

Wk 7: Working with XML, HTML and JSON in R

Dirk Hartog

2023-10-11

```
library(tidyverse)
library(jsonlite)
library(XML)
library(xml2)
library(stringr)
library(rvest)
```

#Pick three of your favorite books on one of your favorite subjects. At least one of the books should have more than one author. For each book, include the title, authors, and two or three other attributes that you find interesting.

Step 1: Create three files which store the book's information in HTML (using an html table), XML, and JSON formats and read the information from the web (in this case Github).

JSON file

```
# libraries used: jsonlite, dplyr, stringr

json_url <- "https://raw.githubusercontent.com/D-hartog/DATA607/main/WK7/books.json"
json_file <- list(read_json(json_url))
books_js <- tibble(books = json_file)

# Unnest the list object from the json file

books_js <- books_js %>% unnest_wider(books) %>%
  unnest_longer(books) %>%
  unnest_wider(books)

# Expand the cell with a list of two authors

# -- Unlist that cell into a vector
authors <- unlist(books_js$author_s[2])

# -- flatten the vector to create a string of the two authors and re-assign it into the cell
books_js$author_s[2] <- str_flatten(authors, ",")

# -- Use separate rows function to separate the rows into two (one with each author)
books_js <- books_js %>% separate_rows(author_s, sep = ",")
```

```
# -- Clean up column names and values
colnames(books_js) <- books_js %>% names() %>% str_to_upper()

books_js$AUTHOR_S <- str_trim(books_js$AUTHOR_S)

books_js

## # A tibble: 5 x 9
##   TITLE      TYPE  AUTHOR_S RATING NO_RATINGS NO_PAGES PUBLISHER PUBLICATION_DATE
##   <chr>      <chr> <chr>      <dbl>      <int>      <int> <chr>      <chr>
## 1 Unfolding Hard~ Matthew~    4.6         27        256 Thames &~ February 16, 20~
## 2 The Infi~ Hard~ Meher M~    4.2         70         96 Tuttle P~ August 6, 2013
## 3 The Infi~ Hard~ Robert ~    4.2         70         96 Tuttle P~ August 6, 2013
## 4 Folding ~ Pape~ Paul Ja~    4.5        516        375 Laurence~ May 2, 2011
## 5 Cut and ~ Pape~ Paul Ja~    4.5         39        128 Laurence~ January 24, 2017
## # i 1 more variable: AMAZON_LINK <chr>
```

HTML file

```
# libraries used: rvest, dplyr, stringr

html_url <- "https://raw.githubusercontent.com/D-hartog/DATA607/main/WK7/books.html"

# read in the file from github
html_file <- read_html(html_url)

books_html <- html_file %>%
  html_table(fill = TRUE)

books_html <- books_html[[1]]

# Use seperate rows function to separate the rows into two (one with each author)
books_html <- books_html %>% separate_rows(author_s, sep = ",")

# Clean up column names and values
colnames(books_html) <- books_html %>% names() %>% str_to_upper()

books_html$AUTHOR_S <- str_trim(books_html$AUTHOR_S)

books_html
```

```
## # A tibble: 5 x 9
##   TITLE      TYPE  AUTHOR_S RATING NO_RATINGS NO_PAGES PUBLISHER PUBLICATION_DATE
##   <chr>      <chr> <chr>      <dbl>      <int>      <int> <chr>      <chr>
## 1 Unfolding Hard~ Matthew~    4.6         27        256 Thames &~ February 16, 20~
## 2 Folding ~ Hard~ Meher M~    4.2         70         96 Tuttle P~ August 6, 2013
## 3 Folding ~ Hard~ Robert ~    4.2         70         96 Tuttle P~ August 6, 2013
## 4 Folding ~ Pape~ Paul Ja~    4.5        516        375 Laurence~ May 2, 2011
## 5 Cut and ~ Pape~ Paul Ja~    4.5         39        128 Laurence~ January 24, 2017
## # i 1 more variable: AMAZON_LINK <chr>
```

XML file

```

# libraries used: xml2, dplyr, stringr

xml_url <- "https://raw.githubusercontent.com/D-hartog/DATA607/main/WK7/books.xml"

# Read in xml file from url
xml_file <- read_xml(xml_url)

# Extract all the book nodes, transform them into a list and bind_rows()
books_xml <- xml_file %>%
  xml_find_all("./book") %>%
  as_list() %>%
  bind_rows()

# Use separate_rows function to separate the rows into two (one with each author)
books_xml <- books_xml %>% separate_rows(author_s, sep = ",")

# Clean up column names and values
colnames(books_xml) <- books_xml %>% names() %>% str_to_upper()

books_xml$AUTHOR_S <- str_trim(books_xml$AUTHOR_S)

books_xml

```

```

## # A tibble: 5 x 9
##   TITLE      TYPE  AUTHOR_S RATING NO_RATINGS NO_PAGES PUBLISHER PUBLICATION_DATE
##   <list>    <lis> <chr>    <list> <list>    <list>  <list>    <list>
## 1 <chr [1]> <chr> Matthew~ <chr>  <chr [1]> <chr>    <chr [1]> <chr [1]>
## 2 <chr [1]> <chr> Meher M~ <chr>  <chr [1]> <chr>    <chr [1]> <chr [1]>
## 3 <chr [1]> <chr> Robert ~ <chr>  <chr [1]> <chr>    <chr [1]> <chr [1]>
## 4 <chr [1]> <chr> Paul Ja~ <chr>  <chr [1]> <chr>    <chr [1]> <chr [1]>
## 5 <chr [1]> <chr> Paul Ja~ <chr>  <chr [1]> <chr>    <chr [1]> <chr [1]>
## # i 1 more variable: AMAZON_LINK <list>

```

For the xml and html files one the data frame was constructed, they did both look similar. When initially constructing the dataframe with the json file, the book with two authors were contain in a list within the author column. An extra step was required to unnest that list. Now all the data frames look the same.