Simon Munzert
Department of Politics and Public Administration
University of Konstanz
simon.munzert@uni.kn
http://r-datacollection.com

A Primer to Web Scraping with R

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 collect data from the Internet, retrieve information from social networks, search engines and dynamic web pages, how to tap web services and finally, clean and process the collected data directly in R? In this two-day workshop, we will learn about the basics of Internet architecture and web scraping practice with R. The sessions are hands-on; we will practice every step of the process with R using various examples.

Software

I strongly recommend to bring your own laptop. Further, although no special knowledge of web technologies or programming languages is required, participants are expected to have applied knowledge of R. Areas you should be familiar include

- data structures and basic vocabulary
- basic data import and export
- data manipulation with base R commands and/or using Hadley Wickham's plyr and dplyr packages
- writing own functions

Before the course starts, you should make several preparations:

1. make sure that the newest version of R (currently 3.1.2; available here) is installed on your computer
2. install the newest stable version of *RStudio* (available here)
3. install the following packages:
   `pkgs <- c('RCurl', 'XML', 'stringr', 'jsonlite', 'httr', 'rvest', 'devtools', 'RSelenium', 'plyr', 'dpylr', 'wikipediatrend', 'twitteR', 'streamR')`
4. install the *Chrome* (from here) and *Firefox* (from here) browsers
5. install *Java* (from here)

Texts

The workshop is accompanied by the following book:

Munzert, Simon, *Christian Rubba*, *Peter Meißner*, und *Dominic Nyhuis*, 2015: Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining. Chichester: John Wiley & Sons.

Outline

Time	Topic
26.03.2015, 09:00 -12:00	Introduction; a first encounter with the Web using R
26.03.2015, 13:15 -15:15	Scraping with regular expressions
26.03.2015, 15:45 -18:00	Scraping via XPath
27.03.2015, 09:00 -12:00	Social Media, APIs and JSON
27.03.2015, 13:15 -15:15	Scraping data from AJAX-enriched webpages
27.03.2015, 15:45 -18:00	Workflow, scraping etiquette, and tricks of the trade

Supplemental Literature

Other useful texts on R and web technologies include:

- *Nolan, Deborah*, und *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*, 2013: Reproducible Research with R and RStudio. Chapman & Hall/CRC.
- *Wickham, Hadley*, 2014: Advanced R. Chapman & Hall/CRC.

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.
- *Flanagan, David*, 2011: JavaScript: The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Holdener III, Anthony T.*, 2008: Ajax: The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Gourley, David*, und *Brian Totty*, 2002: HTTP. The Definitive Guide. Sebastopol, CA: O'Reilly.
- *Crockford, Douglas*, 2008: JavaScript: The Good Parts. Sebastopol, CA: O'Reilly.