

Re: Recursively scraping web pages for embedded links and files

Source:

<http://www.tech-archive.net/Archive/Excel/microsoft.public.excel.programming/2008-02/msg02345.html>

- *From:* "Ker_01" <ker_01@xxxxxxxxxxxxxxxxxxxx>
 - *Date:* Fri, 15 Feb 2008 16:56:37 -0400
-

I've gotten a little farther– the following with the regex works, but it is only debug.printing the first match in the page. There are more target matches in the page, so either my regex is incorrect, or how I am collecting the matches is incorrect. Any advice or corrections welcome!

Keith

Sub GrabPage()

```
Set objHTTP = CreateObject("MSXML2.ServerXMLHTTP")
URL = "http://ourdomain/targetpage"
objHTTP.Open "GET", URL, False
objHTTP.setRequestHeader "User-Agent", "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)"
objHTTP.send ("")
```

```
SourceHTMLText = objHTTP.responseText
```

```
Dim re As RegExp
Set re = New RegExp
Dim s As String
re.Pattern = "^<A HREF=.*>"
re.Global = False
re.IgnoreCase = True
re.MultiLine = True
s = SourceHTMLText
Dim matches As MatchCollection
Set matches = re.Execute(s)
Dim mcmatch As Match
For Each mcmatch In matches
Debug.Print mcmatch.Value
Next
```

End Sub

"Ker_01" <ker_01@xxxxxxxxxxxxxxxxxxxx> wrote in message news:u89JvKBcIHA.4888@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Re: Recursively scraping web pages for embedded links and files

Tim– thank you again for your response. I may send some specific questions via email, but right now my questions are general enough that responses in the public record may help others.

I've managed to get the HTML of a target page (the top page of interest) using the code at the bottom of this post. I have two questions:

(1) The returned object (SourceHTMLText)– what exactly is it? I tried setting Sheet1.range("A1").text equal to it and got an error: either it isn't text, or maybe there is some limit where I can't assign a string over a certain size to a cell? I pushed it to a msgbox, and it did show on the screen (although truncated, because the HTML is long for a messagebox).

(2) What is the best way to parse the content for each instance that starts with ""? In the past, I've read flat files but I could read a line at a time and look for what I wanted. Now I have the whole text at once, so I have to handle it all together. Also, the text may span more than one line, which wouldn't work with the way I used to look for short strings anyway. Should I be using Regex? I saw some "simple" tutorials that tried to cover every aspect of regex, but since they cover a lot more than I need to know, I found it difficult to extract how to look for strings that have a specific start sequence (""). Pointers to any tutorials that are relatively easy to understand would be very helpful.

(2b) In case any referred tutorials don't include the info– if I do use Regex, and I do expect multiple matches within the document, what is best practice for storing those matches and using them– does Regex automatically build an array of matches?

Many thanks,

Keith
XL2003

Sub GrabPage()

```
Set objHTTP = CreateObject("MSXML2.ServerXMLHTTP")  
URL = "http://ourdomain/targetpage  
objHTTP.Open "GET", URL, False  
objHTTP.setRequestHeader "User-Agent", "Mozilla/4.0 (compatible;  
MSIE 6.0; Windows NT 5.0)"  
objHTTP.send ("")
```

SourceHTMLText = objHTTP.responseText

End Sub

Re: Recursively scraping web pages for embedded links and files

"Tim Williams" <timjwilliams at gmail dot com> wrote in message news:u55Lug4bIHA.4684@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Starting from the main page you could identify all of the "folder" links by looking at the URL: each could be clicked in order to drill down into subfolders, and each of these listed etc etc.

Grabbing URL's to the files will be more difficult: you'll have to deconstruct the "openDocument()" javascript code to see how it determines what URL to open. You can't use the javascript href directly in Excel: it depends on having the js function available.

If you're new to working with HTML docs from Excel then it may be a long haul. I can help you with specific points but can't provide a solution. If you prefer you can follow up via email (tim j williams at gmail dot com: no spaces, etc.).

Tim

"Ker 01" <ker_01@xxxxxxxxxxxxxxxxxxxx> wrote in message news:OnBOfG1bIHA.5164@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

This is a followup to a post from yesterday (Thanks to Tim Williams for responding). I have more information now, and felt it warranted a second try to see if there is way to do this now that we've gotten the documents exposed via the web interface. Using XL2003 on WinXP.

We have a corporate web application that exposes various documents in multiple levels of subdirectories. My belief is that these are stored in a database, but now they are directly accessible via web links through this web application, so where they come from hopefully doesn't affect what I am trying to accomplish.

Starting from the main page of the web application, I need to scrape the entire directory tree and capture some of the details (javascript links to .doc and .pdf files that can be opened through IE6 via 'dedicated' URLs for each document). I'm sure I'll have more questions once I start dissecting the HTML, but for starters I need to understand how to even

Re: Recursively scraping web pages for embedded links and files

scrape multiple levels within the directory tree of a website.
I've
copied in some of the URLS (changed slightly for corporate
security) to
give a sense of what I'm working with.

Top of tree:

[http://ourserver.com/rtsa-bin/PermaSite.dll/aaumain.htm?site=omatcone&pagetitle=M%20S%](http://ourserver.com/rtsa-bin/PermaSite.dll/aaumain.htm?site=omatcone&pagetitle=M%20S%20)

I can click a link to go to the next level of subfolder:

<http://ourserver.com/rtsa-bin/PermaSite.dll/aaudisplayfolder.htm?folderpathID=0b00043d800>

Third level of folder:

<http://ourserver.com/rtsa-bin/PermaSite.dll/aaudisplayfolder.htm?folderpathID=0b00043d800>

and so on.

A sample link for a single document within one of the pages
in the web

tree/directory is:

[javascript:openDocument\('0900043d802b3528'\)](javascript:openDocument('0900043d802b3528')):

where clicking that link ultimately opens:

<http://ourserver.com/Documentation/03451TRs142.pdf>

Ultimately I need to recreate all the links in an Excel
workbook so

users can click on a hyperlink and access the relevant
document. An

Excel hyperlink that uses the javascript:opendocument
command is totally

fine with me, but first I need to collect them all. Alternatively

I'll

have to figure out how to cycle through each javascript
command anyway.

then identify the URL it opened (which sounds harder).

Any advice or code snippets greatly appreciated- I haven't
done anything

with HTML at all.

Thanks.

Keith

Re: Recursively scraping web pages for embedded links and files

.