An introduction to
Web Scraping and Text Mining with R

Simon Munzert
University of Konstanz

October 2014

An introduction to Web Scraping with R

Simon Munzert
University of Konstanz

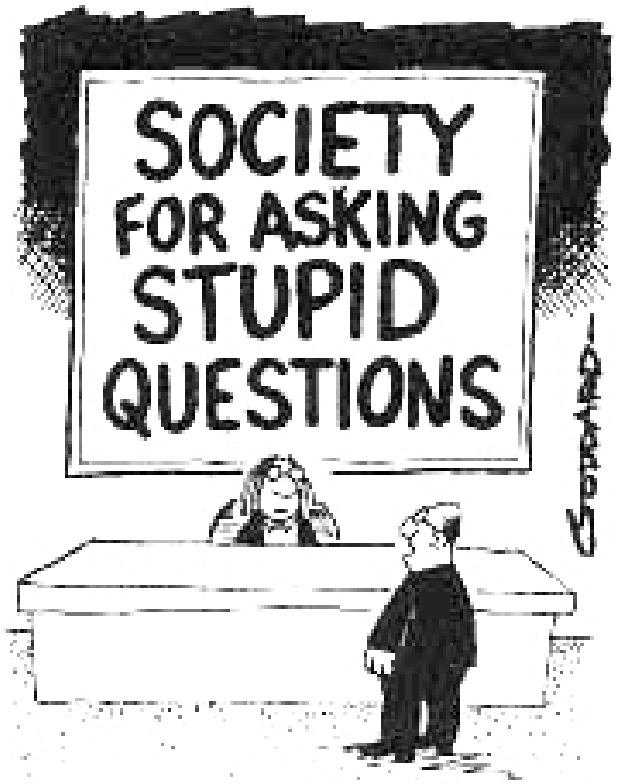
October 2014

Session overview

Session	Topics	Book chapter
Fri, 10/03	Scraping static content using...	
	...XML/HTML parsing	3
	...XPath/SelectorGadget	4
	...Regular expressions	8
Fri, 10/17	Scraping dynamic content + APIs using...	
	...JSON	3
	...APIs	9
	...AJAX	6
	...Selenium	9

What I won't cover: internals of HTTP, complex parsing techniques, OAuth, databases, advanced workflow

First: ask questions! No matter what. . .



"Excuse me, but is this The Society for Asking Stupid Questions?"

Web scraping. What? Why?

The World Wide Web is full of various kinds of new data, e.g.:

- open government data
- search engine data
- services that track social behavior

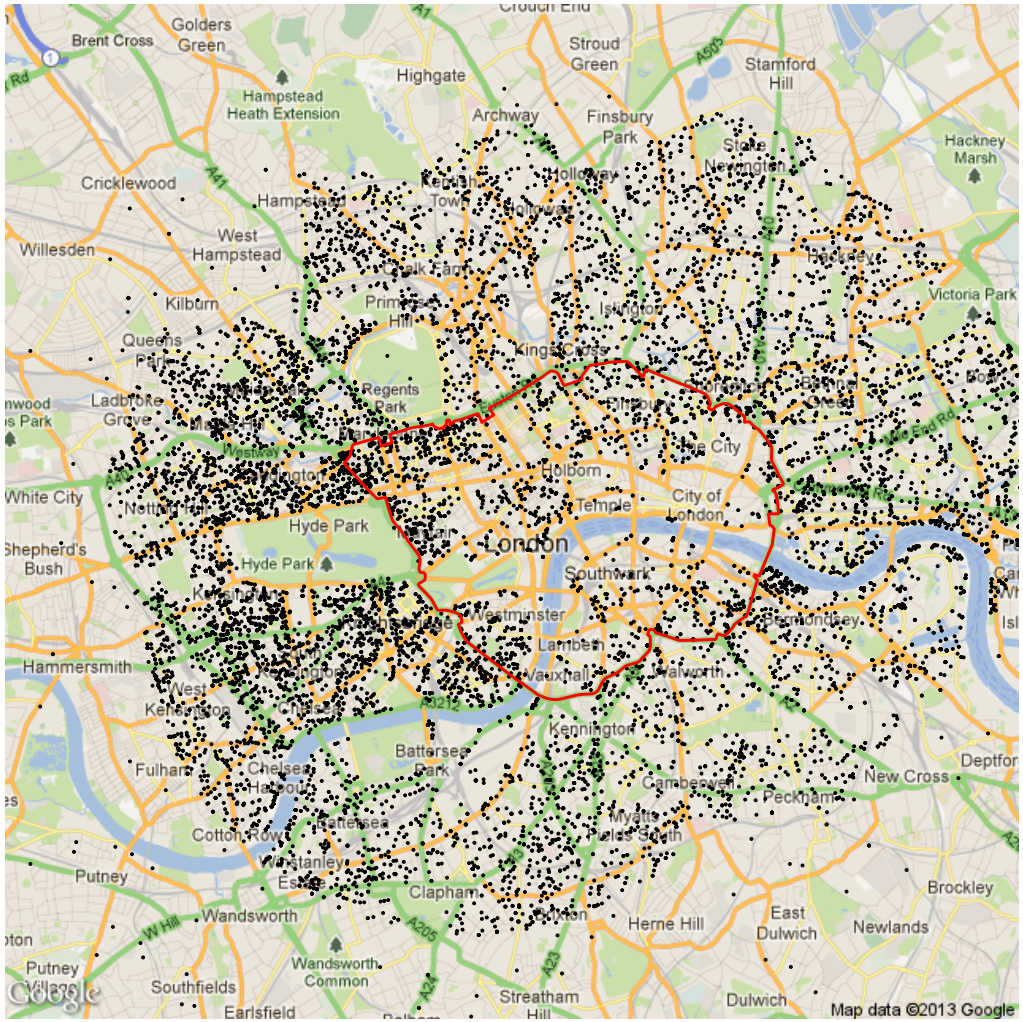
Web scraping

A.k.a. screen scraping, web harvesting. Computer-aided collection of predominantly unstructured data (e.g., from HTML code)

Practical arguments

- financial resources are sparse
- . . . and so is our time
- reproducibility

Real estate prices, London congestion charge



Data retrieved from http://www.zoopla.co.uk

Measuring issue saliency using Wikipedia page view data

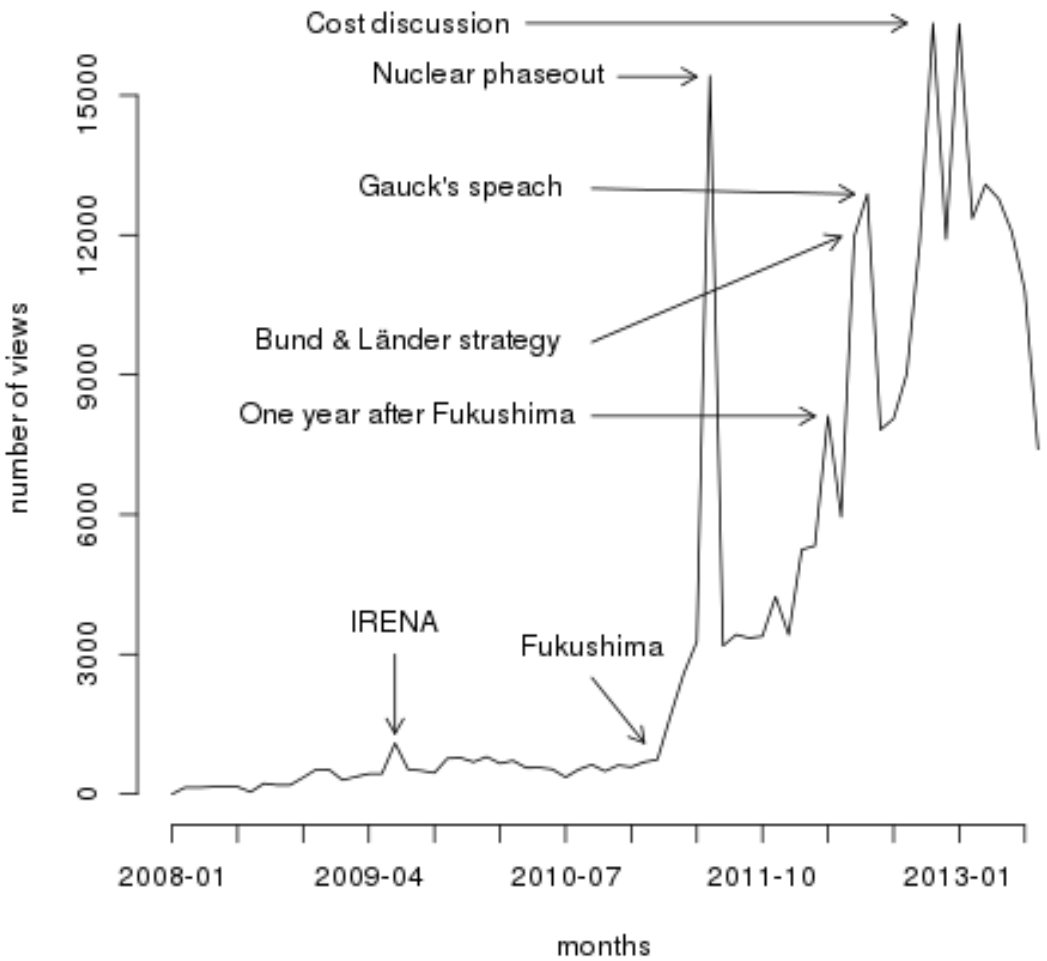
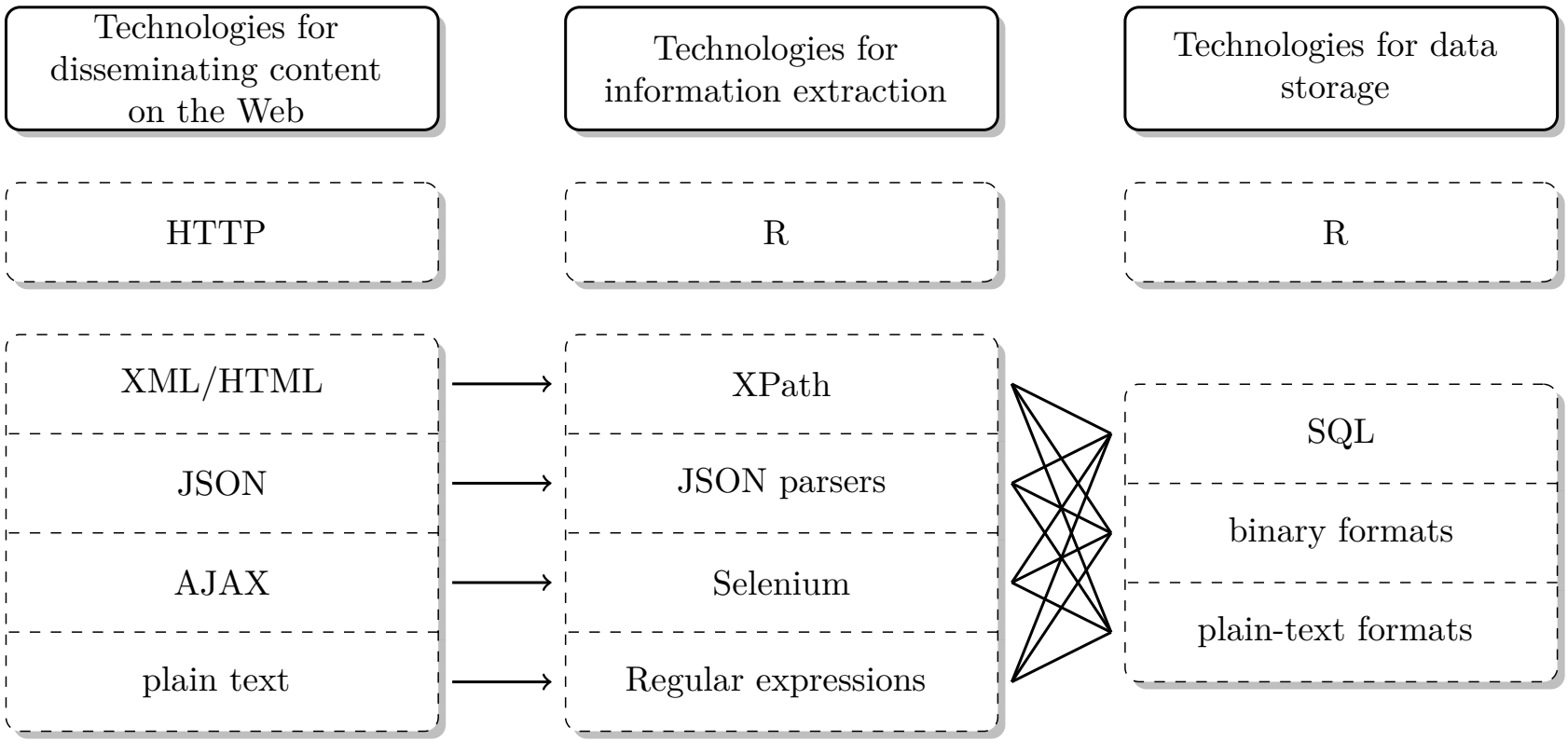


Figure 1: Wikipedia article views for "Energiewende" from January 2008 - July 2013

The philosophy behind web data collection with R

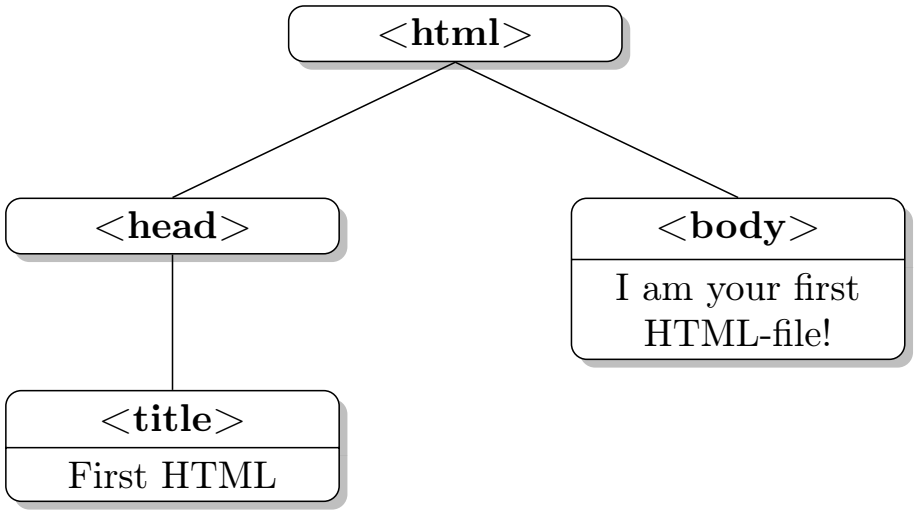
- no point-and-click procedure
- automation of download, parsing, and data extraction procedures
- classical screen scraping
- tapping of web services and APIs
- post-processing of text data
- reproducibility

Technologies of the World Wide Web



XML/HTML: tree structure

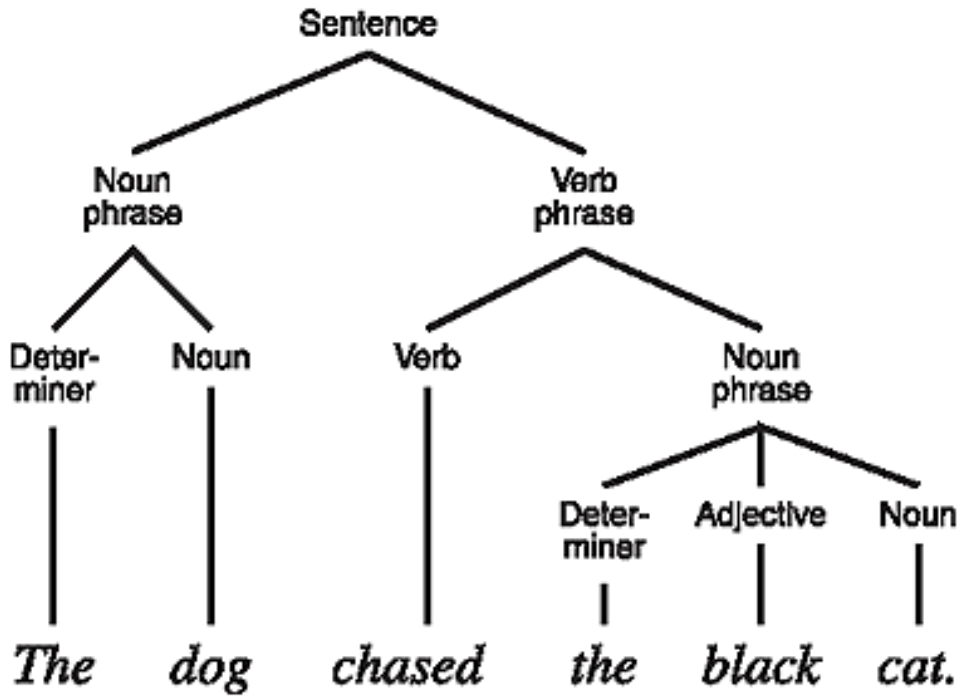
```
1  <!DOCTYPE html>
2    <html>
3      <head>
4        <title id=1>First HTML</title>
5      </head>
6      <body>
7        I am your first HTML file!
8      </body>
9    </html>
```



XML Parsing

Parsing

Syntactic analysis of text according to grammatical rules; analysis of the relationship between single parts of text. In programming, input has to be interpreted (e.g., by R) to process the command.



XML Parsing

- HTML/XML documents are human-readable
- HTML tags structure the document
- web user perspective: the browser interprets the code
- web scraper perspective: use the tags to locate information; document has to be parsed first

Parsing in R

- XML package to parse XML-style documents
- high-level functions: `htmlParse()`, `xmlParse()`
- other packages for other document types
- import via `readLines()` is not parsing - the document's structure is not retained

XPath

Definition

- XML Path language, a W3C standard
- query language for XML-style documents
- used to locate and extract content

Why XPath for web scraping?

- information is structured by layout
- not only content, but context matters
- gold standard of classical screen scraping with R

XPath and R

Definition

- XML Path language, a W3C standard
- query language for XML-style documents
- used to locate and extract content

Why XPath for web scraping?

- information is structured by layout
- not only content, but context matters
- gold standard of classical screen scraping with R

XPath and R

Procedure

- load XML package
- parse document
- query document with XPath
- XML package can 'speak' XPath!

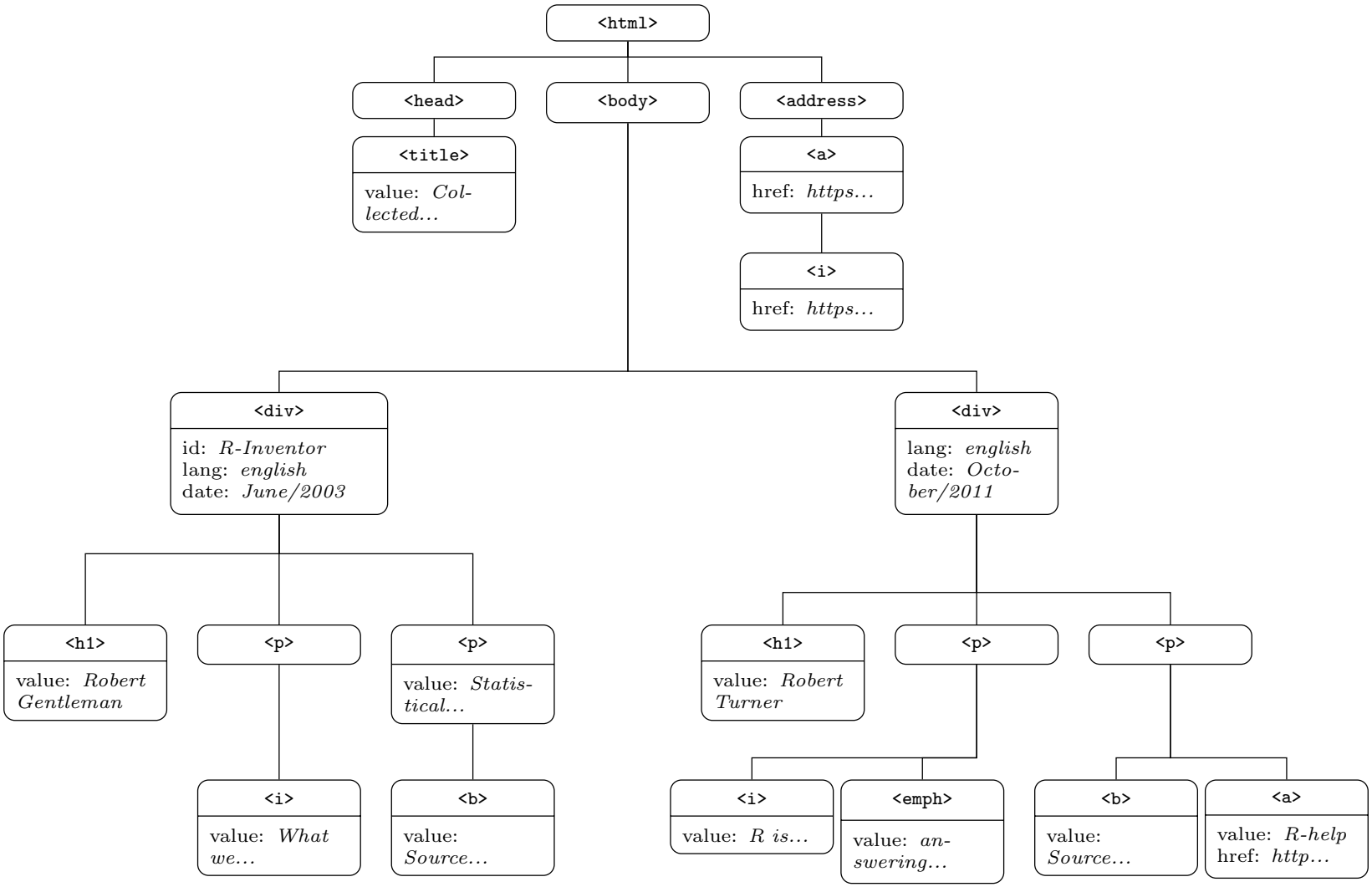
```
R> library(XML)


R> parsed_doc <- htmlParse(file = "materials/fortunes.html")


R> xpathSApply(doc = parsed_doc, path = "/html/body/div/p/i")
[[1]]
<i>'What we have is nice, but we need something very different'</i>

[[2]]
<i>'R is wonderful, but it cannot work magic'</i>
```

```
R> print(parsed_doc)
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>
<head><title>Collected R wisdoms</title></head>
<body>
<div id="R Inventor" lang="english" date="June/2003">
  <h1>Robert Gentleman</h1>
  <p><i>'What we have is nice, but we need something very different'</i
></p>
  <p><b>Source: </b>Statistical Computing 2003, Reisensburg</p>
</div>
<div lang="english" date="October/2011">
  <h1>Rolf Turner</h1>
  <p><i>'R is wonderful, but it cannot work magic'</i> <br><emph>
answering a request for automatic generation of 'data from a known mean
and 95% CI'</emph></p>
  <p><b>Source: </b><a href="https://stat.ethz.ch/mailman/listinfo/r-help
">R-help</a></p>
</div>
<address>
<a href="http://www.r-datacollection.com"><i>The book homepage</i></a><a
></a>
</address>
</body>
</html>
```


<html>
<head>
<body>
<address>
<title>
value: Collected...
<a>
href: https...
<i>
href: https...

id: R-Inventor
lang: english
date: June/2003

lang: english
date: October/2011
<h1>
value: Robert Gentleman
<p>
<p>
value: Statistical...
<h1>
value: Robert Turner
<p>
<p>
<i>
value: What we...
<b>
value: Source...
<i>
value: R is...
<emph>
value: answering...
<b>
value: Source...
<a>
value: R-help
href: http...

R's functionality for working with the Web

- managing file downloads
- import and parsing of XML and JSON content
- tapping REST-based web services
- authentication via OAuth
- communication via HTTP, HTTPS, FTP, ...
- automated browsing

For an extensive and up-to-date overview, see:
http://cran.r-project.org/web/views/WebTechnologies.html

Hands-on web scraping with R

You need

- R + Editor (RStudio)
- R packages: RCurl, XML, stringr, plyr, ggplot2
- R code and data from https://github.com/simonmunzert/rscraping-intro-duke
- Internet access

Web scraping etiquette