Advanced Web Data Collection Workshop

Cornelius Erfort

May 16-17, 2025

Course Overview

The internet is an essential source of data for social science research, providing access to vast amounts of text and structured information. This course introduces students to both basic and advanced methods for automated web data collection, focusing on practical applications in political science and other social sciences. Students will learn web scraping techniques for static and dynamic content, work with APIs, process various data formats, and automate data collection workflows. The course also covers browser automation, error handling, scheduling scraping jobs, and ethical and legal considerations.

Learning Objectives

By the end of this workshop, participants will be able to:

- Understand web scraping fundamentals and best practices
- Use R and Python libraries for web scraping (rvest, RSelenium, requests, BeautifulSoup, httr)
- Handle different types of web content (static, dynamic, JavaScript-rendered)
- Work with APIs and direct HTTP requests
- Store and process scraped data in various formats (CSV, JSON, XML, etc.)
- Implement rate limiting, respect robots.txt, and manage sessions/cookies
- Automate scraping tasks and schedule jobs (cron, Docker)
- Handle common scraping challenges, errors, and logging
- Understand and apply ethical and legal considerations in web data collection

Course Schedule

Day 1: Fundamentals and Basic Scraping

Time	Topic/Activities
9:00	Start, Introduction
	Course overview, setup
	HTML/CSS, Web Structure
12:30–13:30	Lunch Break
	Static Web Scraping, APIs and Data Formats
	Continue with exercises and project work
	Hands-on Project, Practical scraping (static sites, APIs)
	Wrap-up
17:00	Finish

Day 2: Advanced Techniques and Best Practices

Time	Topic/Activities
9:00	Start
	Day 1 Review, Recap and questions
	Dynamic Content, Browser automation (RSelenium)
12:30–13:30	Lunch Break
	HTTP Requests, Automation, Scheduling
	Ethics, Legal Aspects
	Continue with exercises and project work
	Hands-on Project, Project presentations, Q&A
	Wrap-up
17:00	Finish

Course Materials

- Web browser with developer tools (Firefox or Chrome recommended)
- Code editor (RStudio, Cursor, or similar)
- Required R packages:
  - rvest
  - xml2
  - httr
  - RSelenium
  - jsonlite
  - tidyverse

Resources

- rvest Documentation
- RSelenium Documentation
- httr Documentation
- SelectorGadget
- CSS Diner
- List of Free Public APIs
- List of R Packages for APIs

Example Web Scraping Projects

Easy

- Wikipedia: List of political scientists (`https://en.wikipedia.org/wiki/List_of_political_scientists`)
- Parties' press releases (`https://www.spd.de/presse/pressemitteilungen/`)
- Polls (wahlrecht.de) (`https://www.wahlrecht.de/umfragen/`)
- Conference programs (EPSA, DVPW) (`https://epsaweb.org/annual-conference/`)
- Abgeordnetenwatch.de (questions and answers from candidates) (`https://www.abgeordnetenwatch.de/`)

- German Lobby Register (https://www.lobbyregister.bundestag.de/)
- Web Search Results (DuckDuckGo) (https://duckduckgo.com/)
- News articles (https://www.nytimes.com/section/world)

Medium

- Korean election results (backend, JSON) (https://info.nec.go.kr/)
- Parliamentary protocols (PDFs) (https://www.bundestag.de/protokolle)
- US live election data from the New York Times (JSON backend) (https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html)
- Polls (Politico JSON) (https://www.politico.com/interactives/2024/president-election-po
- Doctolib appointment availability (JSON) (https://www.doctolib.de/)
- List of far-right demonstrations from parliamentary query (parsing PDFs and geocoding) (https://dip.bundestag.de/)
- Privatization of state owned companies (Treuhand) map (https://treuhandanstalt.online/karte/)

Difficult

- German Members of Parliament (MPs) websites (parallel scraping/crawling)
- LinkedIn profiles (Python library) (https://pypi.org/project/linkedin-scraper/)
- Air quality sensor data worldwide (encrypted via JavaScript) (https://waqi.info/#/c/3.563/8.145/2.2z)
- Historic shapefiles for Danish parishes (https://dataforsyningen.dk/data/4840)

Research Applications

- Siegel, Alexandra A., and Vivienne Badaan. "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online." *American Political Science Review* 114, no. 3 (2020): 837–55. https://doi.org/10.1017/S0003055420000283
- Mitts, Tamar. "Banned: How Deplatforming Extremists Mobilizes Hate in the Dark Corners of the Internet." *Working Paper* (2021). https://www.dropbox.com/s/iatnxn5gtq48fxu/Mitts_banned.pdf?dl=0
- Boas, Taylor C., and F. Daniel Hidalgo. "Controlling the Airwaves: Incumbency Advantage and Community Radio in Brazil." *American Journal of Political Science* 55, no. 4 (2011): 869–85. https://doi.org/10.1111/j.1540-5907.2011.00532.x

- Bischof, Daniel, and Thomas Kurer. "Place-Based Campaigning: The Political Impact of Real Grassroots Mobilization." *The Journal of Politics* (2023). https://doi.org/10.1086/723985
- Box-Steffensmeier, Janet M., et al. "I Get By with a Little Help from My Friends: Leveraging Campaign Resources to Maximize Congressional Power." *American Journal of Political Science* 64, no. 4 (2020): 1017–33. https://doi.org/10.1111/ajps.12528
- Motolina, Lucia. "Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico." *American Political Science Review* 115, no. 1 (2021): 97–113. https://doi.org/10.1017/S0003055420000672
- Sances, Michael W. "Defund My Police? The Effect of George Floyd's Murder on Support for Local Police Budgets." *The Journal of Politics* (2023). https://doi.org/10.1086/723979
- Lutscher, Philipp M. "Hot Topics: Denial-of-Service Attacks on News Websites in Autocracies." *Political Science Research and Methods* (2021): 1–16. https://doi.org/10.1017/psrm.2021.68
- Erfort, Cornelius, Klüver, Heike, and Stötzer, Lukas F. "The PARTYPRESS Database: A new Comparative Database of Parties' Press Releases." *Research and Politics* (2023).
- Morris, Kevin. "Turnout and Amendment Four: Mobilizing Eligible Voters Close to Formerly Incarcerated Floridians." *American Political Science Review* 115, no. 3 (2021): 805–20. https://doi.org/10.1017/S0003055421000253
- Gessler, Theresa, & Hunger, Sophia. "How the refugee crisis and radical right parties shape party competition on immigration." *Political Science Research and Methods*, 10(3), 524-544 (2022). https://doi.org/10.1017/psrm.2021.64
- Stukal, Denis, et al. "Why Botter: How Pro-Government Bots Fight Opposition in Russia." *American Political Science Review* 116, no. 3 (2022): 843–57. https://doi.org/10.1017/S0003055421001507

Contact Information

cornelius.erfort@uni-wh.de