GrabHTTP

Serge Emond

COLLABORATORS

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

REVISION HISTORY

NUMBER	DATE	DESCRIPTION	NAME

Contents

Chapter 1

GrabHTTP

1.1   GrabHTTP.guide/Contents

_____

GrabHTTP

Version 0.007

Copyright (c) 1996 Serge Emond

_____

1.2   GrabHTTP.guide/About English

A Note About English

You will find in this document sentences incorrectly formed. Bad English. Very bad. My orthograph is bad, my verbs are bad, everything is bad.

I write the documentation of my programs in English because it's the only way to write ONE document that anyone can understand.

Sometimes when I browse new files on AmiNet, I see files that I must grab. So I grab them. It always happens to the one that seems the most interesting: After minutes of downloading, no English documentation. Even the program is in a forein language. The first thing I do after such a discovery is to delete the lot. (Except if the forein language is French, but it never happened to me until now.)

It would be *LOT* faster for me to write this document in French but...

Anyway, if you didn't get the point, all the flafla above means:

There are many grammatical errors in this document and I don't care about them. What is important to me for now is to produce a readable documentation, nothing more. (And I'm sure it's readable...)

1.3   GrabHTTP.guide/Introduction

Introduction

You enter on Yahoo, do a search about "Brain" and find a marvelous site containing lots of data on the Brain. You want to grab everything but all the information is scattered across thousands of .html files. A real nightmare to grab everything with a web browser.

More, one of your hobbies is to view pictures. It relaxes yourself and you have enough space on your harddrive not to bother with selecting the images before you download them. Yesterday you found a superb site containing thousands of sunsets. You don't want to stay 20 hours to click each one and press enter in the requester. Worst some of them are not clickable and your browser don't have any option to save those to disk, you have to find the original filename in the html file and copy the file from the disk cache (if there is one!).

The solution: a robot that can download a file, parse it to find other files, grab them, and so on. More, it should grab some files and forget others. It should allow to download only during x hours, no more than y megs or z files.

GrabHTTP is an ARexx host that can retreive HTTP files. Alone this program is pretty useless but it offers many options those who have the time (and know how :) to create a script in ARexx.

1.4   GrabHTTP.guide/Features

Features

- Not available yet - read the docs!

1.5   GrabHTTP.guide/Legal information

Legal information

License

No warranty

1.6   GrabHTTP.guide/License

License

GrabHTTP 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   GrabHTTP.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   GrabHTTP.guide/Requirements

Requirements

GrabHTTP should work on any Amiga equipped with OS 3.0 or higher. In

fact, you need at least OS v39 (3.0) to use it.

It will also need AmiTCP. I don't know which version you need but it

works with 4.2. GrabHTTP also works with Miami.

1.9   GrabHTTP.guide/History

What has been done

Version 0 Revision 1 - 16.8.1996 and before

- Nothing recorded.

Version 0 Revision 2 - 16.8.1996

- Checks if port already exists

- ARexx Command VERSION

- Read/write to socket not buffered anymore

- A lot of other unlogged stuff

Version 0 Revision 3 - 19.8.1996

- Replaced console output with window

Version 0 Revision 4 - 9.9.1996

- Now fully RFC-1945 compliant

- Use HEAD instead of GET if HeaderOnly opt is used. (Smarter)

- Does not enforce ifmodified anymore

- More information when an error occurs

- Strip CR/LF when using GetHeaderString()

- Bug fixed in SetTimeZone: hours were directly added to mins (and not $\times$ 60)

Version 0 Revision 5 - 10.9.1996

- Does not use TCP: anymore (yé!)

Version 0 Revision 6 - 7.12.1996

- New CLI option: write buffer size

Version 0 Revision 7 - 16.12.1996

- Should not add "index.html" randomly to the filename.

- New ARexx function: TRANSLATE

1.10   GrabHTTP.guide/What's next

What's Next - The To Do List

· Proxy support

· Add program's return code to the guide

· Use a separated process to grab a file

· Allow any number of grab with a single grabhttp started

· Allow different config structs for each process

· Completely font independant progress

1.11   GrabHTTP.guide/Bugs

Known Bugs

- None yet (wow!)

1.12 GrabHTTP.guide/Credits

Credits GrabHTTP was completely written by Serge Emond. Since I could not access any RFC during this summer, I looked in GetURL.rexx to see how HTTP worked. Then I tried it with the only HTTP server I could use in my situation (no internet access :), aws 0.97 which is configured in my AmiTCP. I now have RFC-1945 documentation and GrabHTTP has been completely revised to comply to it instead of the suppositions comming from the analysis of aws or GetURL.

aws is a small but efficient HTTP server written by Mike Meyer. Have a look in FTP://max.physics.sunysb.edu/pub/amiga/amosaic/aws/.

GetURL.rexx is an ARexx script written by James Burton. Have a look on AmiNet...

1.13 GrabHTTP.guide/Author

About the author

If you wish to send me comments, bug reports, gifts, money or whatever, then use this address:

Serge Emond

3392 des Anémones

Jonquière, Québec

Canada, G7X 7V3

I do not live there most of the time but whatever is sent there WILL reach me.

You can also try to email me to:

emonds@ere.umontreal.ca

1.14 GrabHTTP.guide/MIME Types

MIME

What follows come directly from AWeb's HTML documentation:

MIME (Multipurpose Internet Mail Extensions) is a mechanism for specifying and describing the format of Internet message bodies. It was primarily designed for e-mail, but MIME types are used also to identify the type of data in the HTTP protocol, the most widely used protocol on the World Wide Web.

A MIME type consists of a type and a subtype. The type describes the major class of data, like text or image. The subtype is used for a subdivision of the major type into different formats, like GIF or JPEG images.

According to RFC 1521, the following official MIME types are defined:

TEXT/HTML

This is a document in the HTML hypertext format. Virtually all pages on the Web are in this format.

TEXT/PLAIN

This type is used for plain text documents (normally in ASCII).

APPLICATION/OCTET-STREAM

This describes a binary file. The file could be processed by some application. An example of this would be an LHA archive.

APPLICATION/POSTSCRIPT

The document is in PostScript format.

IMAGE/GIF

IMAGE/JPEG

These are images, in GIF and JPEG format.

AUDIO/BASIC

This type is used for audio data encoded using 8-bit ISDN mu-law [PCM].

VIDEO/MPEG

This is an animation in MPEG format.

In addition to these official types and subtypes, it is allowed to define extension MIME types and subtypes. These should start with X- to avoid collisions with future official MIME types.

1.15 GrabHTTP.guide/HTTP Protocol

1.16 GrabHTTP.guide/Progress Indicator

GrabHTTP's progress indicator looks like this:

Url: <url>

File: <file>

Size: <size> Actual: <actual>

Received: <received> Remaining: <remaining>

Status: <status> Chars/sec: <cps>

100%

Here's are the meaning of each field:

<url>

This is the URL GrabHTTP asked for.

<file>

This the name of the file where GrabHTTP is saving the url.

<size>

This is the size of the file. If the remote server does not give this information, the field remains blank.

<received>

This is the number of bytes received.

<status>

This is the status of the transfert. Errors are reported here.

<actual>

This is the time the transfer took up to now.

<remaining>

This is the calculated remaining time.

<cps>

This is the rate at which we are receiving the file - number of bytes/sec.

<100%>

This the progress bar: it fills as we receive the file. It remaining completely blank if the size information is unknown.

You can hit the Close Gadget to abort the tranfert.

1.17 GrabHTTP.guide/Browsing Anonymously

Browsing Anonymously

An HTTP request can have a field containing email and name information about the user. This way, some people send their email address on every server they connect to. Most of the time it won't do anything but... there are twisted webmasters who use this information to send anoying emails ie to sell products.

It you want to see if your Web Browser sends such information without letting you knowing about it, have a look at this page: HTTP://www.uiuc.edu/~ejk/WWW-privacy.html.

By default GrabHTTP won't send any information about the user. In fact, there is absolutely no function allowing to grab any information from your system so you can be sure it won't send anything. :)

If you really want to send your email, there is an ARexx command allowing to do so. (SetEMail)

1.18 GrabHTTP.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.

1.19 GrabHTTP.guide/Arguments

Arguments

GrabHTTP can only be started from a Shell. It is pure and can be made resident.

You can break it at anytime by sending it a CTRL-C break. When receiving a file, it won't break until the file has been completely received or an error occured. When the transfer has been completed, it will exit. You can force the break by clicking on the close gadget.

Text arguments

PORT Name of the ARexx port.

ACCEPT What type of file to accept.

FONT Name of the font to use.

FONTSIZE Size of the font to use.

WBUF Size of write buffer.

1.20 GrabHTTP.guide/Arguments/Port

Port{@uu} (PORT)

It represents the name of the ARexx port GrabHTTP should open at run time. The default port name is "GRABHTTP".

You should use an uppercase port so you won't have to enclose the port's name in quotes in your ARexx scripts.

See also:

ARexx

1.21 GrabHTTP.guide/Arguments/Accept

Accept (ACCEPT/K)

When present, it overwrite the default setting of "*/*". This represent the

MIME types to accept. An asterix means "all".

For example, if you only want to receive images, you can specify "image/*".

If you only want to receive HTML files, specify "text/html".

1.22 GrabHTTP.guide/Arguments/Font

Font (F=FONT)

This is the font to use in the progress display . The default font is "topaz.font".

1.23 GrabHTTP.guide/Arguments/FontSize

FontSize (FS=FONTSIZE/N)

This is the size of the font to use in the progress display . The default size is 8.

1.24 GrabHTTP.guide/Arguments/WriteBufferSize

WBuf (WBUF/N)

This is the size of write buffer to use in bytes. GrabHTTP write to disk only when the buffer is full. If the given value is smaller than 2048 then 2048 is used.

1.25 GrabHTTP.guide/Using ARexx

Introduction

GrabHTTP is above all an ARexx server. This means when started it

sits there and wait for ARexx commands.

Each time a command is received, two things are set in the calling rexx

script: the RC variable and a special variable named "GHRES". GHRES will be

set to a NULL-string if nothing is to be returned so copy it if you want to

use what's in it after calling another command.

At anytime you can receive one of these return-codes:

10 Command unknown/mispelled

15 Error setting secondary result variable (GHRES)

Here's a list of all ARexx commands supported:

GETBYTERECD

GETCPS

GETHEADERSTRING

1.26   GrabHTTP.guide/ARexx/GetByteRecd

NAME

GetByteRecd -- Get the number of bytes received during last transfer.

SYNOPSIS

GetByteRecd

FUNCTION

It returns the number of bytes received (and wrote to disk) during the last transfert.

ARGUMENTS

None.

RC

0 No error.

GHRES

Bytes received.

SEE ALSO

1.27   GrabHTTP.guide/ARexx/GetCPS

NAME

GetCPS -- Get the CPS rating of the last transfer.

SYNOPSIS

GetCPS

FUNCTION

Returns the CPS rating (characters per second) the speed of the last transfer. This is calculated by dividing the number of received bytes per the total time of transfert.

ARGUMENTS

None.

RC

0 No error.

GHRES

Contains the last CPS rating or 0 if no transfer yet or if it failed before any CPS rating could be calculated.

SEE ALSO

1.28 GrabHTTP.guide/ARexx/GetHeaderString

NAME

GetHeaderString -- Get information contained in header.

SYNOPSIS

GetHeaderString STRING

FUNCTION

This returns information contained in the header the remote host sent. An header is built this way:

HTTP/1.0 200 OK

Server: aws/0.97

Date: Sat, 17 Aug 1996 21:20:50 GMT

Content-Type: text/html

Last-Modified: Thu, 05 Oct 1995 23:54:06 GMT

Content-length: 456

(etc....)

The arguments specifies the left part of the string you want to know about. This is case insensitive. Example:

address grabhttp 'GetHeaderString sErV'

if rc ~= 0 then say ghres

else say 'No server information'

The above example would print the line "Server: aws/0.97" if this information is available.

Notes: · There is NO CR & LF.

· You can obtain the size of the file with this command but GetSize is more efficient.

ARGUMENTS

STRING

String to search. Default string: "HTTP/1.0"

RC

0 No error.

1 String not found.

5 Error parsing arguments.

GHRES

Contains the string requested.

SEE ALSO

GetSize

1.29 GrabHTTP.guide/ARexx/GetHTTPFile

NAME

GetHTTPFile -- Grab a file using HTTP transfer protocol.

SYNOPSIS

GetHTTPFile URL/A FileName/K MaxSize/N HeaderOnly/S IfModified/S Progress/S SaveHeader/S

FUNCTION

GetHTTPFile tells GrabHTTP to get a file and save it to disk.

ARGUMENTS

URL (string)

Required. Name of the URL to grab. Must begins with "HTTP://".

FileName (string)

Name of the file to save. If it ends with a slash ('/') then "index.html" will be added. All inexistent directories will be created. If not specified, a filename will be created from the URL given. In that former case there won't be any path, the file will be saved in the current directory . If the filename contains '#', it will be truncated.

MaxSize (int)

This sets the maximal number of bytes a file can have. If the file to receive is bigger, the transfer will be aborted immediately after receiving the header. In such a case, '106' will be returned in GHRES. If this value is not specified, or if it is set to 0, then there won't be any size-checking.

HeaderOnly (switch)

Tells GrabHTTP to get the header and to abort the transfer at the start of the file.

IfModified (switch)

GrabHTTP will send a "If-Modified-Since: " flag when sending the request. If the remote server support this, it will only send the file if it is newer than the file on disk. This flag is useless if the file is not present on disk. You should set correctly the timezone prior to use this flag.

Progress (switch)

When present, this tells GrabHTTP to display progress information about the actual file transfer.

SaveHeader (switch)

This tells GrabHTTP to save the header along with the file, or only the header if HeaderOnly is set. The filename for the header will be the same as for the file, with ".HDR" added.

RC

0 Result of the transfer in GHRES.

5 Failed to parse args.

GHRES

0 Everything went OK.

98 Unable to alloc mem.

99 Unable to alloc mem.

100 Remote did not reported "200 OK". (See GetHeaderString)

101 Bad/unsupported transfert method. (Should begins with HTTP://)

102 Failed to open socket (Ie host does not exists, no answer).

103 Failed to send HTTP request.

105 Failed to create required directories.

106 File is too big (See MSize option).

107 Not enough space left on device (See SetMinSpaceLeft).

150 Error openning output file.

151 Written less bytes than received (error writing?).

161 Error reading header.

162 Error openning/writing header's file.

163 Header do not begin with HTTP/1.0.

164 Error allocating memory for header information (ie to use with GetHeaderString)

201 User pressed CloseWindow gadget

SEE ALSO

1.30   GrabHTTP.guide/ARexx/GetSize

NAME

GetLastModified -- Get the size of the file.

SYNOPSIS

GetSize

FUNCTION

Returns the "Size:" field of the transfer header received during last transfer.

RC

0 No error.

GHRES

Contains the size of the file received or an empty string if this information wasn't given.

SEE ALSO

GetHeaderString

1.31   GrabHTTP.guide/ARexx/GetTicks

NAME

GetTicks -- Get the total time the transfer took.

SYNOPSIS

GetTicks

FUNCTION

Returns the total time the transfer took in ticks. One second contains 50 ticks.

ARGUMENTS

None.

RC

0 No error.

GHRES

Contains the time the transfer took in ticks.

SEE ALSO

1.32 GrabHTTP.guide/ARexx/Quit

NAME

Quit -- Makes GrabHTTP exit.

SYNOPSIS

Quit

FUNCTION

Upon reception, GrabHTTP will free all its memory, close all what is openned, and... quit.

ARGUMENTS

None.

RC

0 No error.

GHRES

Nothing.

SEE ALSO

1.33 GrabHTTP.guide/ARexx/SetDir

NAME

SetDir -- Change GrabHTTP's current directory.

SYNOPSIS

SetDir DIRECTORY

FUNCTION

Makes GrabHTTP change its current directory. Saved files without complete path will be saved relatively to this directory. If no SetDir command is received, the current directory is the one where the program has been started.

ARGUMENTS

Directory (string)

Name of the directory to go to. It will not be created if it do not exists. If no directory is specified, it will go to "PROGDIR:".

RC

0 No error.

1 Failed to change directory.

5 Failed to parse args.

GHRES

Nothing.

SEE ALSO

1.34 GrabHTTP.guide/ARexx/SetEMail

NAME

SetEMail -- Set user information.

SYNOPSIS

SetEMail EMAIL

FUNCTION

Set user's email address. If not set, no user information will be sent and the user will be anonymous (see Browsing anonymously
). The string should be something like:

"UserName <email@address>"

ARGUMENTS

EMail (string)

String containing name/email address of the user. If absent, user will stay anonymous.

RC

0 No error.

1 Failed to alloc memory to store email info, user is forced anonymous.

5 Failed to parse args.

GHRES

Nothing.

SEE ALSO

1.35 GrabHTTP.guide/ARexx/SetMinSpaceLeft

NAME

Set the minmla space to leave on device.

SYNOPSIS

SetMinSpaceLeft BYTES/N

FUNCTION

Set the minimal space that should be left on the device after reception of the file. If there is not enough space on the disk for the
file, GetHTTPFile will return 107 in GHRES. This do not refers to current directory 's device but to the device that will contain
the file. No checking is done if this is set to 0, which is the default.

ARGUMENTS

Bytes (int)

Number of bytes to keep free on the device receiving the file. The default is 0 (feature disabled).

RC

0 No error.

5 Failed to parse args.

GHRES

Nothing.

SEE ALSO

1.36   GrabHTTP.guide/ARexx/SetTimeZone

NAME

SetTimeZone -- Set local timezone information.

SYNOPSIS

SetTimeZone H=HOURS/N M=MINUTES/N

FUNCTION

Set the time you must add to the local time to obtain GMT time. This is currently only used when using the IfModified switch of GetHTTPFile . Minutes can range between -1440 and +1440 and hours between -24 and 24. The total time must not exceed ± 24 hours (± 1440 minutes). The default is EST5.

ARGUMENTS

Hours (int)

Number of hours to add to local time to obtain GMT. If absent, 5 is used.

Minutes (int)

Number of minutes to add to local time to obtain GMT. If absent, 0 is used.

RC

0 No error.

1 GMT cannot be more than 1 day far. Old value kept.

5 Failed to parse args.

GHRES

SEE ALSO

1.37   GrabHTTP.guide/ARexx/Translate

NAME

Translate -- Change some characters in a string.

SYNOPSIS

Translate STRING/A,TOKEN/A,TO/A

FUNCTION

Changes all characters TOKEN into TO in the string STRING. For example, if you give

TRANSLATE "Hey didlo!" "!ed" "_*x"

You would obtain:

H*y xixlo_

ARGUMENTS

String (string)

String to translate.

Tok (string)

"From" characters.

To (string)

"To" characters.

RC

0 No error. New string in GHRES.

1 "TOK" contains more characters than "TO".

5 Failed to parse args.

GHRES

If RC=0, contains translated string.

SEE ALSO

1.38   GrabHTTP.guide/ARexx/SetTimeZone

NAME

SetTimeZone -- Set local timezone information.

SYNOPSIS

SetTimeZone H=HOURS/N M=MINUTES/N

FUNCTION

Set the time you must add to the local time to obtain GMT time. This is currently only used when using the IfModified switch of GetHTTPFile . Minutes can range between -1440 and +1440 and hours between -24 and 24. The total time must not exceed ±24 hours (±1440 minutes). The default is EST5.

ARGUMENTS

Hours (int)

Number of hours to add to local time to obtain GMT. If absent, 5 is used.

Minutes (int)

Number of minutes to add to local time to obtain GMT. If absent, 0 is used.

RC

0 No error.

1 GMT cannot be more than 1 day far. Old value kept.

5 Failed to parse args.

GHRES

SEE ALSO