



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® -->
<?xml version="1.0" encoding="UTF-8" ?>
<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>
  </food>
</breakfast_menu>
```

<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

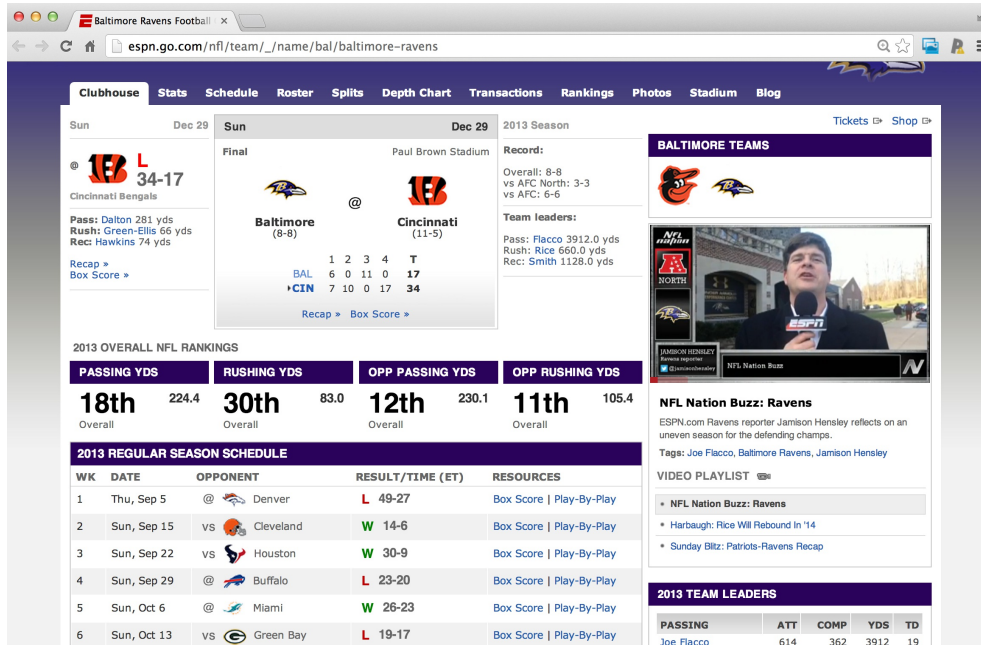
```
xpathSApply(rootNode,"//name",xmlValue)
```

```
[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 the URL 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 Ravens' 34-17 victory over the Cincinnati Bengals; a 2013 season record of 8-8; team leaders for passing and rushing; 2013 overall NFL rankings (18th in passing, 30th in rushing, 12th in opponent passing, and 11th in opponent rushing); a 2013 regular season schedule table; a video player for an NFL Nation Buzz segment; and a 2013 team leaders table.

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

Sun Dec 29

BAL 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 Overall	224.4	30th Overall	83.0
12th Overall	230.1	11th Overall	105.4

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

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

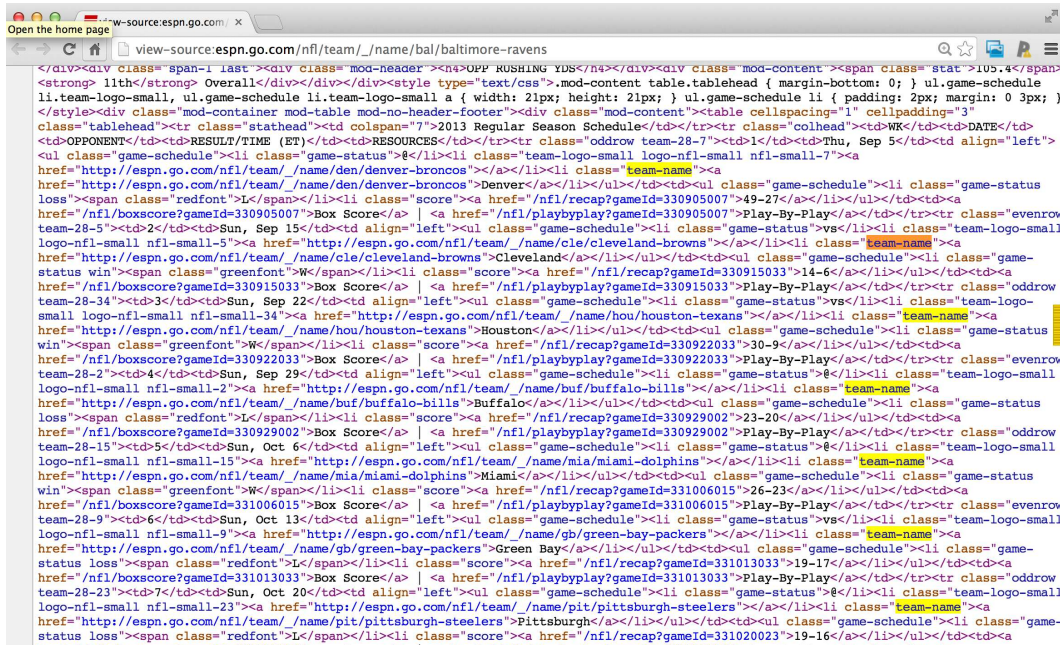
Videos: NFL Nation Buzz: Ravens
Harbaugh: Rice Will Rebound In '14
Sunday Blitz: Patriots-Ravens Recap

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



```
</div><div class="span-1 last"><div class="mod-header"><n>UPP RUSHING YDS</n></div><div class="mod-content"><span class="stat">105.4</span>
<strong> 11th</strong> Overall</div></div><div><style type="text/css">.mod-content table.tablehead { margin-bottom: 0; } ul.game-schedule
li.team-logo-small, ul.game-schedule li.team-logo-small a { width: 21px; height: 21px; } ul.game-schedule li { padding: 2px; margin: 0 3px; }
</style><div class="mod-container mod-table mod-no-header-footer"><div class="mod-content"><table cellpadding="1" cellspacing="3"
class="tablehead"><tr class="thead"><td colspan="7">2013 Regular Season Schedule</td></tr><tr class="colhead"><td>WK</td><td>DATE</td>
<td>OPPONENT</td><td>RESULT</td><td>TIME (ET)</td><td>RESOURCES</td><td></tr><tr class="odddrow team-28-7"><td>1</td><td>10/27</td><td>Thu, Sep 5</td><td><td align="left">
<ul class="game-schedule"><li class="game-status"><li class="team-logo-small nfl-small-7"><a
href="http://espn.go.com/nfl/team/_/name/den/denver-broncos"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/den/denver-broncos">Denver</a></li></ul></td><td><ul class="game-schedule"><li class="game-status
loss"><span class="redfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=330905007">49-27</a></li></ul></td><td><a
href="/nfl/boxscore?gameId=330905007">Box Score</a> | <a href="/nfl/playbyplay?gameId=330905007">Play-By-Play</a></td></tr><tr class="evenrow
team-28-5"><td>2</td><td>10/27</td><td>Sun, Sep 15</td><td><td align="left"><ul class="game-schedule"><li class="game-status"><li class="team-logo-small
logo-nfl-small nfl-small-5"><a href="http://espn.go.com/nfl/team/_/name/cle/cleveland-browns"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/cle/cleveland-browns">Cleveland</a></li></ul></td><td><ul class="game-schedule"><li class="game-
status win"><span class="greenfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=330915033">14-6</a></li></ul></td><td><a
href="/nfl/boxscore?gameId=330915033">Box Score</a> | <a href="/nfl/playbyplay?gameId=330915033">Play-By-Play</a></td></tr><tr class="odddrow
team-28-34"><td>3</td><td>10/27</td><td>Sun, Sep 22</td><td><td align="left"><ul class="game-schedule"><li class="game-status"><li class="team-logo-
small logo-nfl-small nfl-small-34"><a href="http://espn.go.com/nfl/team/_/name/hou/houston-texans"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/hou/houston-texans">Houston</a></li></ul></td><td><ul class="game-schedule"><li class="game-status
win"><span class="greenfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=330922033">30-9</a></li></ul></td><td><a
href="/nfl/boxscore?gameId=330922033">Box Score</a> | <a href="/nfl/playbyplay?gameId=330922033">Play-By-Play</a></td></tr><tr class="evenrow
team-28-2"><td>4</td><td>10/27</td><td>Sun, Sep 29</td><td><td align="left"><ul class="game-schedule"><li class="game-status"><li class="team-logo-small
logo-nfl-small nfl-small-2"><a href="http://espn.go.com/nfl/team/_/name/buf/buffalo-bills"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/buf/buffalo-bills">Buffalo</a></li></ul></td><td><ul class="game-schedule"><li class="game-status
loss"><span class="redfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=330929002">23-20</a></li></ul></td><td><a
href="/nfl/boxscore?gameId=330929002">Box Score</a> | <a href="/nfl/playbyplay?gameId=330929002">Play-By-Play</a></td></tr><tr class="odddrow
team-28-15"><td>5</td><td>10/27</td><td>Sun, Oct 6</td><td><td align="left"><ul class="game-schedule"><li class="game-status"><li class="team-logo-small
logo-nfl-small nfl-small-15"><a href="http://espn.go.com/nfl/team/_/name/mia/miami-dolphins"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/mia/miami-dolphins">Miami</a></li></ul></td><td><ul class="game-schedule"><li class="game-status
win"><span class="greenfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=331006015">26-23</a></li></ul></td><td><a
href="/nfl/boxscore?gameId=331006015">Box Score</a> | <a href="/nfl/playbyplay?gameId=331006015">Play-By-Play</a></td></tr><tr class="evenrow
team-28-9"><td>6</td><td>10/27</td><td>Sun, Oct 13</td><td><td align="left"><ul class="game-schedule"><li class="game-status"><li class="team-logo-small
logo-nfl-small nfl-small-9"><a href="http://espn.go.com/nfl/team/_/name/gb/green-bay-packers"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/gb/green-bay-packers">Green Bay</a></li></ul></td><td><ul class="game-schedule"><li class="game-status
loss"><span class="redfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=331013033">19-17</a></li></ul></td><td><a
href="/nfl/boxscore?gameId=331013033">Box Score</a> | <a href="/nfl/playbyplay?gameId=331013033">Play-By-Play</a></td></tr><tr class="odddrow
team-28-23"><td>7</td><td>10/27</td><td>Sun, Oct 20</td><td><td align="left"><ul class="game-schedule"><li class="game-status"><li class="team-logo-small
logo-nfl-small nfl-small-23"><a href="http://espn.go.com/nfl/team/_/name/pit/pittsburgh-steelers"></a></li><li class="team-name"><a
href="http://espn.go.com/nfl/team/_/name/pit/pittsburgh-steelers">Pittsburgh</a></li></ul></td><td><ul class="game-schedule"><li class="game-
status loss"><span class="redfont"><span></span></li><li class="score"><a href="/nfl/recap?gameId=331020023">19-16</a></li></ul></td><td><a
```

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](#)