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>, , <p>, <table>, jejich atributy (vlastnostmi) a stylováním (CSS).

Pro účely testování jsem vytvořil stránku parsovani.html.

Zdrojový kód:

```
< !DOCTYPE html>
<html>
<head>
   <title>Moje stránka</title>
   <meta charset="UTF-8"/>
   <style>
   table {
         border-collapse: collapse;
      }
      table, td, th {
         border: 1px solid #CCCCCC;
      }
      .moje_trida {
         color: #0066CC;
      }
   </style>
</head>
<body>
   <p><input name="muj_nazev" type="text" value="Excel 2010"/></p>
   <p id="muj_identifikator" style="color: #FF0066">Element P s atributem
   id="muj_identifikator".</p>
   <div class="moje_trida">
      <p>Element P v prvním elementu DIV s atributem class="moje_trida"
      (index 0).</p>
   </div>
   <div class="moje_trida">
      <p>Element P ve druhém elementu DIV s atributem class="moje_trida"
      (index 1).</p>
   </div>
   <table>
      <tr>
         <td>Křížek</td>
         <td>123,45</td>
         <td>20.11.2015</td>
      </tr>
      <tr>
         <td>Bydžovský</td>
         <td>678,90</td>
         <td>1.6.2016</td>
      </tr>
   </table>
   <p><a href="http://www.ceskatelevize.cz/ct1/" target="_blank"><img alt=
  "Logo ČT1" height="83" src="ct1.jpg" width="155"/></a><br />
   <a href="http://www.ceskatelevize.cz/ct2/" target="_blank"><img alt=
  "Logo ČT2" height="83" src="ct2.jpg" width="155"/></a></p>
</body>
</html>
```

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

```vba
Sub ParsovaniHTML()

    'Tools / References / Microsoft HTML Object Library

    Dim objMSHTML As New HTMLDocument
    Dim objDocument As HTMLDocument

    Dim objImages As IHTMLElementCollection
    Dim objLinks As IHTMLElementCollection
    Dim objElements As IHTMLElementCollection
    Dim objTags As IHTMLElementCollection
    Dim objTagsTagName As IHTMLElementCollection
    Dim objTagsClass As IHTMLElementCollection
    Dim objTagsName As IHTMLElementCollection

    Dim objImage As IHTMLImgElement
    Dim objLink As IHTMLAnchorElement

    Dim objTagClass As IHTMLElement
    Dim objTagName As IHTMLElement
    Dim objTagId As IHTMLElement

    'přesnější typy vyplývající z testování
    'Dim objTagClass As IHTMLDivElement
    'Dim objTagName As IHTMLInputElement
    'Dim objTagId As IHTMLParaElement

    Dim objHTML As IHTMLHtmlElement
    Dim objHead As IHTMLHeadElement
    Dim objBody As IHTMLBodyElement

    Dim strURL As String
    Dim strTitulek As String
    Dim strHTML As String
    Dim strHead As String
    Dim strBody As String

    'adresa stránky
    strURL = "http://proexcel.cz/test/parsovani.html"

    'element ... tag

    'dokument
    Set objDocument = objMSHTML.createDocumentFromUrl(strURL, vbNullString)

    'čekání na stažení
    While objDocument.readyState <> "complete"
        DoEvents
    Wend

    'titulek stránky
    strTitulek = objDocument.title

    'objekt HTML (element html)
    Set objHTML = objDocument.documentElement
    strHTML = objHTML.outerHTML

    'hlavička (element head)
    Set objHead = objDocument.head
    strHead = objHead.outerHTML

    'obsah stránky (element body)
    Set objBody = objDocument.body
    strBody = objBody.outerHTML

    'kolekce obrázků (elementy img)
    Set objImages = objDocument.images

    For Each objImage In objImages
        Debug.Print objImage.outerHTML
        Debug.Print objImage.getAttribute("href")
    Next

    'kolekce hypertextových odkazů (elementy a)
    Set objLinks = objDocument.links

    For Each objLink In objLinks
        Debug.Print objLink.outerHTML
        Debug.Print objLink.innerHTML
        Debug.Print objLink.getAttribute("href")
    Next

    'varianta 1 pro tagy

    'kolekce elementů
    Set objElements = objDocument.all

    'kolekce elementů s požadovaným názvem (p)
    Set objTags = objElements.tags("p")

    'varianta 2

    'kolekce elementů s požadovaným názvem (p)
    Set objTagsTagName = objDocument.getElementsByTagName("p")

    'element s atributem id (id="muj_identifikator")
    'ID by mělo být v dokumentu jedinečné
    Set objTagId = objDocument.getElementById("muj_identifikator")

    'typ nalezeného elementu
    'P
    strElement = objTagId.tagName

    'získání barvy atributu style nalezeného elementu (style="color: ...")
    '#ff0066
    strColor = objTagId.style.Color

    'kolekce elementů s požadovaným atributem class (class="moje_trida")
    Set objTagsClass = objDocument.getElementsByClassName("moje_trida")

    For Each objTagClass In objTagsClass
        Debug.Print objTagClass.tagName
        Debug.Print objTagClass.outerHTML
        Debug.Print objTagClass.innerHTML
    Next

    'kolekce elementů (zpravidla elementy input)
    's požadovaným atributem name (name="hledaný řetězec")
    Set objTagsName = objDocument.getElementsByName("muj_nazev")

    For Each objTagName In objTagsName
        Debug.Print objTagName.tagName
        Debug.Print objTagName.outerHTML
        Debug.Print objTagName.getAttribute("value")
    Next

    'odstranění z paměti
    Set objDocument = Nothing
    Set objMSHTML = Nothing

End Sub
```

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

Příklad
nějaký text

nějaký text … vlastnost outerHTML pro element
nějaký text … vlastnost innerHTML pro element
nějaký text … vlastnost innerTEXT pro element

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).

```vba
Sub ParsovaniTabulkyHTML()

    Dim objMSHTML As New HTMLDocument
    Dim objDocument As HTMLDocument

    Dim objTagsRow As IHTMLElementCollection
    Dim objTagsCell As IHTMLElementCollection

    Dim objTagRow As IHTMLTableRow
    Dim objTagCell As IHTMLTableCell

    Dim strURL As String

    'adresa stránky
    strURL = "http://proexcel.cz/test/parsovani.html"

    'dokument
    Set objDocument = objMSHTML.createDocumentFromUrl(strURL, vbNullString)

    'čekání na stažení
    While objDocument.readyState <> "complete"
        DoEvents
    Wend

    'všechny řádky tabulky
    Set objTagsRow = objDocument.getElementsByTagName("tr")

    'pro každý řádek
    For Each objTagRow In objTagsRow

        'všechny buňky řádku
        Set objTagsCell = objTagRow.getElementsByTagName("td")

        'čítač pro řádky
        i = i + 1

        For Each objTagCell In objTagsCell

            'čítač pro sloupce
            j = j + 1

            'zápis do buněk listu
            Select Case j
                Case 1
                    'text
                    Cells(i, j).Value = objTagCell.innerText
                Case 2
                    'desetinné číslo
                    Cells(i, j).Value = CDbl(objTagCell.innerText)
                Case 3
                    'datum
                    Cells(i, j).Value = CDate(objTagCell.innerText)
            End Select

        Next objTagCell

        'reset čítače pro sloupce
        j = 0

    Next objTagRow

    'odstranění z paměti
    Set objDocument = Nothing
    Set objMSHTML = Nothing

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 plácat 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