Collecting Web Data with R

2022 Summer School in Social Research Methods

27 June – 1 July 2022 (week 2)

Hauke Licht*

last updated: 10 June 2022

Overview

*hauke.licht@wiso.uni-koeln.de

Course summary

The increasing availability of large amounts of online data enables new types of research in the social sciences. Over the past years, a variety of data — whether election results, press releases, parliamentary speeches or social media content — has become available online. Although these data has become easier to find, its extraction and reshaping into formats ready for downstream analyses can be challenging. This makes web data collection and cleaning skills essential for researchers.

The goal of this course is to equip participants with the R programming skills necessary to gather online data and process them into formats they can use in their research.

To get the most of the course, participants should have some prior experience with R and be willing to engage with different web technologies.

Participants will learn

- about the characteristics of web data;
- how to extract via Application Programmer Interfaces (APIs), inlcuding those maintained by popular social media platforms such as Twitter;
- how to scrape content from different types of webpages; as well as
- important techniques for cleaning and reshaping web and social media data for downstream analysis.

Course details

Instructor

Hauke Licht is a post-doctoral researcher at the Cologne Center for Comparative Politics, University of Cologne, and has received his PhD from the University of Zurich. He develops and applies text-as-data methods to study political competition and democratic representation with a strong focus on cross-lingual applications. In this research, he frequently relies on collecting textual data at scale by applying different web scraping techniques.

Learning objectives

By the end of the course participants will:

- Know the most important characteristics of web data, including webpage content and social media data.
- Gain an understanding of a variety of scraping scenarios: APIs, static pages, and dynamic pages.
- Be able to parse, clean and process data collected from the web.
- Be able to write reproducible and robust code for web scraping tasks.

Course materials

Disclaimer: I build on materials originally developed by Theresa Gessler we continuously update for our co-taught course at the GESIS Fall Seminar.

Course materials can be found on Github and on *Brightspace*:

- GitHub: https://github.com/haukelicht/methodsnet2022-webscraping-course
- Brightspace: https://brightspace.ru.nl/d2l/home/349712

In addition, I provide interactive R tutorials (also co-developed with Theresa Gessler) that are available through the `learn2scrape` R package (see Listing A.2 in the Appendix for installation instructions). These tutorials complement the course but installing and using the package is *not* a prerequisite.

Organization

We meet **daily** from 27 June to 1 July 2022 for **two 1.5 hour sessions**:

- Monday, 27 June 2022: 10:00 - 11:30 and 11:45 - 13:00 CEST.
- Tuesday - Friday, 28 June - 1 July 2022: 9:00 - 10:30 and 11:00 - 12:30

The course will be organized as a mixture of lectures and exercises. In the lectures, we will focus on explaining core concepts and techniques in web scraping. In the exercises, participants will apply their newly acquired knowledge and I will be available for consultation and support.

We will meet **in person** at `tba`.

Prerequisites

Participants should

- have basic knowledge of and some experience with using the R programming language
- be willing to engage with different web technologies

Participants should install the following programs on their personal computers:

- R (I recommend version $\geq$ 4.0.0)
- RStudio (or a comparable R interface/IDE)
- the Google Chrome and Firefox web browsers

They should install the following **R packages** (see Listing A.1 in the Appendix):

- for web data collection: `httr` ($\geq$ 1.3.0), `xml2` ($\geq$ 1.3.0), `rvest` ($\geq$ 1.0.0), `RSelenium` ($\geq$ 1.7.0), `rtweet` ($\geq$ 0.7.0) `academictwitteR` ($\geq$ 0.3.1)
- for data input/output: `jsonlite` ($\geq$ 1.7.0), `readr` ($\geq$ 1.4.0)
- for data wrangling: `dplyr` ($\geq$ 1.0.0), `tidyr` ($\geq$ 1.1.0), `purrr` ($\geq$ 0.3.0)
- for text wrangling: `stringr` ($\geq$ 1.4.0)
- for interactive tutorials (optional!): `learn2scrape` (see Listing A.2 in the Appendix)

Overview of sessions

Before the course

Before the first day of our course, please

- please complete the pre-course survey at https://forms.gle/DNyvNxaprWawrrsY8
- apply for a Twitter standard developer **and** academic track account

I also recommend you go through tutorials "001-tutorial-how-to" and "002-r-basics" in the `learn2scrape` R package (if you have already programmed in R, this won't take more than 30 minutes) so you will be up to speed when we get into R programming.

Day 1 (Jule 27): Introduction to web data

Preparation

Before we meet, please

- work through the following sections in the W3 school's XML tutorial: "Introduction", "How to use", "Tree", "Syntax", "Elements", and "Attributes" (this won't take more than 45 minutes)
- work through the following sections in the W3 school's HTML tutorial: "Introduction", "Basic", "Elements", "Attributes", and "Links" (this won't take more than 45 minutes)
- watch this video about HTTP
- watch this video about Web architectures (until about 9:26)

Session content

We will cover what web scraping is and how it can be used in social science and digital humanities research. Participants will be asked to share their expectations of the course and how they plan to use web scraping in their research.

We will then introduce most fundamental concepts including HTTP, APIs, and the XML and HTML formats. Finally, we will discuss how websites are commonly organized.

In the practical parts of the session, we will first ensure that all participants have a working R setup (incl. a Twitter developer account). If time permits, we will then also walk through a series of coding exercises designed to ensure that all participants are comfortable with basic R programming concepts and techniques.

Readings

! M. J. Salganik (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, chapter 1 (introduction, pp. 1–5) and 2 (observing behavior, pp. 13–41)

D. M. J. Lazer et al. (2020). "Computational Social Science: Obstacles and Opportunities". In: *Science* 369.6507, pp. 1060–1062. DOI: 10.1126/science.aaz8170. pmid: 32855329

Note: Exclamation marks ("!") in front of references mark mandatory readings.

Day 2 (Jule 28): APIs and social media data

Preparation

Before we meet, please

- install/update the following R packages: `httr` ($\geq$ 1.3.0), `rtweet` ($\geq$ 0.7.0), `academictwitteR` ($\geq$ 0.3.1), `jsonlite` ($\geq$ 1.7.0), `xml2` ($\geq$ 1.3.0)
- skim through the "Getting started with `httr`" vignette
- read H. Wickham (2020). *Managing Secrets*. URL: https://cran.r-project.org/web/packages/httr/vignettes/secrets.html and implement one of the recommend best practices to securely handling secrets for your Twitter credentials
- follow this vignette to obtain Twitter API credentials and to make them accessible in R as `rtweet` access token
- follow this vignette to obtain access to Twitter's research track and to make this access usable in R with `academictwitteR`

Session content

Building on the content discussed on Day 1, we will deepen our understanding of APIs. We will first introduce the `httr` R package, show how to use it to send API requests, and discuss how to work with different content types returned by APIs such as JSON and XML. We will also discuss authentication, pagination, and API rate limits.

We will then use Twitter as an examples to show how to extract social media data and use this example to discuss the specific challenges of social media research.

Readings

! D. Freelon (2018). "Computational Research in the Post-API Age". In: *Political Communication* 35.4, pp. 665–668. DOI: 10.1080/10584609.2018.1477506

! A. Bruns (2019). "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research". In: *Information, Communication & Society* 22.11, pp. 1544–1566. DOI: 10.1080/1369118X.2019.1637447

M. Mancosu and F. Vegetti (2020). "What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data". In: *Social Media + Society* 6.3, p. 2056305120940703. DOI: 10.1177/2056305120940703

A. Halavais (2019). "Overcoming Terms of Service: A Proposal for Ethical Distributed Research". In: *Information, Communication & Society* 22.11, pp. 1567–1581. DOI: 10.1080/1369118X.2019.1627386

M. H. Ribeiro, K. Gligorić, M. Peyrard, F. Lemmerich, M. Strohmaier, and R. West (2021). *Sudden Attention Shifts on Wikipedia During the COVID-19 Crisis*. arXiv: 2005.08505 [cs]. URL: http://arxiv.org/abs/2005.08505

Day 3 (June 29): Scraping static webpages

Preparation

Before we meet, please

- install/update `rvest` ($\geq$ 1.0.0)

Session content

We will learn how to extract data from static websites. Building on what we have learned about HTML (Day 1), we will cover how to systematically extract web data using the `rvest` R package, including a discussion of CSS selectors and the Xpath method to navigate the HTML tree.

Specifically, we will cover

1. how to extract HTML text and attributes as well as other data fro tables and images from web pages
2. how to automatically navigate between and scrape multiple pages of a websites.

Day 4 (June 30): Scraping dynamic webpages

Preparation

Before we meet, please

- install/update `RSelenium` ($\geq$ 1.7.0)

Session content

We will discuss how to scrape dynamic websites. We will first explain what makes a page "dynamic" and show how to recognize dynamic web elements in the wild.

We will then introduce the `RSelenium` package and show how it enables systematic interaction with dynamic web elements. This will include how to instantiate a web driver in R (Google Chrome), how to find web elements, how to navigate dynamic elements (e.g., accordion elements), how to switch between windows (e.g., a main page and a pop-up), and how to automatically download files.

Day 5 (July 1): Advanced topics and web scraping ethics

Session content

We will begin with a recap of what we have learned during the previous four days and collect the best practices that have been taught during the first four days.

Depending on participants' own plans, we can address advanced topics in web scraping, including web sessions, user agents, proxies, login, "iframes", and other topics participants might be interested in. We can also cover advanced techniques for handling webpage content, including regular expressions.

We will also return to the topic of ethics in web scraping.

Recommended literature

Big Data

M. J. Salganik (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press

S. González-Bailón (2017). *Decoding the Social World: Data Science and the Unintended Consequences of Communication*. MIT Press

C. O'neil (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown

About Computational Social Science (CSS)

Y. Theocharis and A. Jungherr (2021). "Computational Social Science and the Study of Political Communication". In: *Political Communication* 38.1-2, pp. 1–22. DOI: 10.1080/10584609.2020.1833121

A. Edelmann, T. Wolff, D. Montagne, and C. A. Bail (2020). "Computational Social Science and Sociology". In: *Annual Review of Sociology* 46.1, pp. 61–81. DOI: 10.1146/annurev-soc-121919-054621

D. M. J. Lazer et al. (2020). "Computational Social Science: Obstacles and Opportunities". In: *Science* 369.6507, pp. 1060–1062. DOI: 10.1126/science.aaz8170. pmid: 32855329

H. Wallach (2018). "Computational Social Science Computer Science + Social Data". In: *Communications of the ACM* 61.3, pp. 42–44. DOI: 10.1145/3132698

A. Jungherr and Y. Theocharis (2017). "The Empiricist's Challenge: Asking Meaningful Questions in Political Science in the Age of Big Data". In: *Journal of Information Technology & Politics* 14.2, pp. 97–109. DOI: 10.1080/19331681.2017.1312187

H. Margetts (2017). "The Data Science of Politics". In: *Political Studies Review* 15.2, pp. 201–209. DOI: 10.1177/1478929917693643

H. Wallach (2016). "Computational Social Science: Toward a Collaborative Future". In: *Computational Social Science: Discovery and Prediction*. Ed. by R. M. Alvarez. Analytical Methods for Social Research. Cambridge: Cambridge University Press, pp. 307–316. DOI: 10.1017/CBO9781316257340.014

J. Grimmer (2015). "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together". In: *PS: Political Science & Politics* 48.1, pp. 80–83. DOI: 10.1017/S1049096514001784

B. L. Monroe, J. Pan, M. E. Roberts, M. Sen, and B. Sinclair (2015). "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political

Science". In: *PS: Political Science & Politics* 48.1, pp. 71–74. DOI: 10.1017/S1049096514001760

J. Nagler and J. A. Tucker (2015). "Drawing Inferences and Testing Theories with Big Data". In: *PS: Political Science & Politics* 48.1, pp. 84–88. DOI: 10.1017/S1049096514001796

J. W. Patty and E. M. Penn (2015). "Analyzing Big Data: Social Choice and Measurement". In: *PS: Political Science & Politics* 48.1, pp. 95–101. DOI: 10.1017/S1049096514001814

S. A. Golder and M. W. Macy (2014). "Digital Footprints: Opportunities and Challenges for Online Social Research". In: *Annual Review of Sociology* 40.1, pp. 129–152. DOI: 10.1146/annurev-soc-071913-043145

M. Strohmaier and C. Wagner (2014). "Computational Social Science for the World Wide Web". In: *IEEE Intelligent Systems* 29.5, pp. 84–88. DOI: 10.1109/MIS.2014.80

Resources for R

B. Anderson, R. Severson, and N. Good (2020). "R Programming for Research". Ebook. Ebook. URL: https://geanders.github.io/RProgrammingForResearch/index.html

C. Gandrud (2018). *Reproducible Research with R and RStudio*. Chapman and Hall/CRC

H. Wickham and G. Grolemund (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc.

H. Wickham (2019). *Advanced r*. chapman and hall/CRC

Y. Xie, J. J. Allaire, and G. Grolemund (2018). *R Markdown: The Definitive Guide*. CRC Press

Resources for web scraping

S. Munzert, C. Rubba, P. Meißner, and D. Nyhuis (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons

Z. C. Steinert-Threlkeld (2018). *Twitter as Data*. Cambridge University Press

Scraping Ethics

F. Gilardi, L. Baumgartner, et al. (2021). "Building Research Infrastructures to Study Digital Technology and Politics: Lessons from Switzerland". In: *Political Science & Politics* forthcoming, p. 10

M. Mancosu and F. Vegetti (2020). "What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data". In: *Social Media + Society* 6.3, p. 2056305120940703. DOI: 10.1177/2056305120940703

G. King and N. Persily (2020). "A New Model for Industry–Academic Partnerships". In: *PS: Political Science & Politics* 53.4, pp. 703–709. DOI: 10.1017/S1049096519001021

A. Bruns (2019). "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research". In: *Information, Communication & Society* 22.11, pp. 1544–1566. DOI: 10.1080/1369118X.2019.1637447

C. Puschmann (2019). "An End to the Wild West of Social Media Research: A Response to Axel Bruns". In: *Information, Communication & Society* 22.11, pp. 1582–1589. DOI: 10.1080/1369118X.2019.1646300

A. Halavais (2019). "Overcoming Terms of Service: A Proposal for Ethical Distributed Research". In: *Information, Communication & Society* 22.11, pp. 1567–1581. DOI: 10.1080/1369118X.2019.1627386

D. Freelon (2018). "Computational Research in the Post-API Age". In: *Political Communication* 35.4, pp. 665–668. DOI: 10.1080/10584609.2018.1477506

Applied articles

Below, you can find a list of articles using web-scraped data that we enjoyed reading because they tackle important questions in new ways, have good ways to measure phenomena, or they reflect on important aspects of applied social science research. This collection is neither complete nor in any way representative of research with web-scraped data. There is no need to read everything on this list — but feel free to skim through some of the articles that sound interesting if you want to see some use-cases.

F. Pradel (2021). "Biased Representation of Politicians in Google and Wikipedia Search? The Joint Effect of Party Identity, Gender Identity and Elections". In: *Political Communication* 38.4, pp. 447–478. DOI: 10.1080/10584609.2020.1793846

H. Le, R. Maragh, B. Ekdale, A. High, T. Havens, and Z. Shafiq (2019). "Measuring Political Personalization of Google News Search". In: *The World Wide Web Conference*. WWW '19. New York, NY, USA: Association for Computing Machinery, pp. 2957–2963. DOI: 10.1145/3308558.3313682

V. Chykina and C. Crabtree (2018). "Using Google Trends to Measure Issue Salience for Hard-to-Survey Populations". In: *Socius* 4, p. 2378023118760414. DOI: 10.1177/2378023118760414

S. Göbel and S. Munzert (2018). "Political Advertising on the Wikipedia Marketplace of Information". In: *Social Science Computer Review* 36.2, pp. 157–175. DOI: 10.1177/0894439317703579

W. R. Hobbs and M. E. Roberts (2018). "How Sudden Censorship Can Increase Access to Information". In: *American Political Science Review* 112.3, pp. 621–636. DOI: 10.1017/S0003055418000084

J. D. Tjaden, C. Schwemmer, and M. Khadjavi (2018). "Ride with Me—Ethnic Discrimination, Social Markets, and the Sharing Economy". In: *European Sociological Review* 34.4, pp. 418–432. DOI: 10.1093/esr/jcy024

A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson (2017). "Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. New York, NY, USA: Association for Computing Machinery, pp. 1914–1933. DOI: 10.1145/2998181.2998327

G. King, B. Schneer, and A. White (2017). "How the News Media Activate Public Expression and Influence National Agendas". In: *Science* 358.6364, pp. 776–780. DOI: 10.1126/science.aao1100. pmid: 29123065

S. González-Bailón and N. Wang (2016). "Networked Discontent: The Anatomy of Protest Campaigns in Social Media". In: *Social Networks* 44, pp. 95–104. DOI: 10.1016/j.socnet.2015.07.003

J. Penney (2016). *Chilling Effects: Online Surveillance and Wikipedia Use*. SSRN Scholarly Paper ID 2769645. Rochester, NY: Social Science Research Network. URL: https://papers.ssrn.com/abstract=2769645

D. H. Chae et al. (2015). "Association between an Internet-Based Measure of Area Racism and Black Mortality". In: *PLOS ONE* 10.4, e0122963. DOI: 10.1371/journal.pone.0122963

A. Datta, M. C. Tschantz, and A. Datta (2015). "Automated Experiments on Ad Privacy Settings". In: *Proceedings on Privacy Enhancing Technologies* 2015.1, pp. 92–112. DOI: 10.1515/popets-2015-0007

A. Street, T. A. Murray, J. Blitzer, and R. S. Patel (2015/ed). "Estimating Voter Registration Deadline Effects with Web Search Data". In: *Political Analysis* 23.2, pp. 225–241. DOI: 10.1093/pan/mpv002

S. Stephens-Davidowitz (2014). "The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data". In: *Journal of Public Economics* 118, pp. 26–40. DOI: 10.1016/j.jpubeco.2014.04.010

A. Hannak et al. (2013). "Measuring Personalization of Web Search". In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. New York, NY, USA: Association for Computing Machinery, pp. 527–538. DOI: 10.1145/2488388.2488435

S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno (2011). "The Dynamics of Protest Recruitment through an Online Network". In: *Scientific Reports* 1.1 (1), p. 197. DOI: 10.1038/srep00197

Applications with Social Media Data

J. Beltran, A. Gallego, A. Huidobro, E. Romero, and L. Padró (2021). "Male and Female Politicians on Twitter: A Machine Learning Approach". In: *European Journal of Political Research* 60.1, pp. 239–251. DOI: 10.1111/1475-6765.12392

guinaudeau_fifteen_2021

F. Gilardi, T. Gessler, M. Kubli, and S. Müller (2021). "Social Media and Political Agenda Setting". In: *Political Communication*, pp. 1–22

P. Barberá and Z. C. Steinert-Threlkeld (2020). "How to Use Social Media Data for Political Science Research". In: *The SAGE Handbook of Research Methods in Political Science and International Relations. London. Sage*, pp. 404–423

S. Shugars and N. Beauchamp (2019). "Why Keep Arguing? Predicting Engagement in Political Conversations Online". In: *SAGE Open* 9.1, p. 2158244019828850. DOI: 10.1177/2158244019828850

C. A. Bail et al. (2018). "Exposure to Opposing Views on Social Media Can Increase Political Polarization". In: *Proceedings of the National Academy of Sciences* 115.37, pp. 9216–9221

S. Stier, A. Bleier, H. Lietz, and M. Strohmaier (2018). "Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter". In: *Political Communication* 35.1, pp. 50–74. DOI: 10.1080/10584609.2017.1334728

G. King, B. Schneer, and A. White (2017). "How the News Media Activate Public Expression and Influence National Agendas". In: *Science* 358.6364, pp. 776–780. DOI: 10.1126/science.aao1100. pmid: 29123065

K. Munger (2017). "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment". In: *Political Behavior* 39.3, pp. 629–649. DOI: 10.1007/s11109-016-9373-5

A. Jungherr, H. Schoen, O. Posegga, and P. Jürgens (2017). "Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support". In: *Social Science Computer Review* 35.3, pp. 336–356. DOI: 10.1177/0894439316631043

J. Burgess and A. Bruns (2015). "Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn". In: *Compromised data: From social media to big data* 95

Appendix

Listing A.1: R code to install required packages.

```r
required_packages <- c(
    "httr" = "1.3.0",
    "xml2" = "1.3.0",
    "rvest" = "1.0.0",
    "RSelenium" = "1.7.0",
    "rtweet" = "0.7.0",
    "academictwitterR" = "0.3.1",
    "jsonlite" = "1.7.0",
    "readr" = "1.4.0",
    "dplyr" = "1.0.0",
    "tidyr" = "1.1.0",
    "purrr" = "0.3.0",
    "stringr" = "1.4.0"
)

# loop over required packages
for (i in sqe_along(required_packages)) {
  # get current package
  pkg <- required_packages[i]

  # get installed packages' versions
  pkgs <- installed.packages()[, "Version"]

  # check if required package already installed
  if (any(idx <- names(pkg) == names(pkgs))) {
    # if so, upgrade if necessary
    if (pkg > pkgs[idx])
      install.packages(names(pkg), quiet = TRUE)
  } else {
    # otherwise, install
    install.packages(names(pkg), quiet = TRUE)
  }
}
```

Listing A.2: R code to install and use the `learn2scrape` package.

```r
# install
remotes::install_github("haukelicht/learn2scrape", ref = "methodsnet2022")

# load the package
library(learn2scrape)
```

```r
# list available tutorials
available_tutorials("learn2scrape")

# run a tutorial (will open in your Browser/Viewer)
run_tutorial("001-tutorial-how-to", package = "learn2scrape")
```