

Parsování HTML představuje slangový výraz pro syntaktickou analýzu obsahu webové stránky. Lidově řečeno porcujeme zdrojový kód stránky a vyzobáváme potřebný obsah. Webovým vývojářům není neznámý pojem HTML DOM (Document Object Model). Ten je popsán konsorciem W3C (http://www.w3schools.com/jsref/dom_obj_document.asp) a umožňuje spravovat obsah stránky s pomocí javascriptu. Ačkoliv je možné brát inspiraci přímo z něj, v případě VBA se odvoláváme na starší knihovnu Microsoft HTML Object Library, v níž jsou některé vlastnosti definovány odlišně (outerHTML, innerText aj.). Každopádně výhodu mají ti, kteří se již potkali s vytvářením stránek v jazyce HTML a ovládají práci s tagy (elementy), jako je např. <body>, <div>, <p>, <table>, jejich atributy (vlastnostmi) a stylováním (CSS).

Pro účely testování jsem vytvořil stránku [parsovani.html](#).

Zdrojový kód:

```

1 <!DOCTYPE html>
2 <html>
3 <head>
4   <title>Moje stránka</title>
5   <meta charset="UTF-8"/>
6   <style>
7     table {
8       border-collapse: collapse;
9     }
10    table, td, th {
11      border: 1px solid #CCCCCC;
12    }
13    .moje_trida {
14      color: #0066CC;
15    }
16  </style>
17 </head>
18 <body>
19   <p><input name="muj_nazev" type="text" value="Excel 2010"/></p>
20   <p id="muj_identifikator" style="color: #FF0066">Element P s atributem
21   id="muj_identifikator".</p>
22   <div class="moje_trida">
23     <p>Element P v prvním elementu DIV s atributem class="moje_trida"
24     (index 0).</p>
25   </div>
26   <div class="moje_trida">
27     <p>Element P ve druhém elementu DIV s atributem class="moje_trida"
28     (index 1).</p>
29   </div>
30   <table>
31     <tr>
32       <td>Křížek</td>
33       <td>123,45</td>
34       <td>20.11.2015</td>
35     </tr>
36     <tr>
37       <td>Bydžovský</td>
38       <td>678,90</td>
39       <td>1.6.2016</td>
40     </tr>
41   </table>
42   <p><a href="http://www.ceskatelevize.cz/ct1/" target="_blank"></a><br />
44   <a href="http://www.ceskatelevize.cz/ct2/" target="_blank"></a></p>
46 </body>
47 </html>

```

A nyní už si pojdme obsah stránky rozebrat programově. Kód je takéž uveden v příloze a doporučuji jej krokovat a studovat v oknech Immediate a Locals.

```

1 Sub ParsovaniHTML()
2
3 'Tools / References / Microsoft HTML Object Library
4
5 Dim objMSHTML As New HTMLDocument
6 Dim objDocument As HTMLDocument
7
8 Dim objImages As IHTMLElementCollection
9 Dim objLinks As IHTMLElementCollection
10 Dim objElements As IHTMLElementCollection
11 Dim objTags As IHTMLElementCollection
12 Dim objTagsTagName As IHTMLElementCollection
13 Dim objTagsClass As IHTMLElementCollection
14 Dim objTagsName As IHTMLElementCollection
15
16 Dim objImage As IHTMLImageElement
17 Dim objLink As IHTMLAnchorElement
18
19 Dim objTagClass As IHTMLElement
20 Dim objTagName As IHTMLElement
21 Dim objTagId As IHTMLElement
22
23 'přesnější typy vyplývající z testování
24 Dim objTagClass As IHTMLDivElement
25 Dim objTagName As IHTMLInputElement
26 Dim objTagId As IHTMLParaElement
27
28 Dim objHTML As IHTMLHtmlElement
29 Dim objHead As IHTMLHeadElement
30 Dim objBody As IHTMLBodyElement
31
32 Dim strURL As String
33 Dim strTitulek As String
34 Dim strHTML As String
35 Dim strHead As String
36 Dim strBody As String
37
38 'adresa stránky
39 strURL = "http://proexcel.cz/test/parsovani.html"
40
41 'element ... tag
42
43 'dokument
44 Set objDocument = objMSHTML.createDocumentFromUrl(strURL, vbNullString)
45
46 'čekání na stažení
47 While objDocument.readyState <> "complete"
48     DoEvents
49 Wend
50
51 'titulek stránky
52 strTitulek = objDocument.title
53
54 'objekt HTML (element html)
55 Set objHTML = objDocument.documentElement
56 strHTML = objHTML.outerHTML
57
58 'hlavička (element head)
59 Set objHead = objDocument.head
60 strHead = objHead.outerHTML
61
62 'obsah stránky (element body)
63 Set objBody = objDocument.body
64 strBody = objBody.outerHTML
65
66 'kolekce obrázků (elementy img)
67 Set objImages = objDocument.images
68
69 For Each objImage In objImages
70     Debug.Print objImage.outerHTML
71     Debug.Print objImage.getAttribute("href")
72 Next
73
74 'kolekce hypertextových odkazů (elementy a)
75 Set objLinks = objDocument.links
76
77 For Each objLink In objLinks
78     Debug.Print objLink.outerHTML
79     Debug.Print objLink.innerHTML
80     Debug.Print objLink.getAttribute("href")
81 Next
82
83 'varianta 1 pro tagy
84
85 'kolekce elementů
86 Set objElements = objDocument.all
87
88 'kolekce elementů s požadovaným názvem (p)
89 Set objTags = objElements.tags("p")
90
91 'varianta 2
92
93 'kolekce elementů s požadovaným názvem (p)
94 Set objTagsTagName = objDocument.getElementsByTagName("p")
95
96 'element s atributem id (id="muj_identifikator")
97 'id by mělo být v dokumentu jedinečné
98 Set objTagId = objDocument.getElementById("muj_identifikator")
99
100 'typ nalezeného elementu
101 'p
102 strElement = objTagId.tagName
103
104 'získání barvy atributu style nalezeného elementu (style="color: ...")
105 '#f0066
106 strColor = objTagId.style.Color
107
108 'kolekce elementů s požadovaným atributem class (class="moje_trida")
109 Set objTagsClass = objDocument.getElementsByClassName("moje_trida")
110
111 For Each objTagClass In objTagsClass
112     Debug.Print objTagClass.tagName
113     Debug.Print objTagClass.outerHTML
114     Debug.Print objTagClass.innerHTML
115 Next
116
117 'kolekce elementů (zpravidla elementy input)
118 's požadovaným atributem name (name="hledaný_řetězec")
119 Set objTagsName = objDocument.getElementsByName("muj_nazev")
120
121 For Each objTagName In objTagsName
122     Debug.Print objTagName.tagName
123     Debug.Print objTagName.outerHTML
124     Debug.Print objTagName.getAttribute("value")
125 Next
126
127 'odstranění z paměti
128 Set objDocument = Nothing
129 Set objMSHTML = Nothing
130
131 End Sub

```

Řádky VBA jsem se snažil komentovat a na tomto místě jen upřesním pojmy innerHTML, innerText a outerHTML.

Příklad

```
<div><p>nějaký text</p></div>
```

`<div><p>nějaký text</p></div>` ... vlastnost outerHTML pro element `<div>`

`<p>nějaký text</p>` ... vlastnost innerHTML pro element `<div>`

`nějaký text` ... vlastnost innerTEXT pro element `<p>`

Pozn. Pokud se chcete odkazovat na členy kolekcí indexem, pak vězte, že číslování začíná nulou.

Přirozeně se sluší na tomto místě ukázat způsob, jak z dané stránky převzít tabulku do listu Excelu (ačkoliv prosté HTML tabulky je lepší načítat prostřednictvím karty Data / Z webu).

```

1  Sub ParsovaniTabulkyHTML()
2
3      Dim objMSHTML As New HTMLDocument
4      Dim objDocument As HTMLDocument
5
6      Dim objTagsRow As IHTMLElementCollection
7      Dim objTagsCell As IHTMLElementCollection
8
9      Dim objTagRow As HTMLTableRow
10     Dim objTagCell As HTMLTableCell
11
12     Dim strURL As String
13
14     'adresa stránky
15     strURL = "http://proexcel.cz/test/parsovani.html"
16
17     'dokument
18     Set objDocument = objMSHTML.createDocumentFromUrl(strURL, vbNullString)
19
20     'čekání na stažení
21     While objDocument.readyState <> "complete"
22         DoEvents
23     Wend
24
25     'všechny řádky tabulky
26     Set objTagsRow = objDocument.getElementsByTagName("tr")
27
28     'pro každý řádek
29     For Each objTagRow In objTagsRow
30
31         'všechny buňky řádku
32         Set objTagsCell = objTagRow.getElementsByTagName("td")
33
34         'čítač pro řádky
35         i = i + 1
36
37         For Each objTagCell In objTagsCell
38
39             'čítač pro sloupce
40             j = j + 1
41
42             'zápis do buněk listu
43             Select Case j
44                 Case 1
45                     'text
46                     Cells(i, j).Value = objTagCell.innerText
47                 Case 2
48                     'desetinné číslo
49                     Cells(i, j).Value = CDbI(objTagCell.innerText)
50                 Case 3
51                     'datum
52                     Cells(i, j).Value = CDate(objTagCell.innerText)
53             End Select
54
55             Next objTagCell
56
57             'reset čítače pro sloupce
58             j = 0
59
60         Next objTagRow
61
62     'odstranění z paměti
63     Set objDocument = Nothing
64     Set objMSHTML = Nothing
65
66 End Sub

```



Tabulka převedená z HTML stránky

HTML stránky by do jisté míry měly dodržovat hierarchii objektů a jejich vnořování do sebe. V praxi tomu tak často není a jejich obsah bývá uspořádán laxně, na rozdíl třeba od XML. Je to jeden z důvodů, proč i já jsem v daném tématu nevyužil skutečnosti, že tagy (elementy) představují jakési „nody“ ve stromové struktuře, kdy uvažujeme vazby rodič (parent) – dítě (child), případně děti (children).

Parsování HTML stránek nepatří k technikám, za které bychom se mohli plácet po ramenou. Pokud máme možnost, vždy sáhneme po přímém přístupu k datům do databáze. Klíčové je slovíčko „pokud“. Až příliš dobře se pamatuji na nutnost zpracovat 60 000 webových stránek z nejmenovaného webu státní správy jen proto, že webová aplikace padala pod deseti minutách nastavování parametrů (bez možnosti uložení). Poměrně solidně se s HTML kódem umí vypořádat i regulární výrazy.

Ke zpracování webových stránek a jejich obsahu se opět někdy vrátíme.

Příloha

[excel_parsovani_html.zip](#)