# Automated Data Collection with R

## A Practical Guide to Web Scraping and Text Mining


Simon Munzert | Christian Rubba | Peter Meißner | Dominic Nyhuis

WILEY

# Automated Data Collection with R

## A Practical Guide to Web Scraping and Text Mining

**Simon Munzert**

*Department of Politics and Public Administration, University of Konstanz, Germany*

**Christian Rubba**

*Department of Political Science, University of Zurich and National Center of Competence in Research, Switzerland*

**Peter Meißner**

*Department of Politics and Public Administration, University of Konstanz, Germany*

**Dominic Nyhuis**

*Department of Political Science, University of Mannheim, Germany*



WILEY

This edition first published 2015


*Registered office*
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom



*Library of Congress Cataloging-in-Publication Data*

Munzert, Simon.
Automated data collection with R : a practical guide to web scraping and text mining / Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis.
pages cm
Summary: "This book provides a unified framework of web scraping and information extraction from text data with R for the social sciences"– Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-118-83481-7 (hardback)
1. Data mining. 2. Automatic data collection systems. 3. Social sciences–Research–Data processing. 4. R (Computer program language) I. Title.
QA76.9.D343M865 2014
006.3′12–dc23

2014032266

A catalogue record for this book is available from the British Library.

ISBN: 9781118834817

Set in 10/12pt Times by Aptara Inc., New Delhi, India.

Printed and bound in Malaysia by Vivar Printing Sdn Bhd

1 2015

# Automated Data Collection with R

To my parents, for their unending support. Also, to Stefanie.
—Simon

To my parents, for their love and encouragement.
—Christian

To Kristin, Buddy, and Paul for love, regular walks, and a final deadline.
—Peter

Meiner Familie.
—Dominic

# Preface

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 researchers and businesses alike.

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. Yet, R is more than a free statistics suite. It also incorporates interfaces to many other programming languages and software solutions, thus greatly simplifying work with data from various sources.

On a personal note, we can say the following about our work with social scientific data:

- our financial resources are sparse;
- we have little time or desire to collect data by hand;
- we are interested in working with up-to-date, high quality, and data-rich sources; and
- we want to document our research from the beginning (data collection) to the end (publication), so that it can be reproduced.

In the past, we frequently found ourselves being inconvenienced by the need to manually assemble data from various sources, thereby hoping that the inevitable coding and copy-and-paste errors are unsystematic. Eventually we grew weary of collecting research data in a non-reproducible manner that is prone to errors, cumbersome, and subject to heightened risks of death by boredom. Consequently, we have increasingly incorporated the data collection and publication processes into our familiar software environment that already helps with statistical analyses—R. The program offers a great infrastructure to expand the daily workflow to steps before and after the actual data analysis.

Although R is not about to collect survey data on its own or conduct experiments any time soon, we do consider the techniques presented in this book as more than the "the poor man's substitute" for costly surveys, experiments, and student-assistant coders. We believe that they are a powerful supplement to the portfolio of modern data analysts. We value the collection of data from online resources not only as a more cost-sensitive solution compared to traditional data acquisition methods, but increasingly think of it as the exclusive approach to assemble datasets from new and developing sources. Moreover, we cherish program-based solutions because they guarantee reliability, reproducibility, time-efficiency, and assembly of higher quality datasets. Beyond productivity, you might find that you enjoy writing code and drafting algorithmic solutions to otherwise tedious manual labor. In short, we are convinced that if you are willing to make the investment and adopt the techniques proposed in this book, you will benefit from a lasting improvement in the ease and quality with which you conduct your data analyses.

If you have identified online data as an appropriate resource for your project, is web scraping or statistical text processing and therefore an automated or semi-automated data collection procedure really necessary? While we cannot hope to offer any definitive guidelines, here are some useful criteria. If you find yourself answering several of these affirmatively, an automated approach might be the right choice:

- Do you plan to repeat the task from time to time, for example, in order to update your database?
- Do you want others to be able to replicate your data collection process?
- Do you deal with online sources of data frequently?
- Is the task non-trivial in terms of scope and complexity?
- If the task can also be accomplished manually—do you lack the resources to let others do the work?
- Are you willing to automate processes by means of programming?

Ideally, the techniques presented in this book enable you to create powerful collections of existing, but unstructured or unsorted data no one has analyzed before at very reasonable cost. In many cases, you will not get far without rethinking, refining, and combining the proposed techniques due to your subjects' specifics. In any case, we hope you find the topics of this book inspiring and perhaps even eye opening: The streets of the Web are paved with data that cannot wait to be collected.

## What you won't learn from reading this book

When you browse the table of contents, you get a first impression of what you can expect to learn from reading this book. As it is hard to identify parts that you might have hoped for but that are in fact not covered in this book, we will name some aspects that you will not find in this volume.

What you will not get in this book is an introduction to the R environment. There are plenty of excellent introductions—both printed and online—and this book won't be just another addition to the pile. In case you have not previously worked with R, there is no reason

to set this book aside in disappointment. In the next section we'll suggest some well-written R introductions.

You should also not expect the definitive guide to web scraping or text mining. First, we focus on a software environment that was not specifically tailored to these purposes. There might be applications where R is not the ideal solution for your task and other software solutions might be more suited. We will not bother you with alternative environments such as PHP, Python, Ruby, or Perl. To find out if this book is helpful for you, you should ask yourself whether you are already using or planning to use R for your daily work. If the answer to both questions is no, you should probably consider your alternatives. But if you already use R or intend to use it, you can spare yourself the effort to learn yet another language and stay within a familiar environment.

This book is not strictly speaking about data science either. There are excellent introductions to the topic like the recently published books by O'Neil and Schutt (2013), Torgo (2010), Zhao (2012), and Zumel and Mount (2014). What is occasionally missing in these introductions is how data for data science applications are actually acquired. In this sense, our book serves as a preparatory step for data analyses but also provides guidance on how to manage available information and keep it up to date.

Finally, what you most certainly will not get is the perfect solution to your specific problem. It is almost inherent in the data collection process that the fields where the data are harvested are never exactly alike, and sometimes rapidly change shape. Our goal is to enable you to adapt the pieces of code provided in the examples and case studies to create new pieces of code to help you succeed in collecting the data you need.

## Why R?

There are many reasons why we think that R is a good solution for the problems that are covered in this book. To us, the most important points are:

1. R is freely and easily accessible. You can download, install, and use it wherever and whenever you want. There are huge benefits to not being a specialist in expensive proprietary programs, as you do not depend on the willingness of employers to pay licensing fees.

2. For a software environment with a primarily statistical focus, R has a large community that continues to flourish. R is used by various disciplines, such as social scientists, medical scientists, psychologists, biologists, geographers, linguists, and also in business. This range allows you to share code with many developers and profit from well-documented applications in diverse settings.

3. R is open source. This means that you can easily retrace how functions work and modify them with little effort. It also means that program modifications are not controlled by an exclusive team of programmers that takes care of the product. Even if you are not interested in contributing to the development of R, you will still reap the benefits from having access to a wide variety of optional extensions—packages. The number of packages is continuously growing and many existing packages are frequently updated. You can find nice overviews of popular themes in R usage on http://cran.r-project.org/web/views/.

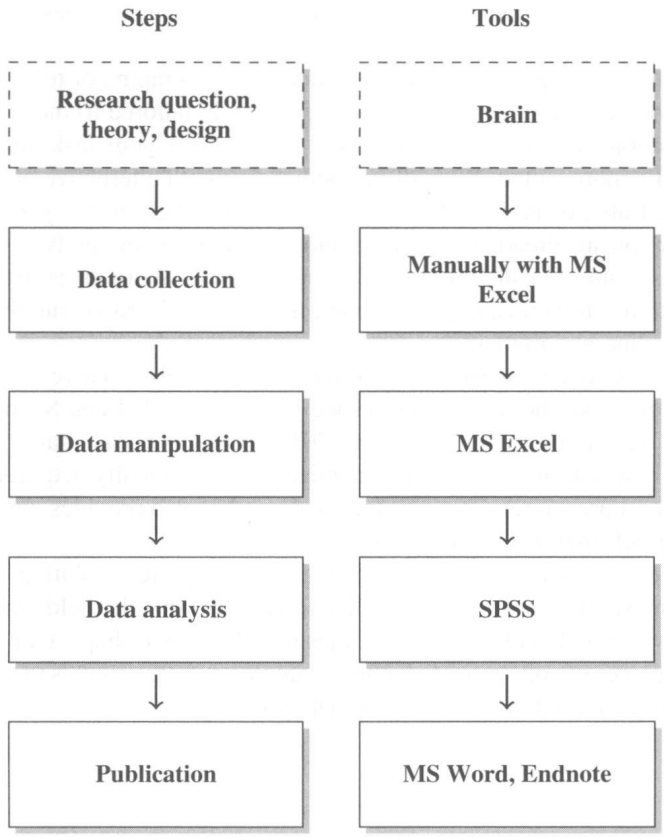


**Figure 1** The research process **not** using R—stylized example

4. R is reasonably fast in ordinary tasks. You will likely agree with this impression if you have used other statistical software like SPSS or Stata and have gotten into the habit of going on holiday when running more complex models—not to mention the pain that is caused by the "one session, one data frame" logic. There are even extensions to speed up R, for example, by making C code available from within R, like the Rcpp package.
5. R is powerful in creating data visualizations. Although this not an obvious plus for data collection, you would not want to miss R's graphics facilities in your daily workflow. We will demonstrate how a visual inspection of collected data can and should be a first step in data validation, and how graphics provide an intuitive way of summarizing large amounts of data.
6. Work in R is mainly command line based. This might sound like a disadvantage to R rookies, but it is the only way to allow for the production of reproducible results compared to point-and-click programs.
7. R is not picky about operating systems. It can generally be run under Windows, Mac OS, and Linux.
8. Finally, R is the entire package from start to finish. If you read this book, you are likely not a dedicated programmer, but hold a substantive interest in a topic or specific

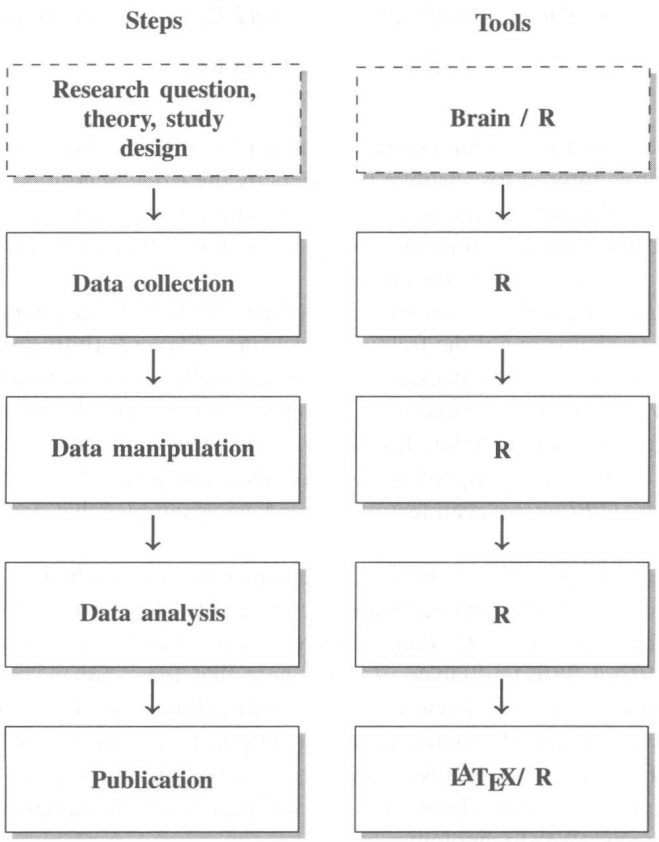


**Figure 2** The research process using R—stylized example

data source that you want to work with. In that case, learning another language will not pay off, but rather prevent you from working on your research. An example of a common research process is displayed in Figure 1. It is characterized by a permanent switching between programs. If you need to make corrections to the data collection process, you have to climb back down the entire ladder. The research process using R, as it is presented in this book, takes place within a single software environment (Figure 2). In the context of web scraping and text processing, this means that you do not have to learn another programming language for the task. What you will need to learn are some basics in the markup languages HTML and XML and the logic of regular expressions and XPath, but the operations are executed from within R.

## Recommended reading to get started with R

There are many well-written books on the market that provide great introductions to R. Among these, we find the following especially helpful:

Crawley, Michael J. 2012. *The R Book*, 2nd edition. Hoboken, NJ: John Wiley & Sons.

Adler, Joseph. 2009. *R in a Nutshell. A Desktop Quick Reference.* Sebastopol, CA: O'Reilly.

Teetor, Paul. 2011. *R Cookbook.* Sebastopol, CA: O'Reilly.

Besides these commercial sources, there is also a lot of free information on the Web. 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. Lastly, you can also find a lot of free resources and examples at http://www.ats.ucla.edu/stat/r/.

R is an ever-growing software, and in order to keep track of the developments you might periodically like to visit some of the following websites: *Planet R* (http://planetr.stderr.org/) provides the history of existing packages and occasionally some interesting applications. *R-Bloggers* (http://www.r-bloggers.com/) is a blog aggregator that collects entries from many R-related blog sites in various fields. It offers a broad view on hundreds of R applications from economics to biology to geography that is mostly accompanied by the necessary code to replicate the posts. *R-Bloggers* even features some basic examples that deal with automated data collection.

When running into problems, R help files are sometimes not too helpful. It is often more enlightening to look for help in online forums like *Stack Overflow* (http://stackoverflow.com) or other sites from the *Stack Exchange* network. For complex problems, consider the R experts on *GitHub* (http://github.com). Also note that there are many Special Interest Group (SIG) mailing lists (http://www.r-project.org/mail.html) on a variety of topics and even local R User Groups all around the world (http://blog.revolutionanalytics.com/local-r-groups.html). Finally, a CRAN Task View has been set up, which gives a nice overview over recent advances in web technologies and services in the R framework: http://cran.r-project.org/web/views/WebTechnologies.html

## Typographic conventions

This is a practical book about coding, and we expect you to often have it sitting somewhere next to the keyboard. We want to facilitate the orientation throughout the book with the following conventions: There are three indices—one for general topics, one for R packages, and one for R functions. Within the text, variables and R (and other) code and functions are set in typewriter typeface, as in `summary()`. Actual R code is also typewriter style and indented. Note that code input is indicated with "R" and a prompt symbol ("`R>`"); R output is printed without the prompt sign, as in

```
R> hello <- "hello, world"
R> hello
[1] "hello, world"
```

## The book's website

The website that accompanies this book can be found at http://www.r-datacollection.com

Among other things, the site provides code from examples and case studies. This means that you do not have to manually copy the code from the book, but can directly access and modify the corresponding R files. You will also find solutions to some of the exercises, as well as a list of errata. If you find any errors, please do not hesitate to let us know.

## Disclaimer

This is not a book about spidering the Web. Spiders are programs that graze the Web for information, rapidly jumping from one page to another, often grabbing the entire page content. If you want to follow in Google's Googlebot's footsteps, you probably hold the wrong book in your hand. The techniques we introduce in this book are meant to serve more specific and more gentle purposes, that is, scraping specific information from specific websites. In the end, you are responsible for what you do with what you learn. It is frequently not a big leap from the code that is presented in this book to programs that might quickly annoy website administrators. So here is some fundamental advice on how to behave as a practitioner of web data collection:

1. Always keep in mind where your data comes from and, whenever possible, give credit to those who originally collected and published it.[1]
2. Do not violate copyrights if you plan to republish data you found on the Web. If the information was not collected by yourself, chances are that you need permission from the owners to reproduce them.
3. Do not do anything illegal! To get an idea of what you can and cannot do in your data collection, check out the Justia BlawgSearch (http://blawgsearch.justia.com/), which is a search site for legal blogs. Looking for entries marked 'web scraping' might help to keep up to date regarding legal developments and recent verdicts. The Electronic Frontier Foundation (http://www.eff.org/) was founded as early as 1990 to defend the digital rights of consumers and the public. We hope, however, that you will never have to rely on their help.

We offer some more detailed recommendations on how to behave when scraping content from the Web in Section 9.3.3.

## Acknowledgments

Many people helped to make this project possible. We would like to take the opportunity to express our gratitude to them. First of all, we would like to say thanks to Peter Selb to whom we owe the idea of creating a course on alternative data collection. It is due to his impulse that we started to assemble our somewhat haphazard experiences in a comprehensive volume. We are also grateful to several people who have provided invaluable feedback on parts of the book. Most importantly we thank Christian Breunig, Holger Döring, Daniel Eckert, Johannes

[1] To lead by example, we owe some of the suggestions to Hemenway and Calishain (2003)'s *Spidering Hacks* (Hack #6).

Kleibl, Philip Leifeld, and Nils Weidmann, whose advice has greatly improved the material. We also thank Kathryn Uhrig for proofreading the manuscript.

Early versions of the book were used in two courses on "Alternative data collection methods" and "Data collection in the World Wide Web" that took place in the summer terms of 2012 and 2013 at the University of Konstanz. We are grateful to students for their comments—and their patience with the topic, with R, and outrageous regular expressions. We would also like to thank the participants of the workshops on "Facilitating empirical research on political reforms: Automating data collection in R" held in Mannheim in December 2012 and the workshop "Automating online data collection in R," which took place in Zurich in April 2013. We thank Bruno Wüest in particular for his assistance in making the Zurich workshop possible, and Fabrizio Gilardi for his support.

It turns out that writing a volume on automating data collection is a surprisingly time-consuming endeavor. We all embarked on this project during our doctoral studies and devoted a lot of time to learning the intricacies of web scraping that could have been spent on the tasks we signed up for. We would like to thank our supervisors Peter Selb, Daniel Bochsler, Ulrich Sieberer, and Thomas Gschwend for their patience and support for our various detours. Christian Rubba is grateful for generous funding by the Swiss National Science Foundation (Grant Number 137805).

We would like to acknowledge that we are heavily indebted to the creators and maintainers of the numerous packages that are applied throughout this volume. Their continuous efforts have opened the door for new ways of scholarly research—and have provided access to vast sources of data to individual researchers. While we cannot possibly hope to mention all the package developers in these paragraphs, we would like to express our gratitude to Duncan Temple Lang and Hadley Wickham for their exceptional work. We would also like to acknowledge the work of Yihui Xie, whose package was crucial in typesetting this book.

We are grateful for the help that was extended from our publisher, particularly from Heather Kay, Debbie Jupe, Jo Taylor, Richard Davies, Baljinder Kaur and others who were responsible for proofreading and formatting and who provided support at various stages of the writing process.

Finally, we happily acknowledge the great support we received from our friends and families. We owe special and heartfelt thanks to: Karima Bousbah, Johanna Flock, Hans-Holger Friedrich, Dirk Heinecke, Stefanie Klingler, Kristin Lindemann, Verena Mack, and Alice Mohr.

*Simon Munzert*
*Christian Rubba*
*Peter Meißner*
*Dominic Nyhuis*

# Contents