and **string manipulation**. You will be given plenty of video and coding material on these as part of the pre-workshop package, but if you want to get started now already, you might want to check out the following sources (as you will see from these recommendations, we will prefer stringr over base R functions to work with strings):

- Regular expressions with stringr:
  https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html
- A stringr tidyverse overview:
  https://stringr.tidyverse.org/
- On strings in R4DS:
  https://r4ds.had.co.nz/strings.html
- A regex cheat sheet:
  https://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf
- Interactive ways to learn regular expressions:
  https://regexcrossword.com/, https://alf.nu/RegexGolf/

Texts and Materials

The workshop is accompanied by the following book:

Munzert, Simon, *Christian Rubba*, *Peter Meißner*, and *Dominic Nyhuis*, 2015: Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining. Chichester: John Wiley & Sons. Some things have changed since this book was published. In fact, watching the videos might just be more convenient than reading the book (although it involves listening to me a lot). Also, I will make sure to cover packages that are most up-to-date in the R environment. In addition, more materials will be made available online on the following GitHub repository (not live yet!):

https://github.com/hertie-data-science-lab/ds-workshop-webscraping

Supplemental Literature

Other useful texts on R and web technologies include:

- *Nolan, Deborah*, and *Duncan Temple Lang*, 2014: XML and Web Technologies for Data Sciences with R. New York: Springer.
- *Murrell, Paul*, 2009: Introduction to Data Technologies. Chapman & Hall/CRC.
- *Gandrud, Christopher*, 2015: Reproducible Research with R and RStudio. Chapman & Hall/CRC, 2nd Ed.
- *Wickham, Hadley*, 2014: Advanced R. Chapman & Hall/CRC.
- *Grolemund, Garrett*, and *Hadley Wickham*, 2016: R for Data Science. O'Reilly.

If you want to dig deeper into web and data technologies, you may want to consider the following books:

- *Beaulieu, Alan*, 2009: Learning SQL. Sebastopol, CA: O'Reilly.
- *Cerami, Ethan*, 2002: Web Services Essentials. Sebastopol, CA: O'Reilly.
- *Holdener III, Anthony T.*, 2008: Ajax: The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Gourley, David*, and *Brian Totty*, 2002: HTTP. The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Crockford, Douglas*, 2008: JavaScript: The Good Parts. Sebastopol, CA: O'Reilly.