GESIS Leibniz Institute for the Social Sciences

gesis training

GESIS Fall Seminar in Computational Social Science 2025
"Web Data Collection with R"

Lecturers:	Iulia Cioroianu
Affiliation:	University of Bath, UK
Email:	iulia.cioroianu@gmail.com

Date: September 8-12, 2025
Time: 09:30-13:00 (CEST / UTC+2)

About the Lecturer

Iulia Cioroianu is a Senior Lecturer (Associate Professor) in the Department of Politics, Languages and International Studies at the University of Bath. She holds a Ph.D. in Political Science from New York University and an M.A. from Central European University. Before joining the University of Bath, she was a research fellow in the Q-Step Centre for Quantitative Social Sciences at the University of Exeter, and a pre-doctoral fellow in the LSE Department of Methodology. Iulia is a data scientist who studies online communication and information exposure using a range of computational methods such as natural language processing and quantitative text analysis, agent-based modelling, machine learning and experiments.

Course Description

The exponential increase in online and social media data offers unprecedented opportunities for advancing research across a variety of fields, both within academia and outside of it. For instance, diverse data such as election results, press releases, or social media posts can inform research questions in the social sciences. Although the availability of data online is steadily increasing, extracting these data is not always straightforward, especially since many popular social media sites have shut down or restricted access to their Application Programming Interfaces (APIs). Furthermore, the heterogeneity of data almost always requires reshaping these data before they can be used effectively for analysis, which can also be challenging. This course provides researchers with the tools needed to collect and pre-process large-scale data from a range of online sources.

Through a combination of lectures, hands-on tutorials and individual/group exercises, participants will develop a theoretical understanding of the challenges associated with online data collection and the best methods and tools for addressing them in R, as well as the practical skills needed to scrape data from both static and dynamic websites and collect data through APIs. The sources used in the examples provided include social media websites, online media outlets and news aggregators, government data portals, and other large-scale online data repositories.

Acknowledging that the most difficult part of a computational project involving the collection of complex and heterogenous data is often the pre-processing needed to prepare the data for subsequent analysis and link it across a variety of sources, the course also covers text-based methods for data cleaning and pre-processing. By the end of the week, participants should be able to apply the methods studied to extract and process data for their own research projects.

Organizational Structure of the Course

The course will be offered **online**, and will be taught in R in the morning. A parallel course taught in Python takes place in the afternoon (14:00-17:30), and **participants interested in taking part in both courses should contact the Fall Seminar team at fallseminar@gesis.org for a discounted rate**. The content and examples used in the lecturer-led tutorials are similar across programming languages, making it easier for those interested in developing new skills in a secondary language that they may not be proficient in to do so by drawing parallels across the two courses.

We will start the daily sessions with a lecture laying out the main notions and providing an overview of the language-specific tools used (approximately 45 minutes), followed by a hands-on lecturer-led tutorial (45 minutes). The second part of the session will consist of several exercises that students are encouraged to solve in small groups or individually (90 minutes). Each exercise will be followed by an instructor-led discussion of the solutions. In the final part of the session, students will complete a short exercise as an individual assignment (30 minutes). The daily schedule is presented in the table below.

Theoretical overview and lecturer-led tutorial	9:30-11:00
Individual or small-group exercises and solutions	11:15-12:00, 12:10-12:30
Individual assignment	12:30-13:00

The lecturer will provide continuous support during the exercise sessions, and will be available for individual consultations on participants' projects during the last day.

Keywords

web scraping, automated data collection, APIs, R

Target Group

You will find the course useful if:

- You want to learn how to collect and process large amounts of data from online sources fast.
- You aim to improve your existing web scraping skills or have so far encountered difficulties trying to scrape data from online sources.
- You have a research idea for which online data might be suitable, but you are not sure of the practical implications.

Course and Learning Objectives

By the end of the course, you will:

- Understand the structure and basic features of different forms of online data.
- Be able to collect data from static and dynamic websites.
- Be able to interact with APIs to access and collect data.
- Be able to parse, clean and process the data collected.
- Be able to apply the methods studied to your own research projects.

Course Prerequisites

- Working knowledge of R, including data structures and control structures.
- Participants attending both the R and the Python course should have working knowledge of each of the two programming languages.
- If you lack basic knowledge of these programming languages, you are encouraged to take the Introduction to Computational Social Science with R or Introduction to Computational Social Science with Python course in week 1 and/or the introductory online workshops (Intro to R, Intro to Python).

Software Requirements

Participants should pre-install the following software and packages:

- RStudio
- required packages (final list of packages to be provided before the course): httr, rvest, RSelenium, dplyr, tidyr, stringr, quanteda

Course Contents

- Forms of online data and data structures
- Scraping static websites
- Scraping dynamic websites, including browser interaction and automation
- Working with APIs
- Storing and pre-processing online data
- Good practices surrounding online data collection

Recommended Literature to Look at in Advance

R refresher:
Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science*. O'Reilly Media, Inc. https://r4ds.hadley.nz/

Or a refresher on R and Python side-by-side:
Scavetta, R. J., & Angelov, B. (2021). *Python and R for the Modern Data Scientist*. O'Reilly Media, Inc. https://www.oreilly.com/library/view/python-and-r/9781492093398/

Other recommended readings:
Luscombe, A., Dick, K., & Walby, K. (2022). Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity, 56*(3), 1023-1044. https://doi.org/10.1007/s11135-021-01164-0
Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060-1062. https://doi.org/10.1126/science.aaz8170
Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*, 61-81. https://doi.org/10.1146/annurev-soc-121919-054621
Theocharis, Y. & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication, *Political Communication*, *38*(1-2), 1-22. https://doi.org/10.1080/10584609.2020.1833121

Day-to-day Schedule and Literature

Day 1: Introduction to Web Data Collection
We will begin with an overview of web data collection, its applications in social science research, and the ethical and legal issues associated with its use. The theoretical overview session will provide a general idea of the structure and logic of websites and APIs and introduce the core concepts that we will be working with, as well as the R packages that we will be using. We will also discuss the structure and features of different forms of online data. Participants will share their expectations and discuss their intended use of web data collection for their own research.

After ensuring that participants' setups are functional, we will go over the required credentials for working with different APIs that we will need later in the course. A series of exercises will reinforce basic programming and data manipulation pre-requisite skills. Participants will then make their first HTTP request, and we will explore the output together in preparation for next day's session.

Literature (required)
Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations (arXiv:2410.23432). arXiv. https://doi.org/10.48550/arXiv.2410.23432

Literature (suggested)
Li, F., Zhou, Y., & Cai, T. (2021). Trails of data: Three cases for collecting web information for social science research. *Social Science Computer Review*, *39*(5), 922-942. https://doi.org/10.1177/0894439319886019

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication, 35*(4), 665-668. https://doi.org/10.1080/10584609.2018.1477506

Luscombe, A., Dick, K. & Walby, K. (2022). Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Qual Quant, 56,* 1023-1044. https://doi.org/10.1007/s11135-021-01164-0

Day 2: Scraping Static Websites

During the second day we will cover methods of data extraction from websites for which the content is fixed and does not change based on user interaction. Building upon the final lab exercise from the previous day, we will discuss the basics of HTTP requests and how web pages are structured using HTML and CSS. We will then cover different methods for locating and extracting various data types from web pages, including text, hyperlinks, tables, images and other media, as well as metadata, and go over some practical examples using *rvest*. An introduction to regular expressions will also be provided.

In the applied session participants will work through a series of exercises designed to test and reinforce their skills, including capturing article titles, authors, and publication dates from a news archive, and collecting indicators from the websites of governments and international organizations. Each exercise will be structured as a mini research project that participants will have to prepare and implement either in a group or individually.

Literature (suggested)

Rvest documentation: https://cran.r-project.org/web/packages/rvest/rvest.pdf

Day 3: Scraping Dynamic Websites

We will begin the third day with an overview of the unique challenges of dynamic websites that make traditional scraping techniques inadequate, and discuss the main ways in which we can overcome them programmatically. We will then demonstrate the use of *RSelenium* to automate and simulate user interactions with web pages through several examples. We will learn to handle dynamic pagination, JavaScript code and infinite scrolling, and to manage browsing sessions, requests to authenticate and rate limits.

A series of exercises will provide the opportunity to apply these skills in different scenarios, including collecting online financial data, automating online database queries, and collecting data from search engines and social media websites that would be otherwise difficult to access.

Literature (suggested)

Selenium documentation: https://www.selenium.dev/documentation/

R: https://cran.r-project.org/web/packages/RSelenium/RSelenium.pdf

Day 4: Working with APIs

Application Programming Interfaces (APIs) can be used to access and collect a range of data that is relevant for social science research. We will start with an introduction to APIs—what they are, how they function, and why they are important tools for researchers looking to access structured data directly from online platforms. We will learn about endpoints, requests, responses, and authentication methods, and go over several examples based on a simple political information API (Vote Smart) as well as the Reddit API.

In the practical session, participants will apply the new skills to extract data from other APIs, such as the Manifesto Project API, the YouTube Data API, the Bluesky API, and the OpenAI API.

Literature (suggested)

Nyhuis, D. (2021). Application programming interfaces and web data for social research. In *Handbook of Computational Social Science, Volume 2*. Routledge. https://doi.org/10.4324/9781003025245-4

Breuer, J., Kmetty, Z., Haim, M. & Stier, S. (2022) User-centric approaches for collecting Facebook data in the 'post-API age': experiences from two studies and recommendations for future research. *Information, Communication & Society*, *26*(14), 2649-2668. https://doi.org/10.1080/1369118X.2022.2097015

Day 5: Processing, Storing, and Starting to Analyze the Collected Data

The final day will provide essential skills for handling the collected web data from initial collection and storage to final analysis. The first part of the session focuses on pre-processing and cleaning. Participants will learn the essential techniques for transforming the raw collected data into formats that are suitable for storage, sharing, and analysis. We will explore different storage options and the types of databases suitable for large-scale web data. We will discuss relational databases such as MySQL, NoSQL databases like MongoDB, and cloud storage solutions, and evaluate the choice of database based on the project's scale, data type, and accessibility requirements. We will briefly consider the process of writing the documentation for the resulting dataset to facilitate research sharing and replicability. Finally, we will focus on text cleaning, processing, and analysis tasks that are commonly used for web data, and provide an overview of text analysis techniques such as sentiment analysis and topic modeling.

In the afternoon sessions participants will have the opportunity to apply these skills to some of the data collected in previous days. By the end of the course participants will have a portfolio of thoroughly documented web scraping examples (from planning stages all the way to basic analysis) that can be replicated or adapted for other use cases.

Literature (suggested)

R Quanteda documentation: http://quanteda.io/
Benoit, K. (2020). Text as Data: An Overview. In *The SAGE Handbook of Research Methods in Political Science and International Relations* (pp. 461-497). SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

Additional Recommended Literature

The following books and articles are useful as reference material and for other/more specialized tasks:

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons. https://doi.org/10.1002/9781118834732
Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691207551/text-as-data
Kopecký, J., Fremantle, P., & Boakes, R. (2014). A history and future of Web APIs. *IT - Information Technology*, *56*(3), 90-97. https://doi.org/10.1515/itit-2013-1035