



Reading XML

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

XML

- Extensible markup language
- Frequently used to store structured data
- Particularly widely used in internet applications
- Extracting XML is the basis for most web scraping
- Components
 - Markup - labels that give the text structure
 - Content - the actual text of the document

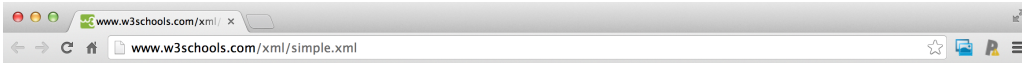
<http://en.wikipedia.org/wiki/XML>

Tags, elements and attributes

- Tags correspond to general labels
 - Start tags `<section>`
 - End tags `</section>`
 - Empty tags `<line-break />`
- Elements are specific examples of tags
 - `<Greeting> Hello, world </Greeting>`
- Attributes are components of the label
 - ``
 - `<step number="3"> Connect A to B. </step>`

<http://en.wikipedia.org/wiki/XML>

Example XML file



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<!-- Edited by XMLSpy® -->
▼ <breakfast_menu>
  ▼ <food>
    <name>Belgian Waffles</name>
    <price>$5.95</price>
    ▼ <description>
      Two of our famous Belgian Waffles with plenty of real maple syrup
    </description>
    <calories>650</calories>
  </food>
  ▼ <food>
    <name>Strawberry Belgian Waffles</name>
    <price>$7.95</price>
    ▼ <description>
      Light Belgian waffles covered with strawberries and whipped cream
    </description>
    <calories>900</calories>
  </food>
  ▼ <food>
    <name>Berry-Berry Belgian Waffles</name>
    <price>$8.95</price>
    ▼ <description>
      Light Belgian waffles covered with an assortment of fresh berries and whipped cream
    </description>
    <calories>900</calories>
  </food>
  ▼ <food>
    <name>French Toast</name>
    <price>$4.50</price>
    ▼ <description>
      Thick slices made from our homemade sourdough bread
    </description>
    <calories>600</calories>
  </food>
  ▼ <food>
    <name>Homestyle Breakfast</name>
    <price>$6.95</price>
    ▼ <description>
      Two eggs, bacon or sausage, toast, and our ever-popular hash browns
    </description>
    <calories>950</calories>
```

<http://www.w3schools.com/xml/simple.xml>

Read the file into R

```
library(XML)
fileUrl <- "http://www.w3schools.com/xml/simple.xml"
doc <- xmlTreeParse(fileUrl,useInternal=TRUE)
rootNode <- xmlRoot(doc)
xmlName(rootNode)
```

```
[1] "breakfast_menu"
```

```
names(rootNode)
```

```
food food food food food
"food" "food" "food" "food" "food"
```

Directly access parts of the XML document

```
rootNode[[1]]
```

```
<food>
  <name>Belgian Waffles</name>
  <price>$5.95</price>
  <description>Two of our famous Belgian Waffles with plenty of real maple syrup</description>
  <calories>650</calories>
</food>
```

```
rootNode[[1]][[1]]
```

```
<name>Belgian Waffles</name>
```

Programmatically extract parts of the file

```
xmlSApply(rootNode,xmlValue)
```

```
"Belgian Waffles$5.95Two of our famous Belgian Waffles with plenty of rea
```

```
"Strawberry Belgian Waffles$7.95Light Belgian waffles covered with strawberries and
```

```
"Berry-Berry Belgian Waffles$8.95Light Belgian waffles covered with an assortment of fresh berries and
```

```
"French Toast$4.50Thick slices made from our homemade so
```

```
"Homestyle Breakfast$6.95Two eggs, bacon or sausage, toast, and our ever-popula
```

Programmatically extract parts of the file

```
xmlSApply(rootNode,xmlValue)
```

```
"Belgian Waffles$5.95Two of our famous Belgian Waffles with plenty of rea
```

```
"Strawberry Belgian Waffles$7.95Light Belgian waffles covered with strawberries and
```

```
"Berry-Berry Belgian Waffles$8.95Light Belgian waffles covered with an assortment of fresh berries and
```

```
"French Toast$4.50Thick slices made from our homemade so
```

```
"Homestyle Breakfast$6.95Two eggs, bacon or sausage, toast, and our ever-popula
```


XPath

- */node* Top level node
- *//node* Node at any level
- *node[@attr-name]* Node with an attribute name
- *node[@attr-name='bob']* Node with attribute name attr-name='bob'

Information from: <http://www.stat.berkeley.edu/~statcur/Workshop2/Presentations/XML.pdf>

Get the items on the menu and prices

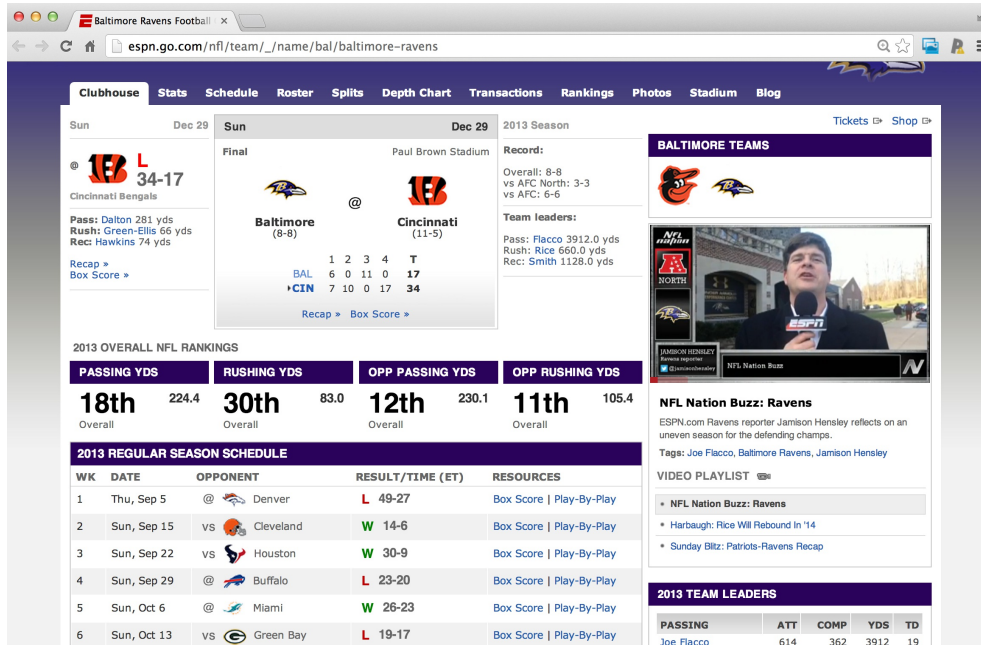
```
xpathSApply(rootNode,"//name",xmlValue)
```

```
[1] "Belgian Waffles"          "Strawberry Belgian Waffles" "Berry-Berry Belgian Waffles"  
[4] "French Toast"             "Homestyle Breakfast"
```

```
xpathSApply(rootNode,"//price",xmlValue)
```

```
[1] "$5.95" "$7.95" "$8.95" "$4.50" "$6.95"
```

Another example



The screenshot shows the ESPN website's Baltimore Ravens team page. The browser address bar displays http://espn.go.com/nfl/team/_/name/bal/baltimore-ravens. The page features a navigation bar with links for Clubhouse, Stats, Schedule, Roster, Splits, Depth Chart, Transactions, Rankings, Photos, Stadium, and Blog. The main content area is divided into several sections: a recent game recap for the Baltimore Ravens vs. Cincinnati Bengals (34-17), a 2013 Season overview with overall and team leaders statistics, a 2013 Overall NFL Rankings section showing the Ravens at 18th in passing yards and 11th in rushing yards, a 2013 Regular Season Schedule table, and a 2013 Team Leaders table. A video player on the right shows a reporter, Jamison Hensley, with the title 'NFL Nation Buzz: Ravens'.

Clubhouse Stats Schedule Roster Splits Depth Chart Transactions Rankings Photos Stadium Blog

Sun Dec 29

BAL **L** **34-17**
Cincinnati Bengals

Pass: Dalton 281 yds
Rush: Green-Ellis 66 yds
Rec: Hawkins 74 yds

Recap > Box Score >

Final Paul Brown Stadium

Baltimore (8-8) @ **Cincinnati** (11-5)

	1	2	3	4	T
BAL	6	0	11	0	17
CIN	7	10	0	17	34

Recap > Box Score >

2013 OVERALL NFL RANKINGS

PASSING YDS	RUSHING YDS	OPP PASSING YDS	OPP RUSHING YDS
18th 224.4 Overall	30th 83.0 Overall	12th 230.1 Overall	11th 105.4 Overall

2013 REGULAR SEASON SCHEDULE

WK	DATE	OPPONENT	RESULT/TIME (ET)	RESOURCES
1	Thu, Sep 5	@ Denver	L 49-27	Box Score Play-By-Play
2	Sun, Sep 15	vs Cleveland	W 14-6	Box Score Play-By-Play
3	Sun, Sep 22	vs Houston	W 30-9	Box Score Play-By-Play
4	Sun, Sep 29	@ Buffalo	L 23-20	Box Score Play-By-Play
5	Sun, Oct 6	@ Miami	W 26-23	Box Score Play-By-Play
6	Sun, Oct 13	vs Green Bay	L 19-17	Box Score Play-By-Play

2013 TEAM LEADERS

PASSING	ATT	COMP	YDS	TD
Joe Flacco	614	362	3912	19

BALTIMORE TEAMS

Record:
Overall: 8-8
vs AFC North: 3-3
vs AFC: 6-6

Team leaders:
Pass: Flacco 3912.0 yds
Rush: Rice 660.0 yds
Rec: Smith 1128.0 yds

NFL Nation Buzz: Ravens
ESPN.com Ravens reporter Jamison Hensley reflects on an uneven season for the defending champs.
Tags: Joe Flacco, Baltimore Ravens, Jamison Hensley

VIDEO PLAYLIST

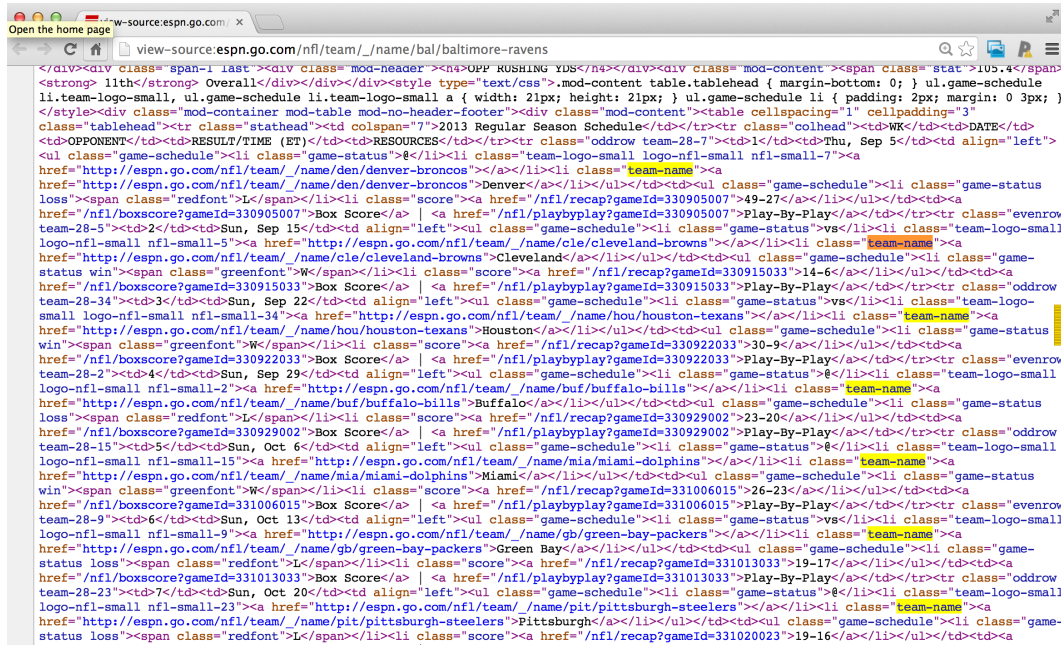
- NFL Nation Buzz: Ravens
- Harbaugh: Rice Will Rebound In '14
- Sunday Blitz: Patriots-Ravens Recap

2013 TEAM LEADERS

PASSING	ATT	COMP	YDS	TD
Joe Flacco	614	362	3912	19

http://espn.go.com/nfl/team/_/name/bal/baltimore-ravens

Viewing the source



http://espn.go.com/nfl/team/_/name/bal/baltimore-ravens

Extract content by attributes

```
fileUrl <- "http://espn.go.com/nfl/team/_/name/bal/baltimore-ravens"
doc <- htmlTreeParse(fileUrl,useInternal=TRUE)
scores <- xpathSApply(doc,"//li[@class='score']",xmlValue)
teams <- xpathSApply(doc,"//li[@class='team-name']",xmlValue)
scores
```

```
[1] "49-27"    "14-6"     "30-9"     "23-20"    "26-23"    "19-17"    "19-16"    "24-18"
[9] "20-17 OT" "23-20 OT" "19-3"     "22-20"    "29-26"    "18-16"    "41-7"     "34-17"
```

teams

```
[1] "Denver"      "Cleveland"  "Houston"    "Buffalo"    "Miami"      "Green Bay"
[7] "Pittsburgh"  "Cleveland"  "Cincinnati" "Chicago"    "New York"   "Pittsburgh"
[13] "Minnesota"   "Detroit"    "New England" "Cincinnati"
```

Notes and further resources

- Official XML tutorials [short](#), [long](#)
- [An outstanding guide to the XML package](#)