Assignment 7: Working with XML and JSON in R

Andrew Carson

October 11, 2016

Task

- 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.
- Take the information that you've selected about these three books, and separately create three files which store the book's information in HTML (using an html table), XML, and JSON formats (e.g. "books.html", "books.xml", and "books.json"). To help you better understand the different file structures, I'd prefer that you create each of these files "by hand" unless you're already very comfortable with the file formats.
- Write R code, using your packages of choice, to load the information from each of the three sources into separate R data frames. Are the three data frames identical?
- Your deliverable is the three source files and the R code. If you can, package your assignment solution up into an .Rmd file and publish to rpubs.com. [This will also require finding a way to make your three text files accessible from the web].

Solution

To begin, I created a target table that the other tables should match when completed. The subject matter is philosophy, and the four books I have selected are some of those that I read in my undergraduate and graduate philosophy studies. Here is the table that we want to mimic:

```
library(DT)
library(stringr)

#load the target data
target<-read.csv("https://raw.githubusercontent.com/anrcarson/CUNY-MSDA/master/DATA607/Assignment7/Assi

#show target data
datatable(target)
```

Now we work on creating the identical tables using HTML, XML, and JSON.

HTML

I created the html file by hand. Below is the code for the table:

```
<html>
    <head>
        <title>books</title>
    </head>
    <body>
        <table>
```

```
            <tr>    <th>BookTitle</th> <th>Author</th> <th>PhilosophySubject</th> <th>Pages</th> <th>Ti
            <tr>    <td>The Republic</td> <td>Plato</td> <td>Political Philosophy</td> <td>416</td> <td
            <tr>    <td>Metaphysics</td> <td>Peter Van Inwagen</td> <td>Metaphysics</td> <td>329</td> <
            <tr>    <td>Consolation of Philosophy</td> <td>Boethius</td> <td>Philosophy of Religion</td
            <tr>    <td>What's Wrong?  Applied Ethicists and their Critics</td> <td>David Boonin, Graha
        </table>
    </body>
</html>
```

I saved the file to GitHub, so we can download it from there. Then we can import the file from the local copy of the file. I use the `XML` package to do so:

```
#load package
library(XML)


#download file from:
# https://raw.githubusercontent.com/anrcarson/CUNY-MSDA/master/DATA607/Assignment7/books.html

#parse file from local location
html_parsed<-htmlParse(file = "C:/Users/Andy/Desktop/Personal/Learning/CUNY/DATA607/books.html")

#create data frame
html_table<-readHTMLTable(html_parsed, stringsAsFactors = FALSE)
html_table<-html_table[[1]]

#view
datatable(html_table)
```

XML

I created an XML file with the same information. Below is the code for the table:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<books>
    <book id="1">
        <BookTitle>The Republic</BookTitle>
        <Author>Plato</Author>
        <PhilosophySubject>Political Philosophy</PhilosophySubject>
        <Pages>416</Pages>
        <TimePeriod>Ancient</TimePeriod>
    </book>
    <book id="2">
        <BookTitle>Metaphysics</BookTitle>
        <Author>Peter Van Inwagen</Author>
        <PhilosophySubject>Metaphysics</PhilosophySubject>
        <Pages>329</Pages>
        <TimePeriod>Contemporary</TimePeriod>
    </book>
    <book id="3">
        <BookTitle>Consolation of Philosophy</BookTitle>
        <Author>Boethius</Author>
        <PhilosophySubject>Philosophy of Religion</PhilosophySubject>
```

```
        <Pages>216</Pages>
        <TimePeriod>Medieval</TimePeriod>
    </book>
    <book id="4">
        <BookTitle>What's Wrong?  Applied Ethicists and their Critics</BookTitle>
        <Author>David Boonin, Graham Oddie</Author>
        <PhilosophySubject>Applied Ethics</PhilosophySubject>
        <Pages>618</Pages>
        <TimePeriod>Contemporary</TimePeriod>
    </book>
</books>
```

I saved the file to GitHub, so we can download it from there. Then we can import the file from the local copy of the file. I use the `XML` package to do so:

```
#download file from:
# https://raw.githubusercontent.com/anrcarson/CUNY-MSDA/master/DATA607/Assignment7/books.xml

#parse file from local location
xml_parsed<-xmlParse(file = "C:/Users/Andy/Desktop/Personal/Learning/CUNY/DATA607/books.xml")

#create data frame
xml_table<-xmlToDataFrame(xml_parsed, stringsAsFactors = FALSE)

#view
datatable(xml_table)
```

JSON

I created a JSON file with the same information. Below is the code for the table:

```
{"books" :[
    {
    "BookTitle" : "The Republic",
    "Author" : "Plato",
    "PhilosophySubject" : "Political Philosophy",
    "Pages" : 416,
    "TimePeriod" : "Ancient"
    },
    {
    "BookTitle" : "Metaphysics",
    "Author" : "Peter Van Inwagen",
    "PhilosophySubject" : "Metaphysics",
    "Pages" : 329,
    "TimePeriod" : "Contemporary"
    },
    {
    "BookTitle" : Consolation of Philosophy",
    "Author" : "Boethius",
    "PhilosophySubject" : "Philosophy of Religion",
    "Pages" : 216,
    "TimePeriod" : "Medieval"
    },
```

```
    {
    "BookTitle" : "What's Wrong?  Applied Ethicists and their Critics",
    "Author" : "David Boonin, Graham Oddie",
    "PhilosophySubject" : "Applied Ethics",
    "Pages" : 618,
    "TimePeriod" : "Contemporary"
    }]
}
```

I saved the file to GitHub, so we can download it from there. Then we can import the file from the local copy of the file. I use the `RJSONIO` package to do so:

```
#load package
library(RJSONIO)

#download file from:
# https://github.com/anrcarson/CUNY-MSDA/blob/master/DATA607/Assignment7/books

#parse file from local location
json_parsed<-fromJSON(content = "C:/Users/Andy/Desktop/Personal/Learning/CUNY/DATA607/books.json")

#create data frame.  Credit: page 74 in "Automated Data Collection with R"
json_table<-do.call("rbind", lapply(json_parsed[[1]], data.frame, stringsAsFactors = FALSE))

#view
datatable(json_table)
```

Identical?

Are the three tables identical with the original table?

```
all.equal(target, html_table)
```

```
## [1] "Component \"Pages\": Modes: numeric, character"
## [2] "Component \"Pages\": target is numeric, current is character"
```

```
all.equal(target, xml_table)
```

```
## [1] "Component \"Pages\": Modes: numeric, character"
## [2] "Component \"Pages\": target is numeric, current is character"
```

```
all.equal(target, json_table)
```

```
## [1] TRUE
```

The HTML and XML tables imported the numeric values in `Pages` as characters. We can change this by casting each to integer.

```
#cast to integer
html_table$Pages<-as.integer(html_table$Pages)
xml_table$Pages<-as.integer(xml_table$Pages)

all.equal(target, html_table)
```

```
## [1] TRUE
```

```
all.equal(target, xml_table)
```

```
## [1] TRUE
```

```
all.equal(target, json_table)
```

```
## [1] TRUE
```

Now they are all identical to the target table.

Additional

The task text asks us to include at least one book with multiple authors. In the above, I have placed both authors for "What's Wrong? Applied Ethicists and their Critics" in a single character string. I doubt I'd see this separated in an HTML table, but in an XML or JSON file, this is very likely. So I repeat the above for the XML and JSON files, this time, with the authors separated.

XML

```
#download file from:
# https://raw.githubusercontent.com/anrcarson/CUNY-MSDA/master/DATA607/Assignment7/books2.xml

#parse file from local location
xml2_parsed<-xmlParse(file = "C:/Users/Andy/Desktop/Personal/Learning/CUNY/DATA607/books2.xml")

#view parsed file
xml2_parsed
```

```
## <?xml version="1.0"?>
## <books>
##   <book id="1">
##     <BookTitle>The Republic</BookTitle>
##     <Author>Plato</Author>
##     <PhilosophySubject>Political Philosophy</PhilosophySubject>
##     <Pages>416</Pages>
##     <TimePeriod>Ancient</TimePeriod>
##   </book>
##   <book id="2">
##     <BookTitle>Metaphysics</BookTitle>
##     <Author>Peter Van Inwagen</Author>
##     <PhilosophySubject>Metaphysics</PhilosophySubject>
##     <Pages>329</Pages>
##     <TimePeriod>Contemporary</TimePeriod>
##   </book>
##   <book id="3">
##     <BookTitle>Consolation of Philosophy</BookTitle>
##     <Author>Boethius</Author>
##     <PhilosophySubject>Philosophy of Religion</PhilosophySubject>
```

```
##     <Pages>216</Pages>
##     <TimePeriod>Medieval</TimePeriod>
##   </book>
##   <book id="4">
##     <BookTitle>What's Wrong?  Applied Ethicists and their Critics</BookTitle>
##     <Author>
##       <Author1>David Boonin</Author1>
##       <Author2>Graham Oddie</Author2>
##     </Author>
##     <PhilosophySubject>Applied Ethics</PhilosophySubject>
##     <Pages>618</Pages>
##     <TimePeriod>Contemporary</TimePeriod>
##   </book>
## </books>
##
```

```
#create data frame
xml2_table<-xmlToDataFrame(xml2_parsed, stringsAsFactors = FALSE)

#split author for book 4 and assign back to table
xml2_table$Author[4]<-
  str_c(xmlValue(xpathSApply(xml2_parsed,"//Author1")[[1]])
    ,xmlValue(xpathSApply(xml2_parsed,"//Author2")[[1]])
    , sep=", ")

#view
datatable(xml2_table)
```

JSON

```
#download file from:
# https://raw.githubusercontent.com/anrcarson/CUNY-MSDA/master/DATA607/Assignment7/books2.json

#parse file from local location
json2_parsed<-fromJSON(content = "C:/Users/Andy/Desktop/Personal/Learning/CUNY/DATA607/books2.json")

#view parsed data
json2_parsed
```

```
## $books
## $books[[1]]
## $books[[1]]$BookTitle
## [1] "The Republic"
##
## $books[[1]]$Author
## [1] "Plato"
##
## $books[[1]]$PhilosophySubject
## [1] "Political Philosophy"
##
## $books[[1]]$Pages
## [1] 416
##
```

```
## $books[[1]]$TimePeriod
## [1] "Ancient"
##
##
## $books[[2]]
## $books[[2]]$BookTitle
## [1] "Metaphysics"
##
## $books[[2]]$Author
## [1] "Peter Van Inwagen"
##
## $books[[2]]$PhilosophySubject
## [1] "Metaphysics"
##
## $books[[2]]$Pages
## [1] 329
##
## $books[[2]]$TimePeriod
## [1] "Contemporary"
##
##
## $books[[3]]
## $books[[3]]$BookTitle
## [1] "Consolation of Philosophy"
##
## $books[[3]]$Author
## [1] "Boethius"
##
## $books[[3]]$PhilosophySubject
## [1] "Philosophy of Religion"
##
## $books[[3]]$Pages
## [1] 216
##
## $books[[3]]$TimePeriod
## [1] "Medieval"
##
##
## $books[[4]]
## $books[[4]]$BookTitle
## [1] "What's Wrong?  Applied Ethicists and their Critics"
##
## $books[[4]]$Author
## [1] "David Boonin" "Graham Oddie"
##
## $books[[4]]$PhilosophySubject
## [1] "Applied Ethics"
##
## $books[[4]]$Pages
## [1] 618
##
## $books[[4]]$TimePeriod
## [1] "Contemporary"
```

```
#create data frame.
json2_table<-do.call("rbind", lapply(json2_parsed[[1]], data.frame, stringsAsFactors = FALSE))

#get both authors for book 4 and assign to one row.  delete other row
json2_table$Author[4]<-
  str_c(json2_table$Author[which(json2_table$BookTitle=="What's Wrong?  Applied Ethicists and their Cri
json2_table<-json2_table[1:4,]

#view
datatable(json2_table)
```

We change the character values to numeric and check if they are identical.

```
#cast to integer
xml2_table$Pages<-as.integer(xml2_table$Pages)

all.equal(target, xml2_table)
```

```
## [1] TRUE
```

```
all.equal(target, json2_table)
```

```
## [1] TRUE
```

We have now completed the task. All of the data is in R data frames.