Subject: Re: Download files from the web
Posted by Helder Marchetto on Fri, 10 Jan 2014 11:08:00 GMT
View Forum Message <> Reply to Message

On Friday, January 10, 2014 11:41:20 AM UTC+1, Helder wrote:
> On Friday, January 10, 2014 11:22:16 AM UTC+1, Mats Löfdahl wrote:
>
>> Den tisdagen den 17:e december 2013 kl. 10:39:42 UTC+1 skrev Mats Löfdahl:
>
>>
>
>>> Den måndagen den 16:e december 2013 kl. 22:27:02 UTC+1 skrev Mike Galloy:
>
>>
>
>>>
>
>>
>
>>>> On 12/16/13, 7:46 AM, Mats Lï¿½fdahl wrote:
>
>>
>
>>>
>
>>
>
>>>> > Den mï¿½ndagen den 16:e december 2013 kl. 15:14:10 UTC+1 skrev Mats
Lï¿½fdahl:
>
>>
>
>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> Thanks. But it seems it has the same problem as the webget
>
>>
>

>>>
>
>>
>
>>>> >> function, in that it can't tell the difference between a proper
>
>>
>
>>>
>
>>
>
>>>> >> download and a 404 error web page.
>
>>
>
>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> I simplified your code a bit (because I don't need the progress
>
>>
>
>>>
>
>>
>
>>>> >> bar) and came up with this:
>
>>
>
>>>
>
>>
>
>>>> >>
>
>>
>

```
>>>
>
>>
>
>>>>  >> function downloadurl, url, file
>
>>
>
>>>
>
>>
>
>>>>  >>
>
>>
>
>>>
>
>>
>
>>>>  >> url_scheme = (strsplit(url, ':',/extract))[0]
>
>>
>
>>>
>
>>
>
>>>>  >>
>
>>
>
>>>
>
>>
>
>>>>  >> url_hostname = strjoin((strsplit(url,'/',/extract))[1:*],'/')
>
>>
>
>>>
>
>>
>
>>>>  >>
>
>>
>
```

```
>>>
>
>>
>
>>>>  >> oUrl = OBJ_NEW('IDLnetUrl', URL_SCHEME = url_scheme, URL_HOSTNAME =
>
>>
>
>>>
>
>>
>
>>>>  >> url_hostname)
>
>>
>
>>>
>
>>
>
>>>>  >>
>
>>
>
>>>
>
>>
>
>>>>  >> retrievedFilePath = oUrl->Get(FILENAME=file)
>
>>
>
>>>
>
>>
>
>>>>  >>
>
>>
>
>>>
>
>>
>
>>>>  >> oUrl->GetProperty, RESPONSE_CODE=RespCode ; 200 = OK
>
>>
>
```

```
>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> oUrl->CloseConnections
>
>>
>
>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> OBJ_DESTROY, oUrl
>
>>
>
>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> return, RespCode eq 200       ; True if OK
>
>>
>
```

>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> end
>
>>
>
>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> I tried it both with a url pointing to an existing web page and to
>
>>
>
>>>
>
>>
>
>>>> >> a non-existing page. In both cases I get RespCode eq 200. With the
>
>>
>
>>>
>
>>
>
>>>> >> non-existing page I again had downloaded a 404 error page.
>
>>
>

>>>
>
>>
>
>>>> >>
>
>>
>
>>>
>
>>
>
>>>> >> I got the value 200 for OK from the list at
>
>>
>
>>>
>
>>
>
>>>> >> http://www.exelisvis.com/docs/IDLnetURL.html#objects_network _1009015_1417867
>
>>
>
>>>
>
>>
>
>>>>
>
>>
>
>>>
>
>>
>
>>>> > Thought it might work better to use spawn and wget and read its exit
>
>>
>
>>>
>
>>
>
>>>> > status. But that seems to have the same problem: 404 error page
>
>>
>

>>>
>
>>
>
>>>> > downloaded in case the remote file doesn't exist, but exit status 0
>
>>
>
>>>
>
>>
>
>>>> > (=OK) regardless.
>
>>
>
>>>
>
>>
>
>>>>
>
>>
>
>>>
>
>>
>
>>>> > So this does not seem to be an IDL problem. It is just hard to get
>
>>
>
>>>
>
>>
>
>>>> > the information I want from the download process.
>
>>
>
>>>
>
>>
>
>>>>
>
>>
>

>>>
>
>>
>
>>>> > The web server obviously knows the requested file does not exist but
>
>>
>
>>>
>
>>
>
>>>> > isn't there a way to make it tell the downloading process this in a
>
>>
>
>>>
>
>>
>
>>>> > more condensed way than constructing a web page with a 404 error?
>
>>
>
>>>
>
>>
>
>>>>
>
>>
>
>>>
>
>>
>
>>>> I use IDLnetURL in my routine and it is able to tell if there is a 404
>
>>
>
>>>
>
>>
>
>>>> error:
>
>>
>

```
>>>
>
>>
>
>>>>
>
>>
>
>>>
>
>>
>
>>>>  IDL> c = mg_get_url_content('http://michaelgalloy.com/nothing', $
>
>>
>
>>>
>
>>
>
>>>>  IDL>                    error_message=em, $
>
>>
>
>>>
>
>>
>
>>>>  IDL>                    response_code=rc, response_header=rh)
>
>>
>
>>>
>
>>
>
>>>>  IDL> help, em
>
>>
>
>>>
>
>>
>
>>>>  EM              STRING    = 'IDLNETURL::GET:  CCurlException:  Error:
>
>>
>
```

>>>
>
>>
>
>>>> Http Get Request Fai'...
>
>>
>
>>>
>
>>
>
>>>> IDL> help, rc
>
>>
>
>>>
>
>>
>
>>>> RC   LONG   =   404
>
>>
>
>>>
>
>>
>
>>>
>
>>
>
>>>
>
>>
>
>>> Aha. This is an important clue! It seems to be a property of the web server and not of the way we try to download. With your url, http://michaelgalloy.com/nothing, I also get response code = 404 with the code I wrote but with, e.g, http://www.exelisvis.com/docs/nothing, I get 200.
>
>>
>
>>>
>
>>
>
>>>
>

>>
>
>>>
>
>>
>
>>> I guess I should really try it on the server I will be downloading from. Problem is just that it is down right now and it is remote enough that nobody is there to turn it on until Christmas. I also have some influence over that server so I should be able to request that it is set up so that it returns the proper code in case the file does not exist.
>
>>
>
>>>
>
>>
>
>>>
>
>>
>
>>>
>
>>
>
>>> Hm. But how to test what happens when the file exists on the server but does not get downloaded properly? Would be great if one could get the file size or even better a checksum from the server and then compare that with the same property of the downloaded file. But I don't see any of those properties listed here http://www.exelisvis.com/docs/IDLnetURL_Properties.html.
>
>>
>
>>
>
>>
>
>> I thought I had this working but now I suddenly get into another kind of problem when the file I'm trying to download does not exist on the server. The get method of IDLnetUrl seems to crash rather than just returning a status that is NE 200.
>
>>
>
>>
>
>>
>
>> What I do now is:
>

```
>>
>
>>
>
>>
>
>>
>
>>
>   urlComponents = parse_url(url)
>
>>
>
>>   oUrl = OBJ_NEW('IDLnetUrl' $
>
>>
>
>>                  , URL_SCHEME = urlComponents.scheme $
>
>>
>
>>                  , URL_HOSTNAME = urlComponents.host $
>
>>
>
>>                  , URL_PATH = urlComponents.path $
>
>>
>
>>                  , URL_PORT = urlComponents.port $
>
>>
>
>>                 )
>
>>
>
>>   tmpfile = String('tmp_', Bin_Date(SysTime()), format='(A, I4, 5I2.2)')
>
>>
>
>>   retrievedFilePath = oUrl -> Get(FILENAME=tmpfile)
>
>>
>
>>   oUrl -> GetProperty, RESPONSE_CODE=RespCode
>
```

>>
>
>>
>
>>
>
>> The urlComponents look fine, tmpfile is a string as it should (e.g., 'tmp_20140110l1025') but I get the following error message:
>
>>
>
>>
>
>>
>
>> % IDLNETURL::GET: CCurlException: Error: Http Get Request Failed. Error =
>
>>
>
>> http: Client Error. Remote Host(www.royac.iac.es), Http
>
>>
>
>> ErrCode(404), Http Err(Not Found) Http ErrMsg(No HTML found).
>
>>
>
>> % Execution halted at: RED_GETURL 159
>
>>
>
>>
>
>>
>
>> where red_geturl is the function where I put the code above (with line 159 being the one where the get method is called). retrievedFilePath is undefined after this of course.
>
>>
>
>>
>
>>
>
>> I can execute the last line above after this and get a response code of 404 but at that point the program has stopped. I suppose I could add some error handling code but isn't the IDLnetUrl object supposed to take care of this?
>

>
>
> Hi Mats,
>
> this will not really help, but at least you feel my pain :-)
>
> I'm working around this problem with a call to catch and then check the error code (404). You can then either report that the file is missing if the error is 404, otherwise simply give out the error code.
>
>
>
> I found that this is the only way if you're using http. With ftp you can of course check the list of available files with the method GetFtpDirList.
>
>
>
> But I guess you already know all of this.
>
>
>
> If there's a better way... happy to hear about it.
>
>
>
> Regards,
>
> Helder

Hi Mats,
one more dirty trick. It is a one liner that uses windows powershell. Dunno about Linux, but you also have on linux the wGet command and it looks like it uses the same syntax. Otherwise curl should also work (--head option).

SPAWN, 'powershell -WindowStyle Hidden "wget --server-response --spider -o OutputFile.txt http://yourserver.com/MissingFile.txt"', wGetResult, /NOSHELL

wGetResult will unfortunately be empty.
The resulting file (OutputFile.txt) will contain a lot of stuff, among which lots of 404 mentions if the file does not exist. You can then search for that or check if the last line of the file contains "Remote file exists."

Does this help?

I'm always on the look for better/cleaner solutions.

Regards,
Helder