ScanHTML

Serge Emond

COLLABORATORS			
	TITLE : ScanHTML		
ACTION	*NAME*	*DATE*	*SIGNATURE*
WRITTEN BY	Serge Emond	February 12, 2023	

REVISION HISTORY			
NUMBER	DATE	DESCRIPTION	NAME

Contents

Chapter 1

ScanHTML

1.1 ScanHTML.guide/Contents

ScanHTML

Version 1.04

1.2 ScanHTML.guide/About English

A word or two about English

My primary language is French. I read English very well but I'm sure what I write is awful. Sorry about that.

I could write this in French but maybe you could not read it!

When I will have some free time I will write the docs in French too but that's not now!

1.3 ScanHTML.guide/Introduction

Introduction

I'm writing ScanHTML because I want to build an HTTP robot.

An HTTP robot is a program that can automatically download one or more files using HTTP. Then it can scan those files to search for references to other files that it download upon some conditions and so on...

ScanHTML scans HTML files and creates a list of urls. That's all!

1.4 ScanHTML.guide/Features

Features

- Written in C, which means QUICK compared with ARexx scripts

- Can append to or overwrite an existing url list

- Can complete relative URLs (".", "..", & "/" are understood)

- Accepts two independant AmigaDOS patterns to select which URLs to keep

- Select which type(s) of URLs to keep (backgrounds, images, references)

Not a very long list but, there IS a list! :)

1.5 ScanHTML.guide/Legal information

Legal information

License

No warranty

1.6 ScanHTML.guide/License

License

ScanHTML is released under the concept of freeware. This means you are allowed to use and copy this program freely, as long as the following requirements are fulfilled:

All files are copied without any alterations or modifications. If any

extra files are added, it must be obvious that they do not belong to

the original distribution, and that they do not need to be included in

any redistribution. Exception: So called "BBS ads" may not be added.

The copying is done on a non-commercial basis. A small fee to cover

media costs etc. may be charged.

The copier is not claiming the copyright of this program.

Any exceptions from the above require a written permission from the author.

If you want to publish this program on a cover disk or similar, contact me first for approval (to make sure you have the latest version etc). I then expect a copy of the issue in question in return (additional contributions are welcomed :).

1.7 ScanHTML.guide/Legal issues/No warranty

No warranty

There is no warranty for the programs, to the extent permitted by applicable law. Except when otherwise stated in writing the copyright holder and/or other parties provide the programs "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the programs is with you. Should the programs prove defective, you assume the cost of all necessary servicing, repair or correction.

In no event unless required by applicable law or agreed to in writing will any copyright holder, or any other party who may redistribute the programs as permitted above, be liable to you for damages, including any general, special, incidental or consequential damages arising out of the use or inability to use the programs (including but not limited to loss of data or data being rendered inaccurate or losses sustained by you or third parties or a failure of the programs to operate with any other programs), even if such holder or other party has been advised of the possibility of such damages.

1.8 ScanHTML.guide/Requirements

Requirements

ScanHTML requires at least OS v37 (2.0 release). You must know however that if used on a pre-v39 (3.0) machine, there is a bug in dos.library: character-classes ([x-y]) are not converted properly to uppercase. ScanHTML won't uppercase them either. I suppose it will work OK if you use uppercase [x-y] patterns. (See you AmigaDOS manual or dos.lib's autodocs about patterns)

The only other thing required is an HTML file. :)

1.9 ScanHTML.guide/History

The Past

Version 0 Revision 1 - 16.8.1996 and before

Nothing recorded.

Version 0 Revision 2 - 21.8.1996

· Changed NOIMG to NOSRC.

· Now recognize tags not in uppercase

Version 0 Revision 3- 15.12.1996

· Now works when there's no '"'.

14.1.1997 - Since it seems stable, 0.3 is relabelled 1.0. Also renamed ScanHTML (was ParseHTML).

Version 1.01 - 17.1.1997

· Fixed enforcer hit

Version 1.02 - 22.2.1997

· Replaced "NO#" argument with "STRIP#": The URL is truncated at the "#" instead of discarded.

Version 1.03 - 13.4.1997

· Now handle situations where an url is "." or "..". In past, '/' was required (ie "./" & "../").

Version 1.04 - 15.4.1997

· Now complains about ".." followed by something else than nothing or than "/" but do not abort scanning. This is a quick and dirty patch but I don't have the time to do more :)

1.10 ScanHTML.guide/What's next

What's Next - The To Do List

· Nothing for now.

1.11 ScanHTML.guide/Bugs

Known Bugs

· Patterns, see Requirements .

1.12 ScanHTML.guide/Credits

Credits

ScanHTML was entirely written in 'C' by Serge Emond .

1.13 ScanHTML.guide/Author

About the author

Name: Emond

First Name: Serge

Current email: emonds@jsp.umontreal.ca

Current web: http://www.jsp.umontreal.ca/~emonds/

1.14 ScanHTML.guide/This is BETA

This is a BETA release

This release is BETA. This means it has not been fully tested. I release it only because it seems stable and it works well on my computer. As stated in legal issues I may not be responsible for any dammage caused by the use of this program.

1.15 ScanHTML.guide/Arguments

Arguments

ScanHTML can only be used from a Shell. It is pure and ca be made resident.

You can stop it whatever it is doing by sending it CTRL-C.

Warning: ScanHTML needs stack.. if it crashs, increase it. 12k seems enough, 16k and more is safe...

ReadArgs() Template:

I=IN/A, O=OUT/A, A=APPEND/S, B=BASE/K, P=PATTERN/K, P2=PATTERN2/K, NOBG/S, NOF#=NOFIRST#/S, NOHREF/S, NOSRC/S, NQ=NOQUERY/S, STRIP#/S

Text arguments

IN Input HTML file.

OUT Output URL list.

BASE URL to use when completing URLs.

PATTERN Pattern saved URLs must match.

PATTERN2 Another pattern that must be matched.

Switches

APPEND Don't overwrite output file.

NOBG Don't add background images.

NOFIRST# Don't add urls beginning with '#'.

NOHREF Don't add HRef'd urls.

NOSRC Don't add SRC urls.

NOQUERY Don't add urls containing '?'.

STRIP# Truncate URLs at "#".

1.16 ScanHTML.guide/Arguments/IN

I=IN/A

Required. Name of the HTML file to scan.

If you are running OS v40+ the input will be buffered (4k).

1.17 ScanHTML.guide/Arguments/OUT

O=OUT/A

Required. Name of the file to save urls to.

If you are running OS v40+ the output will be buffered (16k).

1.18 ScanHTML.guide/Arguments/APPEND

A=APPEND/S

Tells ScanHTML to append to an existing file instead of overwriting it.

If the file do not exists then it is created.

1.19 ScanHTML.guide/Arguments/BASE

B=BASE/K

If not given, the urls in the output file will be identical to those in the HTML file.

If given, this represents the name of the file ScanHTML is actually scanning. In that case, the incomplete urls will be completed using this name. For example, using "BASE http://www.z.com/dizz/index.html" would do the following conversions on urls not beginning with "HTTP://":

buz.gif -> http://www.z.com/dizz/buz.gif

../imgs/moon.png -> http://www.z.com/imgs/moon.png

Actually ScanHTML can convert the following:

/ Replace the complete path

./ Do nothing (current directory)

../ Back one directory

None of the above:

Replace the existing name (if any).

"./" & "../" can be in any number and any order.

1.20 ScanHTML.guide/Arguments/PATTERN

P=PATTERN/K

If present, all urls not matching this pattern will be discarded. If the argument BASE is also supplied, the url is completed before testing the pattern.

Any AmigaDOS wildcard can be used.

1.21 ScanHTML.guide/Arguments/PATTERN2

P2=PATTERN2/K

If present, all urls not matching this pattern will be discarded. If the argument BASE is also supplied, the url is completed before testing the pattern.

Any AmigaDOS wildcard can be used.

This pattern matching is totally independant of the one specified with PATTERN .

1.22 ScanHTML.guide/Arguments/STRIP#

STRIP#/S

When present, if a '#' is present in the URL, it is truncated.

Generally what follows '#' simply means to download the file before the '#' and then search for a tag named from what follows '#'.

For example: http://www.voila.com/index.html#DATA

This means to download http://www.voila.com/index.html and then to search a tag named "DATA".

1.23 ScanHTML.guide/Arguments/NOBG

NOBG/S

When given, the url specified with

<BODY BACKGROUND="url">

won't be kept. This url represent the background image to display when viewing an HTML file.

1.24 ScanHTML.guide/Arguments/NOFIRST#

NOF#=NOFIRST#/S

This will discard all urls beginning with '#'. Such an URL simply means to search for a tag in the current HTML file.

This argument is verified before Strip# . In other words, an URL beginning with "#" will always be discarded regardless of the Strip# switch.

1.25 ScanHTML.guide/Arguments/NOHREF

NOHREF/S

Links from one HTML file to another are specified this way in an HTML file:

<A HREF="url"> Clickable text </A>

"url" can be an image, an HTML file, a movie or any other file.

If this switch is given, urls referenced this way won't be kept.

1.26 ScanHTML.guide/Arguments/NOSRC

NOSRC/S

When you view an HTML file with a graphical web browser, there are images that are shown with the text. These images are specified this way in an HTML file:

<IMG SRC="image_file_name">

SRC is also used in some non-standard kludges. The only one I know about for now are FRAMEs used by NetScape.

If this switch is given, there urls (FRAMEs, IMGs, etc..) won't be kept.

1.27 ScanHTML.guide/Arguments/NOQUERY

NQ=NOQUERY/S

Urls may contain '?' which is used to pass arguments to a program executed on the server. For example it can be used by search engines to search keywords in their databases. Some counters (ie "You are the 131352e to visit this page") also use this to know which page is "counted".

If given, urls containing "?" are discarded.