

The *Advantage* Consultancy HTML Processor

This program was written to meet an in-house requirement for a program to:

Process HTML files to remove HTML tags.

Process HTML files to produce a list of http and ftp URLs within that document.

The program is not terribly complex and I've released it as Shareware for other people to use as well as ourselves. If you do find the program useful, then a payment of \$20 or £10 (Sterling) would be appreciated. In return, you will receive upgrades of the software. The next upgrade is likely to include improvements to error trapping, the ability to URLs other than ftp and http, the ability to process e-mails and news messages and an improved handling of line lengths. Also, I reckon that I'll probably make the system use a 'real' .INI file, proper Help and offer more flexibility in the naming of output files, possibly including an 'output file directory' setting. I'll even send you the source!

If you wish to do the honourable thing, then please send your payment to The Advantage Consultancy, Sheffield Science Park, Arundel Street, Sheffield, S1 2NS, United Kingdom. If you send a cheque, make it payable to 'The Advantage Consultancy', though Sterling Cheques or Money Orders are preferable - the banks take a bit of a cut off of dollar cheques! I'd also be generally interested in what you think of this program, whether it's any use, what other improvements could be made, etc. If you are a registered charity then please register for upgrades on a letterhead but don't send payment.

Feel free to distribute this package but distribute all files. The copyright of this software remains with The Advantage Consultancy; it is not being put in the Public Domain.

Use and Installation of the program

The program will process HTML files to produce two files. A .URL file with the same name as the source file, which contains a list of any http or ftp URLs found in the source file, and a .ASC file with the same name as the source file that contains the HTML.

Installation

To set the program up, after unzipping the HTMLASCI.ZIP file, copy all the files into a directory called C:\HTMLASC. Now MOVE the .VBX and .DLL files (if you downloaded the version with the VBRUN300.DLL file) into your WINDOWS\SYSTEM directory.

Set up a group file from Windows using the NEW, Program Group options from the File menu, and then set up a program item for the HTMLASCI.EXE.

Running the Program

On running the program, you may, if you have not installed a full file set, receive a message requesting you to recreate the System Options file. If this occurs, it indicates that the file HTMLASC.DAT, which contains option information for the system such as whether or not to build an out URL List (see later) has been damaged or is not present. Enter the Options tab and rebuild the file.

Setting the Options Up

Click on the 'Options' Tab and you will see a form containing the following options. Fill them in, then Save. This will create the options file and also set the values for the options for this and

future sessions.

ASCII Output File Extension

This is the three letter extension added to ASCII files created by the system. Enter the three letter suffix and a '.'. So, if you want files created to have the suffix 'asc' enter '.asc' in this box. The program will then convert 'index.htm' to 'index.asc' and so on.

Note that the program does not warn if a file already exists; take care with naming these extensions.

URL List Output File Extension

This is the three letter extension added to URL List files created by the system. Enter the three letter suffix and a '.'. So, if you want files created to have the suffix 'url' enter '.url' in this box. The program will then produce a URL list from 'index.htm' called 'index.url' and so on.

Note that the program does not warn if a file already exists; take care with naming these extensions.

Create URL List Output File

Check this box to create a list of URLs found in a particular HTML file. The program will output to a file any line containing a URL of the form 'http' or 'ftp'. This can be useful for summarising lists of URL information.

When checked, processing a single HTML file will create both an ASCII file and a file of URLs.

Include Source Document Name

When checked includes the file name of the source document in the URL List output file.

Processing Files

Processing Files is very simple. Use the Drive and Directory boxes to select the directory of interest. All files in that directory with the suffix .htm will be listed in the file list box. Simply select the files that you want to process by clicking the file name with the mouse. If you wish to process more than one file at once, hold the control key down and click on each file that you wish to process.

The files marked for processing will be displayed in reverse video on screen.

Press the 'Process' button. The status line at the bottom of the screen will display the name of the file currently being processed.

Output files will be created as follows:

<filename.htm> will be processed to create:

<filename.asc> containing an ASCII representation of the file.

<filename.url> containing a list of ftp and http URLs in that file.

assuming the default file suffixes are used.

Thanks

Joe Pritchard, BSc(Hons), FIAP, MBA
joe@vantage.demon.co.uk
<http://www.demon.co.uk/advantage>