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

A Primer to Web Scraping with R

Max-Planck-Institut für Gesellschaftsforschung
31.01./01.02.2018

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. This turn of events does not come without problems. For example, traditional techniques for collecting and analyzing data may no longer suffice to overcome the tangled masses of data. One consequence of the need to make sense of such data has been the inception of 'data scientists', who sift through data and are greatly sought after by research and business alike.

But how to efficiently collect data from the Internet, retrieve information from social networks, search engines, and dynamic web pages, tap web services, and, finally, process the collected data with statistical software? We will learn how to scrape content from static and dynamic web pages, connect to APIs from popular web services such as Twitter to read out and process user data, and set up automatically working scraper programs. The sessions are hands-on; we will practice every step of the process with R using various examples.

Course Objectives

By the end of the course, you will. . .

- be able to assess the feasibility of conducting scraping projects in diverse settings,
- be able to scrape information from static and dynamic websites as well as web APIs, and
- be able to tackle current research questions with original data in your own work.

Instructor

Simon Munzert is Lecturer in Political Data Science at the Hertie School of Governance. He received his doctoral degree in Political Science from the University of Konstanz. His research interests include measuring and forecasting public opinion, political representation, and the use of new media in society. He is author of the textbook *Automated Data Collection with R* (Wiley). His research has been published in the *American Political Science Review, Political Analysis*, and *Political Science Research and Methods*, among others. Furthermore, he is an enthusiastic user of the statistical software R.

Schedule

Date	Time	Topic
	09:30 -10:30	Introduction; a first encounter with the Web using R
	10:30 - 10:45	*Coffee break*
	10:45 -12:15	Regular expressions and string manipulation
	12:15 - 13:30	*Lunch break*
July 16	13:30 -15:30	Scraping static webpages
	15:30 - 15:45	*Coffee break*
	15:45 -17:30	Advanced scraping of static webpages
	09:30 -10:30	Scraping dynamic webpages
	10:30 - 10:45	*Coffee break*
	10:45 -12:15	Tapping APIs and gathering social media data
	12:15 - 13:30	*Lunch break*
Feb 01	13:30 -15:30	Legal and ethical issues in web scraping
	15:30 - 15:45	*Coffee break*
	15:45 -17:30	Scraping workflow and tricks of the trade

Prerequisites and Software

I strongly recommend to bring your own laptop. Furthermore, although no special knowledge of web technologies or programming languages is required, participants are expected to have applied knowledge of R. Ideally, areas you are familiar with include

- data structures and basic vocabulary
- data import and export with readr and haven or rio
- data manipulation with dplyr
- writing own functions

Before the course starts, you should make several preparations:

1. make sure that the newest version of R (available here) is installed on your computer
2. install the newest stable version of *RStudio* (available here)
3. install the needed packages as outlined on the GitHub repository (see below)

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. 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:

https://github.com/simonmunzert/rscraping-mpi-cologne-2019

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.