Course title
WA108 Automated web data collection with R

Instructor details
First name, last name: Peter Meißner
Department/Unit: Department of Politics and Public Administration
Institution: University of Konstanz
E-mail: peter.meissner@uni-konstanz.de

Short Bio
I am a researcher in the Institutional Design in European Democracies (IDEP) working group and doctoral student at the University of Konstanz. Part of my work for IDEP is the provision of semi-automated data gathering infrastructure for which I extensively use R and database systems. My substantive research interests include political institutions in general and parliaments, the legislative process as well as political careers in particular. As a co-author of the book ‘Automated Data Collection With R’, forthcoming at Wiley, I have been working on web scraping techniques with R over the past years.

Prerequisite knowledge
Note from the Academic Convenors to prospective participants: by registering to this course, you certify that you possess the prerequisite knowledge that is requested to be able to follow this course. The instructor will not teach again these prerequisite items. If you doubt whether you possess that knowledge to a sufficient extent, we suggest you contact the instructor before you proceed to your registration.

For the course a no less than intermediate practical knowledge of R is an indispensable prerequisite, as there is no time for an R refresher. The following questions help you to find out if your level of R knowledge suffices to follow the course.

- Do you know how to subsets vectors, data frames and lists?
- Can you transform lists into data frames, matrices into vectors and factors into character?
- Are you familiar with the apply functionality (apply, lapply, sapply, mapply) or, alternatively, the plyr package?
- Can you write a loop?
- Can you write a function?

If the answer to all of these questions is yes, you are well prepared to get the best out of this course.

Disclaimer: the information contained in this course description form may be subject to subsequent adaptations (e.g. taking into account new developments in the field, specific participant demands, group size etc.). Registered participants will be informed in due time in case of adaptations.
If you are not too familiar with R yet but still want to join the course, there are many well-written books on the market that provide great introductions to R:


Besides these commercial sources, there is also a lot of free information on the Web. Visit http://cran.r-project.org/manuals.html and http://cran.r-project.org/other-docs.html to get R introductions and manuals. A truly amazing online tutorial for absolute beginners by the Code School is made available at http://tryr.codeschool.com/. Additionally, Quick-R (http://www.statmethods.net/) is a good reference site for many basic commands. You can also find a lot of free resources and examples at http://www.ats.ucla.edu/stat/r/ and if you want to dive deeper into R have a look at Hadley Wickham’s page at http://adv-r.had.co.nz/.

**Short course outline**

Are you interested in the analysis of social media data? Do you want to extract information from websites to build your own data set of, e.g., press releases, online newspaper headlines, parliamentary speeches, or politicians’ biographies? The rapid growth of the World Wide Web over the past two decades made firms, public institutions and private users provide every imaginable type of information and new channels of communication generate vast amounts of data on human behaviour.

Along with the triumphant entry of the World Wide Web, we have witnessed a second trend, the increasing popularity and power of open source software like R. For quantitative social scientists, R is among the most important statistical software. It is growing rapidly due to an active community that constantly publishes new packages. An extraordinarily useful feature of R is that we can use it as a powerful tool for collecting, extracting, cleaning and storing data from web sources. In combination with R’s well known strengths in data analyses, data manipulation and data visualization, R facilitates staying in a familiar programming environment throughout the course of the research and helps researchers to focus on substantive problems instead of investing too much time in learning other software like PHP, Python or Pearl.

**Long course outline**

*Automated web data collection with R* is a course designed to give you a primer on web scraping with R. Web scraping is a general term for all kinds of activities that involve (automated) gathering of data and texts from the web: starting with tiny bits of information – like the current time or geographical location of say Ulan Bator; or maybe the current headline of the *The Sun* – up to retrieving hundreds of speeches, texts and other documents from dozens of different web pages. Around the world. Each day. Every 5 minutes. For the next ten months.

Web scraping can give new insights into your social science research project or indeed help you develop a complete new subject by making available data that just recently started to exist – e.g. Twitter and blog posts or webpage linkage networks – or it simply might push the scales of the amount of data that can be used from traditional sources to new limits – e.g. working with thousands of speeches, news articles, or law proposals instead of just a handful.

Although, web scraping is not a new technique, it just starts to get recognized by a broad social science audience. So far, statistical software used by social scientists like SPSS, Stata or R did not have web scraping capabilities while software that had the means for web scraping like PHP, Pearl or Python was alien to social scientists. Furthermore, web scraping tends to get presented either purely technical in form of ‘from programmers for programmers’ manuals or as specialized single-issue case studies that make it hard to get the general picture. Being well aware that social scientists usually want to do substantive research instead of learning yet another software, the course builds on a
thorough but gentle, hands on web scraping textbook written by social scientists – my colleagues and me – for social scientists: *Automated Data Collection with R - A Practical Guide to Web Scraping and Text Mining*. The book rests on three essential pillars: (1) providing introductions to web technologies and associated tools, (2) presenting handy R packages and howtos for web scraping and (3) illustrating real life applications with case studies. In addition each chapter (except case studies) comes with a set of exercises and solutions making it ‘the book we would have wanted to have when starting to scrape the web ourselves’.

As preparatory course, the main goal of the course is to give you a solid overview on the most important web technologies, how they are connected and which tools are available in R to handle them. Having completed the course you should be able to cope with simple scraping scenarios: downloading various file formats, extracting lists and tables from HTML documents or retrieve information that is easily accessible. Although, you will not become a scraping master after only two days, more complicated scraping scenarios – e.g. accessing Twitter data, handling APIs, handling dynamic web pages, coping with large amounts of data and so on – should come in reach with continuous exercise and further study of the book.

**Day-to-day schedule (Friday 13 February to Saturday 14 February)**

<table>
<thead>
<tr>
<th></th>
<th>Topic(s)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday afternoon</td>
<td>Introduction.</td>
<td>In the first part of the session we will work on getting an overview of the universe of web technologies and how they are interrelated. The second part of the session is reserved for exploring R’s base capabilities to gather data from the web.</td>
</tr>
<tr>
<td>Saturday morning</td>
<td>Basic web scraping.</td>
<td>In the first part of the session we make our selves familiar with THE most common forms of getting data presented in the web – HTML and XML. The second part of the session is reserved for extracting information from those documents.</td>
</tr>
<tr>
<td>Saturday afternoon</td>
<td>Advanced web scraping</td>
<td>In the last session we refine our scraping skills, tap into more refined and powerful tools for data extraction and will look into further problems and where we can find R specific solutions.</td>
</tr>
</tbody>
</table>

**Day-to-day reading list**

All readings are based on: Munzert, Rubba, Meißner, Nyhuis (2014): *Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining*. Wiley, numbers in parentheses indicate chapter numbers

<table>
<thead>
<tr>
<th></th>
<th>Friday afternoon</th>
<th>Saturday morning</th>
<th>Saturday afternoon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Preface, Introduction (1), Regular Expressions and String Functions (8), all Introductions to chapter 2-8 (everything until the first headline)</td>
<td>HTML (2), XML (3)</td>
<td>XPATH (4), AJAX (6), Scraping the Web (9), Managing Projects (11)</td>
</tr>
</tbody>
</table>

**Software and hardware requirements**

**Software programme**

R. and your editor of choice (e.g., RStudio)

**Hardware requirements**

Any modern computer should do – please bring your own Laptop.
Literature

Lab requirement
Reliable WirelessLAN access allowing the participants to use their own Laptops would be great. (If not possible, please contact me! Peter.Meissner@uni-konstanz.de, 0049 7531885665,
Each computer should have Internet connection, R, Rstudio, all rights to install additional R packages, some 500 Mb of personal disk space, and Google’s Chrome as browser)

Other recommended courses (before or after this course)
The following other ECPR Methods School courses could be useful in combination with this one in a ‘training track’. NB this is an indicative list.

Before this course:
Introductory or advanced courses on R or any course providing guidelines on data handling. Although, no statistical pre-knowledge is required, as this is a course on data collection, courses on data analyses can be complementary.

After this course:
As this is a course on data collection, courses on data analyses can be complementary.