Journal of Statistical Software

October 2014, Volume 61, Book Review 1. *http://www.jstatsoft.org/*

Reviewer: Simon Munzert
University of Konstanz

XML and Web Technologies for Data Sciences with R

Deborah Nolan and Duncan Temple Lang
Springer-Verlag, New York, 2014.
ISBN 978-1-4614-7899-7. 663 pp. USD 69.99 (P), USD 49.99 (E).
http://www.rxmlwebtech.org/

With their book *XML and Web Technologies for Data Sciences with R*, Deborah Nolan and Duncan Temple Lang provide an extensive introduction to the collection and processing of XML and other web data within the R programming environment. The authors pick up a remarkable development that has been taking place over the last years: We observe a rapid and profound change in what is possible in data-based science and business by exploiting new (and often web-based) sources of information. From the software perspective, R has become one of the major workhorses for modern data science work. Temple Lang has contributed several important R packages over the past years, including **RCurl** for network communication, **XML** for parsing and creating XML-style documents and **RJSONIO** for handling JSON content.

Why is this book important? In the last chapters, the authors suggest to use R to draw on the flexibility of XML for text editing, visualization and code generation purposes. For instance, they demonstrate that XML in combination with R is a powerful tool to author documents, extending what is already possible with tools like **knitr** (see Xie 2014; McNamara 2014). The book itself is their best example, as it is written with DocBook, a markup language derived from XML. However, there is more to that: The range and scope of scientifically relevant data are rapidly increasing and research on computer-based methods for the analysis of existing large data is booming across all disciplines. Consequently, knowledge about XML and web data standards enables us to tap this vast and ever-growing amount of data. Furthermore, scientific work becomes more and more interdisciplinary, which calls for consistent standards with regards to the exchange and structure of data and documents. Therefore, this book is important because it provides one of the most comprehensive introductions to these technologies and is the first one to demonstrate their use in the R environment.

The authors refrain from a clear statement who the book is aimed at, but the fact that they do not offer an introduction into the basic functionality of R (which is a good choice, given the abundance of books of that kind on the market) implies that readers are supposed to have basic familiarity with the language. The authors propagate the use of R as a programming environment that can be used for virtually every step in a data scientist's workflow – starting from data acquisition on the web to data manipulation, visual and statistical processing and finally publication and sharing. In this sense, this book is good news for any R user who is interested in working with new data sources – particularly unstructured and structured data from the web – but has so far shied away from the time-consuming barrier that learning new technologies always poses.

The first part of the book (Chapters 1–7) is dedicated to fundamentals of the data formats XML and JSON. Chapter 1 offers a very gentle introduction to some core concepts of the book, that is the import of HTML, XML and JSON content into the R workspace and a set of basic data extraction tasks. It provides a basic understanding of these languages and formats and illustrates why these are essential for data scraping tasks and the work with web APIs. Chapter 2 dives deeper into various aspects of XML, covering basic syntax rules, conceptual issues like the hierarchical structure and extensibility to other grammars, but also namespaces, schema, and DTD. Chapters 3–6 elaborate on how to deal with XML documents in R, discussing the functionality of the **XML** package from A to Z. We are introduced to various parsing techniques, to using XPath to extract data from specific nodes of an XML document, and to generating XML documents from R.

My favorite in the first part is Chapter 5. It provides an overview of **XML**'s high-level functions that quickly turn out to be very useful in practice, especially when dealing with HTML content. Furthermore, it highlights with different examples that scraping from HTML pages is a step-by-step procedure and often involves a considerable amount of trial-and-error and complex modifications of the document tree even for seasoned programmers. In that sense, one of the few complaints I have about this book is that it offers examples of failure and unexpected difficulties a bit too rarely – after all, dealing with XML and web-generated data can be very complex and cause a lot of frustration for beginners. Seeing that complex XPath expressions do not simply fall into anybody's lap can be a good lesson, too. Finally, this chapter offers a thorough view on more sophisticated tools like handler functions and SAX parsing. This should also be of interest for readers who are already familiar with the basic technologies described in the book, as these techniques are usually not explained in any of the countless introductions to web scraping and XML parsing that circulate on the web.

In Chapter 7 we are shown how to read and create JSON data using the **RJSONIO** package. We also learn how to pull JSON content from a web service into R and process it, which is probably the most common case of an encounter with JSON. Given the book's publication date, the authors cannot be blamed to not introduce the recently published **jsonlite** package (Ooms, Temple Lang, and Wallace 2014), which builds upon **RJSONIO** but improves mapping between JSON and R data structures and thus makes working with web APIs easier.

The book's second part (Chapters 8–13) covers technologies that are essential to pull data from the web. We learn how to specify HTTP requests with the **RCurl** package (Chapter 8), how we can use R to submit HTML forms (Chapter 9), how to tap REST, XML-RPC or SOAP-based web services (Chapters 10–12), and how we can authorize access to an API via OAuth using the **ROAuth** package (Chapter 13). The authors' efforts to make all of these important procedures for data collection from the web possible with R cannot be understated. The book puts the finishing touches on this achievement, as it provides markedly better access to the massive set of functions, predominantly available with the **RCurl** package, than the original documentation in the packages' R help files. Especially Sections 8.4–8.8 serve as a useful manual for more complex HTTP requests that have to exhaust the functionality of the underlying C-based **libcurl** library. What I missed though in this part is a short discussion of other work in the R community on this buzzing field. The task view on *Web Technologies and*

Services at the Comprehensive R Archive Network (Scott Chamberlain, Gandrud, and Mair 2014) lists plenty other useful tools – many of which are from the authors themselves or build upon their work, like **httr** (Wickham 2014), but also a vast amount of original contributions.

In the book's final part (Chapters 14–18), the authors offer a set of XML applications. In essence, these case studies demonstrate XML's flexibility and omnipresence by showing, for example, how we can process spreadsheets from Open or (modern) Microsoft Office which are both based on the Office Open XML standard, or how we can manipulate and visualize SVG and KML data. While each of these applications is an impressive demonstration of the capabilities of the R-XML interplay, readers who are particularly interested in web data collection might miss applications on what the other half of the book was about, e.g., the programming of R bindings to web APIs or more sophisticated scraping tasks.

The authors provide access to their code in an exemplary manner. All code snippets are made available on the accompanying website `http://www.rxmlwebtech.org/`. They even offer a complete version of the book content in DocBook format for free. In that sense, the book itself serves as a practical showcase for what it teaches. For most of the examples, corresponding data files are stored on the book's website, too. However, web data are a moving target. Like everything that rests on real-life data from the web, Nolan's and Temple Lang's book is continuously running the risk of getting outdated when the web pages that underlie many of the examples are updated, re-designed, or even removed. Nonetheless, to me, the book benefits from applications that are not artificially constructed by the authors but can be found on the web. And after all, the underlying technologies that are treated extensively will be valid in the future anyway.

As an 'unofficial' second companion to the book, `http://www.omegahat.org/` provides guidance for the dozens of packages for R-web communication that have been developed within the 'Omega Project for Statistical Computing' (with Temple Lang as its most diligent contributor) over the last years. Furthermore, the authors have tried everything to make the book's content, which is naturally characterized with text frequently interrupted by various code snippets, as accessible as possible. They use typographic conventions (using different colors and fonts) to discriminate between R, XML, JSON, XPath, JavaScript and other types of code. A glossary of functions is attached to many of the fundamental chapters, which is useful to internalize the enormous complexity of the **XML** and other packages.

In general, the book leaves not much to be desired for anyone who is interested in working with XML-based data and is familiar with the R environment. Readers who are looking for a more basic primer on the covered technologies might have a look at the excellent *Introduction to Data Technologies* by Murrell (2009), which also provides some fundamentals on database technology and regular expressions but lacks profound treatment of **RCurl** and **XML**.

Recapitulating my review, *XML and Web Technologies for Data Sciences with R* is an exceptional textbook for several reasons. It closes a remarkable gap in R-based data science literature by offering a condensed introduction to important data technologies for applied statisticians who do not want to study a multitude of computer science books. It builds upon an extensive set of packages, mostly written by Temple Lang himself, which can be seen as the gold standard software of web data handling with R. It is mostly written in a hands-on fashion, demonstrating the packages' functionality by example. And finally, it offers plenty of opportunities to see the bigger picture of data formats on the web and their immense potential by offering a multitude of examples and applications. While a weighty tome covering more than

600 pages might discourage practitioners who are generally interested in working with new data types that proliferate on the web, every prospective data scientist feeling comfortable within the R programming environment should consider this book essential reading.

References

McNamara A (2014). "Dynamic Documents with R and **knitr**." *Journal of Statistical Software, Book Reviews*, **56**(2), 1–4. URL http://www.jstatsoft.org/v56/b02/.

Murrell P (2009). *Introduction to Data Technologies*. Chapman & Hall/CRC, Boca Raton.

Ooms J, Temple Lang D, Wallace J (2014). ***jsonlite**: A Smarter JSON Encoder/Decoder for R*. R package version 0.9.11, URL http://CRAN.R-project.org/package=jsonlite.

Scott Chamberlain KR, Gandrud C, Mair P (2014). "CRAN Task View: Web Technologies and Services." Version 2014-08-29, URL http://CRAN.R-project.org/view=WebTechnologies.

Wickham H (2014). ***httr**: Tools for Working with URLs and HTTP*. R package version 0.5, URL http://CRAN.R-project.org/package=httr.

Xie Y (2014). *Dynamic Documents with R and **knitr***. Chapman & Hall/CRC, Boca Raton.

Reviewer:

Simon Munzert
University of Konstanz
Department of Politics and Public Administration
P.O. Box D85
D-78462 Konstanz, Germany
E-mail: simon.munzert@uni.kn

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 61, Book Review 1 *Published:* 2014-10-16
October 2014