🌎 WebGrab Help Topics

Product Information:
License and Warranty Information
Registering Your Software
Technical Support
Product Overview

Usage Information:
Grabbing Websites
Smart Grabbing
Dialup Friendly Features
Skipping Dynamic Pages
Following External Links
Excluding Files by Type
Excluding Hyperlinks
Grabbing a List of Hyperlinks
Grabbing a Password Protected Site
Changing File Associations
Logging Events
Clocking a Download
Windows/UNIX Command Line Console Utility

Configuration Information:
Changing External Application Associations
Changing Column Names
Turning Off Confirmations Messages
Setting the Maximum Thread Limit
Changing the Timeout Value
Setting the Number of Cache Segments

Database Operations:
Changing Your Grab Directory
Browsing the Grab Directory
Adding Entries
Deleting Entries
Modifying an Entry
Search Operations
Printing a Database
Opening an Existing Database
Creating a New Database
Renaming an Existing Database

Database Utilities:

<u>Importing a Database</u>
<u>Exporting a Database</u>
<u>Creating Import Templates</u>
<u>Merging a Database</u>
<u>Backing up a Database</u>
<u>Troubleshooting</u>

General Information:
<u>Using the Toolbar</u>
<u>Using the Edit Menu</u>
<u>Using the View Menu</u>
<u>Using the Help Menu</u>

Documentation produced by: glNET Software

http://www.glnetsoftware.com
glnet@glnetsoftware.com

glNET Software
P.O. Box 820314
South Florida, FL. 33082-0314
FAX: 1-800-339-7068

License and Warranty Information

WebGrab version 5.0 for Windows 95/98/ME/NT/2000/XP
WebGrab version 5.0 for Linux 2.3.1 and higher
Copyright (c) 1999-2002 glNET Software
All Rights Reserved

Thank you for your interest in WebGrab.

PLEASE READ THIS SOFTWARE LICENSE AND DISCLAIMER AGREEMENT CAREFULLY. ALL USERS OF WEBGRAB MUST ACCEPT THIS DISCLAIMER OF WARRANTY AND LICENSE AGREEMENT STATED BELOW IN ITS ENTIRETY. BY USING THIS SOFTWARE AND/OR ANY OF ITS COMPONENTS, YOU ARE AGREEING TO BE AND ARE HEREBY BOUND TO ALL THE TERMS AND CONDITIONS OF THIS AGREEMENT, WITHOUT EXCEPTION. IF YOU DO NOT AGREE WITH ANY OF THE TERMS AND CONDITIONS OF THIS LICENSE AND DISCLAIMER AGREEMENT, PLEASE REMOVE THIS SOFTWARE AND ALL OF ITS COMPONENTS FROM YOUR COMPUTER. PLEASE NOTE THAT THE TERMS OF THIS AGREEMENT ARE SUBJECT TO CHANGE WITHOUT PRIOR NOTIFICATION.

DISCLAIMER – **AGREEMENT**:
1. This software, documentation, and technical support is provided "AS IS" and without warranty of any kind. Should any element of this software prove defective, you will assume the cost of all-necessary servicing, repair, or correction. Licensee expressly acknowledges and agrees that the use of this software and its documentation is at Licensee's sole risk. glNET Software disclaims all warranties, expressed or implied, including, without limitation, the warranties of merchantability and of fitness for any purpose.

2. glNET Software assumes no liability for damages, direct or consequential, which may result from the use of WebGrab. glNET Software does not warrant that the functions contained in this software will meet Licensee's requirements, or that the operation of the software will be uninterrupted or error-free, or that any defects in this software will be corrected. Furthermore, glNET Software does not warrant or make any representations regarding the use or the results of the use of this software or its documentation in terms of their correctness, accuracy, reliability, or otherwise.   No oral or written information or advice given by glNET Software shall create a warranty or in anyway increase the scope of this warranty.

3. Limitation of Liability: Under no circumstances, including negligence, shall glNET Software, glNET Software affiliates or their directors, officers, employees or agents be liable for any incidental, special or consequential damages. This includes damages for loss of business profits, business interruption, loss of business information and the like arising out of the use or inability to use this software or its documentation, even if glNET Software or an authorized representative has been advised of the possibility of such damages. In no event shall glNET Software's total liability to licensee for all or any

damages, losses, and causes of action (whether in contract, tort, including negligence, or otherwise) exceed the amount paid by licensee for this software and its documentation.

<u>LICENSE – AGREEMENT FOR USE AND DISTRIBUTION</u>:
1. The terms of this agreement apply to both the DemoWare and retail versions of WebGrab with distinctions for each noted. The WebGrab demo is a "shareware program" and is provided at no charge to the user for an unspecified evaluation period. Anyone may use the DemoWare version of WebGrab for demonstration and evaluation purposes only. The DemoWare manager limits the number of grabs and the number of items that can be entered into the database. If you purchase a retail version of WebGrab you will be supplied with a fully functional upgrade that has no preset limitations.

2. You may not use, copy, emulate, clone, rent, lease, sell, modify, decompile, disassemble, otherwise reverse engineer, or transfer the DemoWare or retail version of WebGrab, or any subset of it.   Any such unauthorized use shall result in immediate and automatic termination of this license and may result in criminal and/or civil prosecution. Please note that tacit agreements or spoken words do not constitute consent to use this software in any manner defined or not defined in this agreement. All agreements must be made in writing in either hardcopy or electronic form. If you are in doubt about any part of this agreement please contact gINET Software.

3. The WebGrab DemoWare version may be freely distributed provided the distribution package is not modified in any fashion. Feel free to share the DemoWare version with your friends, but please do not give it away altered or as part of another system. The essence of "user-supported" software is to provide personal computer users with quality software without high prices, and yet to provide incentive for programmers to continue developing new products. You may not decompile the DemoWare version or make any attempt to turn the DemoWare version into a fully functional retail version.

4. If you find this program useful and wish to use upgrade to the retail version please remit a registration payment. The one time registration fee will license one copy of WebGrab to a single individual for use on multiple computer systems at any time. Each licensed copy may be used by a single individual and may be freely moved from one computer location to another, so long as there is no possibility of it being used in place of an unlicensed DemoWare version. You may not distribute, post, upload, attach, or otherwise give away a retail version of WebGrab or any subset of the upgrade package. You may distribute, post, upload, attach, or forward any information provided to you during the registration process.

5. Anyone distributing a DemoWare version of WebGrab for any kind of remuneration must first contact gINET Software for authorization.   Otherwise, you are encouraged to pass a copy of the DemoWare version along to your friends for demonstration and evaluation. Please encourage them to upgrade to the retail version if they find that they can use WebGrab. All registered users are entitled to free upgrades of WebGrab for the

life of the product, notification of new releases, and technical support if you should require it.

<u>WARRANTY FOR REMOVABLE MEDIA</u>:
The warranty for removable media is only applicable to this license agreement if WebGrab has been supplied to you on a CDROM or floppy disk. There are no returns or refunds for CDROM or floppy disk distributions removed for the original packing material. Returns or refunds will only be honored for unopened CDROM or floppy disk distributions for a period of 30 days after the date of purchase. If the CDROM or floppy disk is defective it may be substituted up to 30 days after the date of purchase. glNET Software is not responsible for shipping and handling costs on returned merchandise.

Registering Your Software

Thank you for using WebGrab. The WebGrab demo is a "shareware program" and is provided at no charge to the user for an unspecified evaluation period. Anyone may use the DemoWare version of this software for demonstration and evaluation purposes only. The DemoWare manager limits the number of grabs and the number of items that can be entered into the database. If you find this program useful and wish to upgrade to the retail version please remit a registration payment. All registered users will receive free upgrades of WebGrab as they become available, automatic notification of new releases, and technical support should you require it. **NOTE**: All registered users of WebGrab version 5 for Windows will receive a free version of the WebGrab command line utility. All registered users of WebGrab version 5 for Linux will receive a free version of the WebGrab command line utility for UNIX.

For purchasing information and secure server on-line ordering please visit the WebGrab Home Page at:

http://www.glnetsoftware.com/webgrab

After your order is processed a glNET Software representative will contact you within 24 to 48 hours with instructions for downloading the retail version of WebGrab.

After receiving your retail copy of the software you may upgrade to the full version by installing the retail version over the DemoWare copy. All databases created with the DemoWare copy will not be affected during the upgrade installation process.

Technical Support

Technical support and database recovery is freely available to all registered users and to unregistered users for installation questions only. Should you require technical support on this product or to request custom modifications contact glNET Software technical support at:

glnet@glnetsoftware.com

Subject: WEBGRAB SUPPORT

Please include your name and return email address in the body of the message. When requesting technical support, please be sure to state the exact problem and all the circumstances surrounding it. If possible, include the version of your operating system, hardware platform, and the installed version of this program. To obtain this program's version number, select the "About WebGrab" option from the Help Menu.

For the latest version of WebGrab and new product releases, please visit the WebGrab Home Page at:

http://www.glnetsoftware.com/webgrab

To contact glNET Software via U.S. mail please send written inquires to:

glNET Software
P.O. Box 820314
South Florida, FL. 33082-0314
FAX: 1-800-339-7068

Product Overview

WebGrab

The simple way to download and organize Web pages.

WebGrab is a simple and convenient way to save and organize Web Pages for off-line viewing. How many times have you bookmarked Web pages for future reference only to find that page is no longer in existence? The information you were relying on is gone and the search and collection process starts all over again. Don't let this happen to you! Using WebGrab you can easily grab files and images from a Web Server and store them on your local hard drive burn them to
CD. WebGrab is also an excellent tool for maintaining your Web Site.

Program Features:
- Easy to use graphical interface
- Dialup friendly features included for slow Internet connections
- Allows the capture of a single Web page or an entire Web site
- Grab selectively using hyperlink, object, and file exclusions
- Follow any number of external hyperlinks
- Multithreaded interface allowing you to capture multiple sites simultaneously
- Automatically recreates Web directory structures on your local drive
- Allows captured Web documents to be viewed offline using your favorite Web Browser
- Records each grab to a searchable database file
- The WebGrab status window displays information about each transfer
- Optional logging features record detailed information about each capture
- The clipboard options allow you to copy and paste URLs from your Web Browser
- Includes print and print preview facilities
- Includes context sensitive help
- Includes command line grab utility

Program Requirements:
- 486 DX4 75 or higher
- 16MB of memory or higher
- Windows 95/98/ME/2000/XP
- 5MB of disk space
- Microsoft TCP/IP stack
- WinSock version 1.1 or higher

Grabbing Websites

WebGrab has three different capture modes, which gives you the option to capture a single Web page, a single Web page plus all its relative links, or an entire Web site. The WebGrab engine will recreate the Web server's directory structure on your local hard drive. Select the desired grab type prior to each capture:


- **Single Page** - Select this option to grab a single Web document. A single page download will not following any links to other documents or image files contained in the page being captured.
- **Page Plus Links** - Select this option to grab a single Web document plus any links to other documents or image files contained in the page being captured.
- **Entire Web Site** - Select this option to download an entire Web site to your local hard drive. The download will start at the page specified in the location box and recursively follow links down the directory tree to documents and image files contained on the Web site.

To grab a Web page simply enter or paste a URL (Uniform Resource Locator) in the WebGrab location box:

After entering the URL press the "Enter" key or click on the "Start" button to initiate the Web page capture:
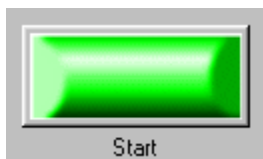
The WebGrab status bar will continually update the progress of the download as the page and/or any of its links are captured:

Ending grab: http://www.glnetsoftware.com/webgrab

If for any reason the download stalls or you wish to terminate this capture click on the stop button. Please note that the stop button will stop all grabs.

To view the WebGrab status window click on the status button located on the WebGrab status bar:

If a grab fails for any reason the status button will turn red to indicate and error. All detailed error messages will be printed in the WebGrab status window.


After successfully capturing a Web page a hyperlink to the document or image file will appear in the WebGrab database:

Site Name	Hyperlink
www.glnetsoftware.com/webgrab	file:///C:/webgrab/www.glnetsoftware.com/webgrab/index.htm

Double click on the hyperlink to view the document using your favorite Web browser.

The WebGrab history file is a short-term record of all attempted page captures.

Grab History

http://www.glnetsoftware.com/webgrab - Oct 11, 2001 20:58:35
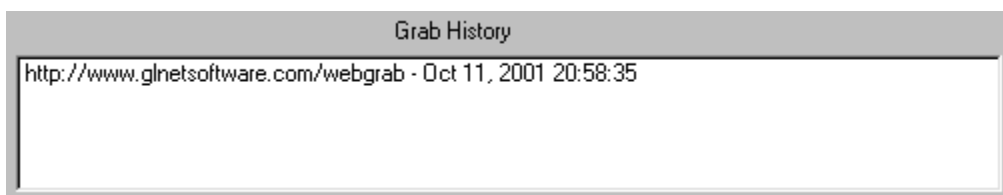
To clear the WebGrab history file select the "Clear History File" option from the "Database" menu.

Smart Grabbing

The smart grab feature enables WebGrab to download files that have only changed on the Website since the last grab. Smart grabbing speeds up downloads considerably and allows you to resume a stopped or stalled grab without starting over.

The smart grab feature can be enabled or disabled from the Grab menu:

The smart grab feature works by recording page header information for each object grabbed from a Website. There are some additional disk space requirements when the smart grab feature is enabled. If your disk space is limited you can disable the smart grab feature by clicking on the Enable Smart Grab menu option. When disabled this option will be unchecked. To enable this option click on the option again and this option will be checked.

When enabled the smart grab feature creates a database file in the URL's grab subdirectory. The smart grab database file can only be read by the WebGrab application. In addition to the smart grab database a plain text log file will be created in the URL's grab subdirectory. The smart grab log file is a human readable version of the database file and can be used for troubleshooting and statistical information. If you wish to disable smart grab logging click on the Enable Smart Grab Log menu option. When disabled this option will be unchecked. To enable this option click on the option again and this option will be checked.

Please note that not all Web servers return page information when a Web object is requested. Page information and the content of page information are dependent on the HTTP server used by the Website and how the HTTP server is configured for that Website. Neither WebGrab nor any Web browser has control over the page information returned by a Web server. The smart grab feature will not work if the server does not return any page information or returns incomplete page information. In these situations Web objects will be downloaded each time the Web object is requested.

Dialup Friendly Features

WebGrab version 5.0 and higher offers some new features to make the program friendlier when used with dialup Internet connections. Dialup connections are very slow compared to DSL and direct Internet connections. Another downside to dialup connections is the fact that a dialup connection is prone to frequent disconnects and other erratic service interruptions. But despite the slow connections speeds and unpredictable behavior dialup connections are widely used because the cost of a dialup connection is significantly less than DSL and direct Internet connections.

The dialup friendly features can be found on the Grab menu:
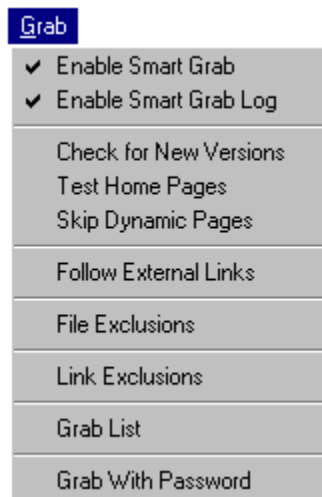
For optimal speed use the configuration shown above. In this configuration you can resume grabs that have been abruptly terminated without having to start over.

For more information about the smart grab feature please refer to the Smart Grabbing section of this document.

The Check for New Version option is used to tell the program whether or not to check for new versions of a page. When this feature is disabled the WebGrab will not attempt to download a Web object if it exists in the smart grab cache. When disabled this option will be unchecked as shown above. To enable this option click on the option this option will be checked. The Check for New Version option should be enabled if you are downloading Websites that continuously change page content.

The Test Home Pages option is used to tell the WebGrab whether or not to check for a new version of a home page. Any hyperlink that does not specify a file is considered a home page, for example:

http://www.glnetsoftware.com
http://www.glnetsoftware.com/webgrab

If a record of the home page exists in the smart grab cache from a previous grab all subsequent grabs will not attempt to download the home again if the Test Home Pages option is disabled as shown above. To enable this option click on the option this option will be checked. The Test Home Pages option should be enabled if you are downloading Websites that continuously change page content.

For more information about the Skip Dynamic Pages feature please refer to the Skipping Dynamic Pages section of this document.

Skipping Dynamic Pages

A dynamic page refers to any Web resource that executes a program on a Web server to generate a Web page. Generally pages with dynamic content can by identified by the dot extension used in file name and/or directory name where files are stored. For example, older dynamic Web technologies such as CGI (Common Gateway Interface) will store executable files in a directory named cgi-bin. Hyperlinks containing a cgi-bin path will be executing a CGI program, which may be an executable script file or binary executable. New Web technologies such as Active Server Page technology, ColdFusion markup language, and Hypertext Preprocessor files can be identified by the dot extensions used in hyperlink's file name, for example:

index.asp - Active Server Page
index.cfm - ColdFusion markup language
index.php - Hypertext Preprocessor

WebGrab gives you the option of skipping dynamic pages or downloading them. **NOTE**: Since a dynamic page is regenerated every time the page is visited dynamic pages will not be stored in the smart grab cache. For dialup Internet users it is recommend to skip dynamic pages during repetitive grabs or grabs that were resumed after a dialup disconnect. To enable or disable dynamic page downloads select the Skip Dynamic Pages option from the Grab menu:
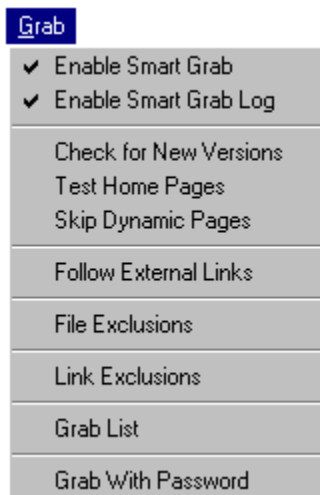
When the Skip Dynamic Pages option is unchecked as shown above dynamic pages will be downloaded. To skip dynamic pages click on the Skip Dynamic Pages option this option will be checked. When this option is checked dynamic pages will be skipped.

As dynamic Web technologies are continuously changing WebGrab allows you to modify how dynamic content is identified inside a hyperlink. To change the dynamic page options select Edit from the main menu and click on the File Types menu item. The "Edit WebGrab file types" dialog will appear:

Adding additional dot extensions to the Dynamic File Types section will inform the grab engine of new dynamic file types.

The URL Query/Execute Strings section informs the grab engine of special characters in a hyperlink used by dynamic Web pages for queries and other executable switches.

The URL Dynamic Directory Names section is used by the grab engine to identify pages with dynamic content by a directory name used in the hyperlink.

Following External Links

When WebGrab downloads a Website the grab engine limits its visibility to the pages contained within in that Website starting with the initial hyperlink grabbed and below. Sometimes Websites contain hyperlinks to other Websites that contain the actual Web resources you were attempting to download. The only way to allow the grab engine to extend the download to other Websites is to enable the external link feature.

To enable external links select Grab from the main menu and click on the Follow External Links menu item. The "External Links Setup" dialog will appear:
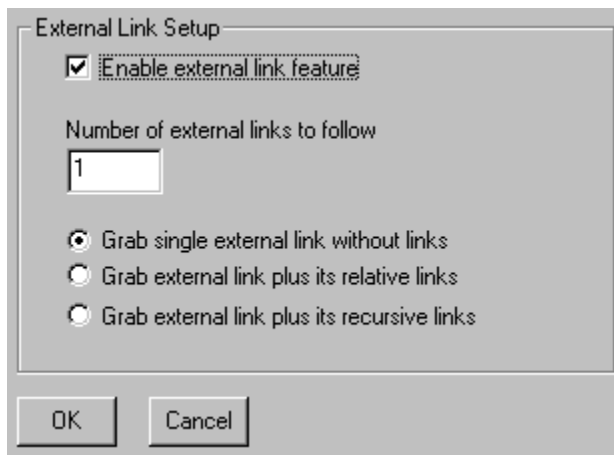
When grabbing external links ensure that the Enable external link feature is checked. To disable the external link feature uncheck this option. **NOTE**: When enabled the external link feature will be applied to the current grab and all subsequent grabs until the external link feature is disabled.

External links can be limited to one layer of external hyperlinks or expanded to any number of external hyperlinks. By default the number of external links the grab engine will follow is limited to one external Website and the first page of that Website. If you wish to follow more than one external link change the Number of external links to follow in the "External Links Setup" dialog. **NOTE**: Always be aware that a large number of external links to follow has the potential of consuming a large amount of disk space and bandwidth.

The "External Links Setup" dialog has three download options when following external links. You can limit the download to a single page, follow the page and its relative links, or follow the page and all the links below it.

Excluding Files by Type

The WebGrab grab engine identifies file types by the character string following the last dot in the file. A file type refers to what is actually contained in a file and how it is used by the operating system and associated applications. For example, some common graphics files end with .bmp .jpg and .gif dot extensions and some common multimedia files end with .mpg .avi .mov .wav and .mp3. Since graphics, multimedia, compressed archives, and others have a tendency to be very large there may be times when you wish to exclude these file types from a grab.

To exclude specific file types select Grab from the main menu and click on the File Exclusions menu item. The "Exclude Files" dialog will appear:

Enter the dot extension of all files to exclude.
Click on OK to save or Cancel to clear this list.

.zip .Z .tar .tgz .gz
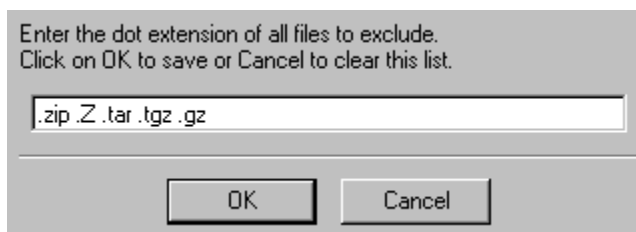
OK   Cancel

As shown above this grab will exclude the most common types of compressed archive files. After entering the file types you wish to exclude click on OK to save the entries or click on the Cancel button to clear the list. **NOTE**: The files types entered in the "Exclude Files" dialog will affect the current grab and all subsequent grabs until the list is cleared.

Excluding Hyperlinks

The link exclusion feature allows you to exclude hyperlinks from a grab. The hyperlinks you wish to exclude from a grab session can be denoted by a complete hyperlink or a partial hyperlink with a * wildcard character prefix, suffix, or midpoint insertion.

To exclude specific hyperlinks select Grab from the main menu and click on the Link Exclusions menu item. The "Exclude Links" dialog will appear:
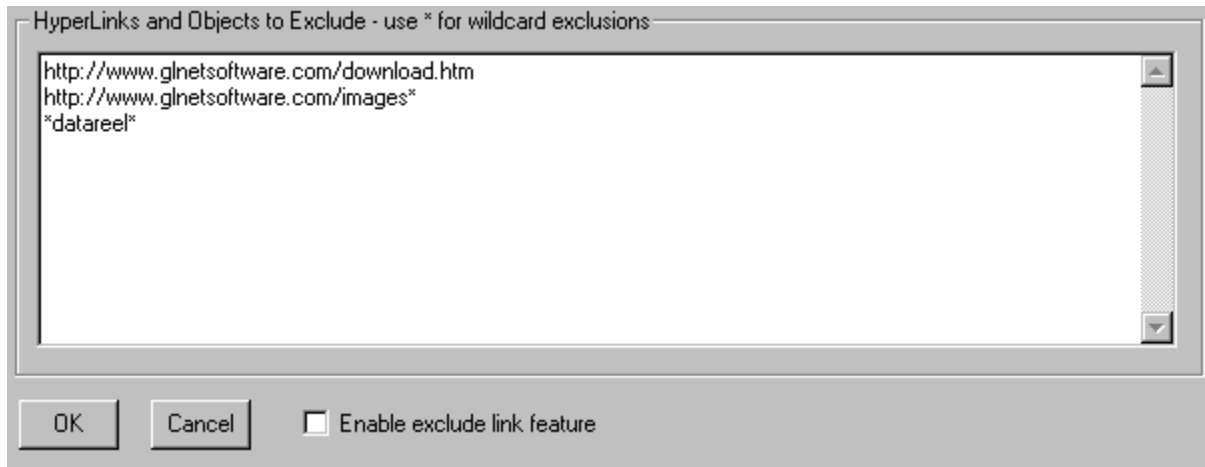
As shown above you may enter the complete hyperlink or use a wildcard character to exclude a group of files or hyperlinks. After entering the list of hyperlinks you wish to exclude you must click on the Enable exclude link feature in order to enable link exclusions. **NOTE**: The hyperlinks entered in the "Exclude Links" dialog will affect the current grab and all subsequent grabs until the link exclusion feature is disabled.

Grabbing a List of Hyperlinks

The grab list feature allows you to grab a list of hyperlinks. This feature is useful when you have a list of hyperlinks that you would like to grab. The list of hyperlinks can be manually entered or imported from an existing text or HTML file.

To use the grab list feature select Grab from the main menu and click on the Grab List menu item. The "Hyperlink List" dialog will appear:
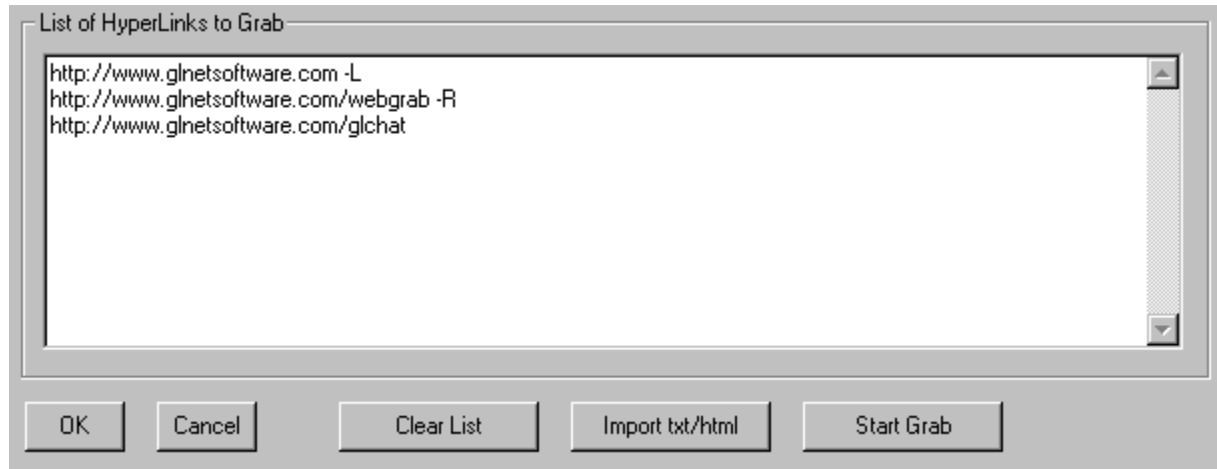
As shown above all the hyperlinks in the list will be grabbed when you click on the Start Grab button. **NOTE**: During a grab list session you will not be able to grab hyperlinks from the main panel.

To import an existing text file or HTML file click on the Import txt/hml button and the "Browsing for import file" dialog will appear. From here you can select a file to import. The import utility will search the import file for all hyperlinks and insert them into the grab list.

After an import or manual entry insertion there some additional options you can set to control how the hyperlinks are downloaded. By default the grab option set on the main panel will be used to download single files, relative files, or recursive files. In the grab list these options can be changed using the -S, -L and -R switches:

```
http://www.glnetsoftware.com -S # Grab a single page
http://www.glnetsoftware.com –L # Grab page and its links
http://www.glnetsoftware.com –R # Grab the whole website
```

To skip a hyperlink in the grab list use the –K switch or place a # character at the beginning of the list entry:

```
http://www.glnetsoftware.com -K # Skip this entry
```

`#http://www.glnetsoftware.com`
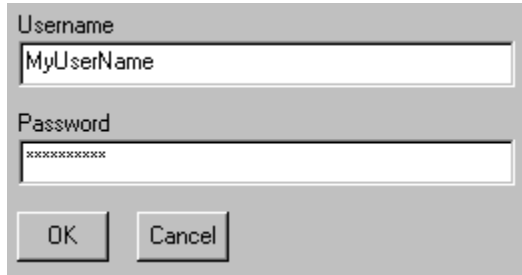
To grab a password protected site from the grab list use the following syntax:

`http://username:password@www.glnetsoftware.com`

NOTE: The username and password entered in the grab list will be stored and transmitted in clear text. Clear text passwords could possibly be intercepted and read if transmitted over untrusted networks. If this username and password combination can in anyway jeopardize your home or business information security please do not enter this information in the grab list.

Grabbing a Password Protected Site

The WebGrab password option allows you to enter a plain text user name and password for sites that require basic HTTP authentication. To grab a password-protected site you must first enter the hyperlink you wish to grab in the main panel's Location: box. After entering the hyperlink select Grab from the main menu and click on the Grab With Password menu item. The "Enter a username and password" dialog will appear:

After entering your username and password click on the OK button and start the grab by clicking on the main panel Start button. The username and password entered will be used for the specified hyperlink until you exit the program. For local security reasons usernames and passwords are only good for the current WebGrab session and will not be stored in any database files, log files, or displayed in any status frames.

NOTE: The username and password submitted with the download request will be transmitted in clear text. Clear text passwords could possibly be intercepted and read if transmitted over untrusted networks. If this username and password combination can in anyway jeopardize your home or business information security please do not submit the download request.

Changing File Associations

The WebGrab grab engine identifies file types by making associations with the dot extension and/or unique character strings contained in file and directory names. A file type refers to what is actually contained in a file and how the operating system and other applications will process the file. WebGrab allows you to customize its file associations to account for changes in Web technologies and customizations as required by the end user.

The "Edit WebGrab file types" dialog is used to change file type and dynamic page associations in the grab engine. To open the dialog select Edit from the main menu and click on the File Types menu item.

The "Edit WebGrab file types" dialog allows you to change any of the following file types and dynamic page associations:

- HTML File Types - Dot extensions used to identify text based Web pages.
- Text File Types - Dot extensions used to identify text file types.
- Dynamic File Types - Dot extensions used to identify dynamic pages.
- URL Query/Execute Strings - Special characters in a hyperlink used by dynamic Web pages for queries and other executable switches.
- URL Dynamic Directory Names – Directory names used to identify pages with dynamic content.

After changing any of the file types and/or dynamic page associations you must click on the Accept button to enter the changes. To cancel the changes click on the Cancel

button. To revert back to the default WebGrab settings click on the Default button.

Logging Events

WebGrab allows you record detailed information about each download by logging events to a disk file. Log files are text files that can be viewed in any text editor or word processor. The information recorded in the log file will also be displayed in the WebGrab status window. **NOTE**: Any information stored in the WebGrab status window will be reset each time the program is started and can also be cleared by selecting Clear Status Window from the View Menu.

Logging is disabled by default and should only be enabled for troubleshooting purposes. Always be aware that log files can grow extremely large if they are not periodically purged. To purge the log file select the Purge Log File option from the Database menu. You can also manually delete a WebGrab log using Windows Explorer, Nautilus, Konqueror, or from the command line. A manual delete can be preformed anytime the program running or when the program is not running. The WebGrab log file will automatically be regenerated if it does not exist.

Clocking a Download

The download speed of every file download is recorded in the WebGrab status window. To open the WebGrab status window select View from the main menu and click on the View Status Window option. The WebGrab Status Window frame will open and entries such as the following will scroll up the window as files are downloaded:

Grabbing: www.glnetsoftware.com/webgrab/index.htm
Downloading 11.64K bytes
Downloaded 11.64K bytes in 2.0 seconds
Download bytes per second rate = 5.97K BytesPS
LAN/WAN connection speed = 47.78K BPS

As the file is downloading the grab engine clocks the download speed. The LAN/WAN connection speed is printed when the download is complete. This value is the number of bits per second, which can be used to measure the speed of the connection path to the Web server.

Windows/UNIX Command Line Utility

The Windows and UNIX command line utility is used to grab Web pages from the command line, within batch files, or script files.

The command line utility is a standalone executable that operates independently of the GUI based WebGrab program. Under Windows the command line utility is installed in the C:\Program Files\webgrab directory. Under Linux the command line utility is installed in the ~/webgrab/bin directory.

To use the command line utility you must enter the name of the executable followed by one of several setup switches and the name of the hyperlink. For a list of command line switches enter the name of the program with no command line arguments:

Windows XP:
C:\"Program Files"\webgrab\webgrab.exe

Linux:
% ~/webgrab/bin/webgrab

```
Usage: webgrab [switches] URL
Example: webgrab -l http://www.xyz.com/index.html

Switches:
-d[dir] = Specify a grab directory.
-h = Grab a document header.
-l = Grab a file plus its links.
-r = Grab a file and recursively follow each link.
-s = Grab a single Web page without links (default).
-L = Disable logging when smart grab feature is enabled.
-S = Disable smart grabbing. Do not check website for newer files.
-X = Follow external links encountered during a grab.
-xs = Grab single external link without links (default).
-xl = Grab external link plus its links.
-xr = Grab external link and recursively follow each link.
-e[num] = Layer of external links to follow (default = 1 layer).
-k = List of file types to skip: -k".zip, .Z, .gz, .tgz"
-K = Text file containing a list of hyperlinks to skip:
     -K"skiplinks.txt"
```

When using the command line utility Websites will be downloaded and stored in the program's working directory. Use the '-d' switch to specify the grab directory used to store downloaded Web objects:

Windows XP:
C:\"Program Files"\webgrab\webgrab.exe -dc:\webgrab www.glnetsoftware.com

Linux:
% ~/webgrab/bin/webgrab -d~/webgrab/database www.glnetsoftware.com

The command line utility can also be used as a troubleshooting tool allowing you to download and analyze HTTP header information returned by a Web server when a page

is requested. Use the '-h' switch to grab and display a page header.

Windows XP:
C:\"Program Files"\webgrab\webgrab.exe -h www.glnetsoftware.com

Linux:
% ~/webgrab/bin/webgrab -h www.glnetsoftware.com

Changing External Application Associations

External applications associations are used by WebGrab to launch applications for specific file types. Under Windows the associations are made in the Windows registry and no changes to the external applications associations are required unless you choose to use another instead of the Windows default application. A change to the WebGrab external application associations will not change any Windows registry information. Under Linux external applications associations are not handled by a central registry and may be dependent on the desktop you are using.

The WebGrab grab engine identifies file types by making associations with the dot extensions and/or unique character strings contained in file and directory names. A file type refers to what is actually contained in a file and how the operating system and other applications will process the file. WebGrab allows you to customize its file associations to account for changes in Web technologies and customizations as required by the end user.

The "Edit external applications" dialog is used to change external application associations. To open the dialog select Edit from the main menu and click on the Applications menu item.

The "Edit external applications" dialog allows you to change any of the following application associations:

- Web Browser - Web broswer used to view hyperlinks.
- Mail Client - Mail client used to send email messages.

- File Manger - File manager used to view files and directories.
- Help Program - Program used to launch help files.
- Help File - Pointer to the program's help file.

The Browse button is used to select the application. NOTE: Under Windows WebGrab expects help files to be in a Windows help file format or .hlp files. Under Linux WebGrab expects help files to be in an HTML format.
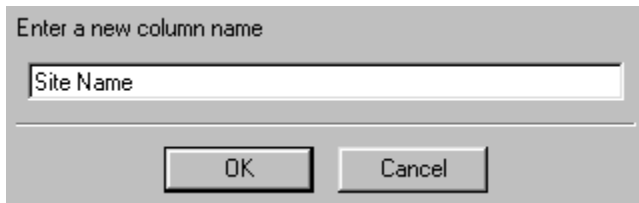
After changing any of the application associations you must click on the Accept button to enter the changes. To cancel the changes click on the Cancel button. To revert back to the default WebGrab settings click on the Default button.

Changing Column Names

Column names are used to label each column in the database. The initial column names are set when a new database is created. By clicking on a column name with your right mouse button you can change a column name:

Site Name	Hyperlink

After right clicking on the column name a dialog box will appear:

Enter a new column name

Site Name

OK   Cancel

Enter the new column name and click on the OK button to commit the changes or the Cancel button to cancel this action.

Turning Off Confirmations Messages

Confirmation messages are dialog boxes used to inform you of certain program settings before executing program operations that may require further attention. When operations are repeated there might be times when you no longer wish to be interrupted by confirmation messages.

To turn off program confirmation messages select Edit from the main menu and click on the Preferences option.

☐ Turn off program confirmation notices

Click inside the Turn off program confirmation notices check box. When this box is checked program confirmation messages will be turned off. Click on the Accept button to enter the changes.

Setting the Maximum Thread Limit

The maximum thread limit is used to set a limit on how many grabs may be active at one time. Each grab requires network bandwidth, CPU time, memory and hard disk space. The maximum number of threads is dependent on the amount of these resources you have available. The default limit is 10 grabs per session but there may be times when you want to increase or decrease this limit. **NOTE**: During external link grabs and list grabs only one thread may be active no matter what the maximum thread limit is set to.

To change the maximum thread limit select Edit from the main menu and click on the Preferences option.

Num Threads	Timeout Value	Cache Segments
10	300	10

Set the number of threads inside the Num Threads box. Click on the Accept button to enter the changes.

Changing the Timeout Value

The timeout value is used to set how many seconds the grab engine will wait for a Web server to respond. The default value of 7200 seconds will allow a Web server 1 hour and twenty minutes to reply before WebGrab disconnects and closes the connection. The repsonse time of a Web server is dependent on the speed of the connection path you have to the Web server. Please note that a connection path may not always be determined by the speed of your Internet connection. A Web server on a slow network may take several minutes to complete a download request even though you have a fast Internet connection. Also be aware that connection speeds may vary many times over the course of a download. To be safe you should not set a timeout value below 300 seconds. **NOTE**: If you lower this number below 300 seconds some Web servers may not have enough time to complete the download request.

To change the timeout value select Edit from the main menu and click on the Preferences option.

Set the timeout value inside the Timeout Value box. Click on the Accept button to enter the changes.

Setting the Number of Cache Segments

The WebGrab grab engine uses cache segments during a download to buffer segments of a file before writing the file to disk. By buffering segments of a file to memory the grab engine can ensure that no data will be lost due to slow disk writes. The access time to your hard drive will vary depending on the operating system and number of processes reading and writing to and from the hard drive. Writing directly from the network to disk may result in data loss and corruption if the operating system does not allow the disk write to complete before the next network read. Generally only a small number of cache segments are require to allow network reads and disk writes to complete without any data loss or corruption.

 The default cache segment size of 10 will allocate 10 cache segments for every file downloaded. When the download is complete the memory used by the cache segments is released back to the operating system. The amount of memory allocated for the number cache segments is dependent on the size of the file being downloaded. The number of cache segments sets a number of potential placeholders that may or may not require system memory. When a file is downloaded a network read will return a small chunk of the file. Most Web servers will return between 1024 and 4096 bytes per network read. The first cache segment is used to hold this chunk in system memory. The next network read will allocate another cache segment. This process continues until the maximum number of cache segments is reached. If the entire file has not been downloaded by the time all the cache segments are used the first cache segment is written to disk and reused. Upon completion of the file download all cache segments are written to disk and all system memory is release.

Changing the number of cache segments is a performance issue relative to the amount of system memory available, the speed at which disk writes occur, and the size of the file being downloaded. The default cache segment size of 10 balances these resources in a manner that is acceptable on variety of hardware platforms. The only time you might need to change the number of cache segments is if you are continuously downloading very large files ranging from 100MB and higher. In these situations a large number of cache segments will consume a large amount of system memory as the file is being downloaded. So be sure that your computer has an adequate amount of system memory to handle a large download request. Calculating the exact amount of system memory required is difficult because the cache segments vary in length according to how many bytes are returned by the Web server following a network read. For a conservative estimate you can multiply 4096 times the number of cache segments. This estimate represents the maximum amount of memory required.
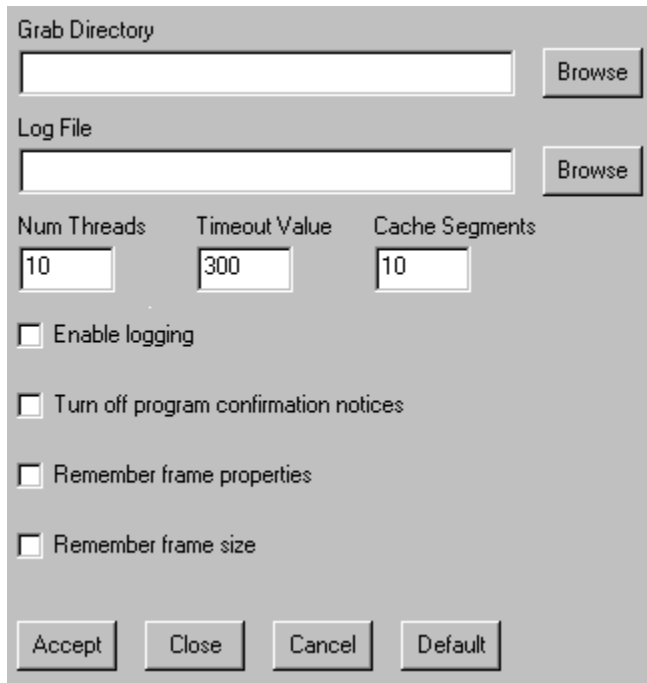
To change the number of cache segments select Edit from the main menu and click on the Preferences option.

Num Threads	Timeout Value	Cache Segments
10	300	10

Set the number of cache segments inside the Cache Segments box. Click on the Accept button to enter the changes.

Changing Your Grab Directory

The Webgrab grab directory is used as the default location for all Web pages, database files, and text exports. To change the default grab directory click on the 🔧 tool bar icon or select the Preferences option from the edit menu. Once selected the Edit user preference panel will appear.

Grab Directory

```
[                              ]  [ Browse ]
```

Log File

```
[                              ]  [ Browse ]
```

Num Threads	Timeout Value	Cache Segments
10	300	10

☐ Enable logging

☐ Turn off program confirmation notices

☐ Remember frame properties

☐ Remember frame size

[Accept] [Close] [Cancel] [Default]

Click on the Browse button to navigate the file system or enter the full path name of the new home directory. After selecting a new home directory click on the Accept button to commit the changes.

Browsing the Grab Directory

When Web objects are downloaded an entry for the first object is inserted in the WebGrab database. However, there may be times when the WebGrab database does not contain enough information about the Web objects. In these situations it is preferable to browse the grab directory where Web objects are stored on disk.

To browse the grab directory without using the WebGrab database select Database from the main menu and click on the Browse DB Directory option. This will open the database directory using the <u>associated file manager</u>.

Adding Entries

To add an entry to the database click on the ✚ tool bar icon using your left mouse button or select the Add option from the database menu. You may also add an entry by pressing **Alt A** on your keyboard.

Once selected the Add Panel will appear:

In the Add Panel enter the site name and hyperlink. **NOTE**: The "Site Name" is the only field required to add an entry. All the other entry fields can be filled in at a later time.

After filling in the appropriate entry fields click on the Redo button to insert the entry. **NOTE**: All entries will be added to the table in alphabetical order.

If you are adding multiple entries with similar fields click on the Redo button to reuse data from previous entries. The Redo function will store up to 255 previous entries.

To navigate the file system for a hyperlink click on the Browse button.

To cancel an add operation click on the Cancel button.

To close the Add Panel click on the Close button.

Deleting Entries
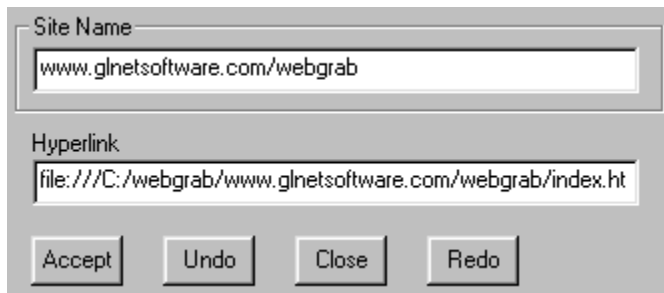
To delete an entry from the database select the item(s) you wish to delete and click on the ⊠ tool bar icon using your left mouse button or select the Delete option from the database menu. You may also delete an entry by pressing **Alt D** on your keyboard.

To select a single item simply left click on the "Site Name" using your mouse.

To select a contiguous list of items press and hold the Shift key and select the items by pressing the Arrow Down key. To select a list of random items press and hold the Ctrl key and left click on the item names using your mouse.

Modifying an Entry

All the entry fields with the exception of the site name can be changed within the database table or by using the change panel. To change an entry within the table, double click on the entry field using your left mouse button. To change the site name or to use the change panel to modify an entry double click on the site name field with your left mouse button.

After modifying the desired entry fields click on the Accept button to commit the changes.

If you are changing multiple entries with similar fields click on the Redo button to reuse data from previous modifications. The Redo function will store up to 255 previous entries.

To cancel a change operation click on the Cancel button.

To close the Change Panel click on the Close button.

Search Operations

To search for an entry in the database click on the tool bar icon using your left mouse button or select the Find option from the database menu. A search operation may also be preformed by pressing **Alt F** on your keyboard.

Once selected the Find Panel will appear:
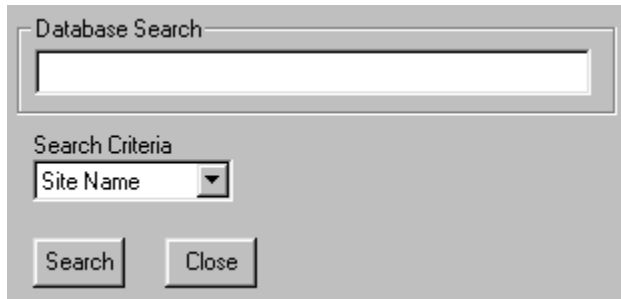
To search for an item by name, type the name of the entry in the "Database Search" box and press Enter or click on the Search button using your mouse.

Click on the "Search Criteria" choice box to search for an item by its site name or hyperlink:

Two types of database searches can be performed, a plain search or a wild card search.

A plain search is not case sensitive, but it requires you to enter the complete string. The plain search will find a single match if searching by item name or will find every match of the complete string if searching on any of the other entry fields.

A wildcard search is performed by entering the wildcard character "*" to the right side of the string. Wildcard searches are not case sensitive, and only require you to enter a few characters of the string. The wild card search will find every occurrence of the string in the specified field.

After a plain search or wildcard search, you will be given the option to display the object(s) that matched the string. If no matches are found, the program will alert you with a message stating that no matches were found. If you elect to display the object found following a plain or wildcard search the display panel will appear:

Use the "Next" and "Prior" panel buttons to navigate through the list in a linear progression.

To modify an entry, click on the "Change" button. **NOTE**: None of the display panel fields are editable. You must always use the "Change" Button to modify an entry.

To delete the current display panel entry click on the "Delete" button. **NOTE**: The program will always prompt you for conformation before performing a delete operation.

To print the list of display panel entries, click on the "Print" button. NOTE: This print function assumes that the appropriate printer and page options have been previously set. See the Printing a Database section for printer and page setup operations.

To export the list of display panel entries to a tab delimited text file, click on the "Export" button. See the Exporting a Database section for detailed information about exporting a database.

To close the display panel click on the "Close" button. **NOTE**: The display panel will not be displayed again until another search is preformed.

Printing a Database

Webgrab includes printing and preview options plus page and printer setup options:

To print all the items in the database select the "Print" option from the print menu. To preview all the items in the database select the "Preview" option from the print menu.

The print preview facility allows you to preview a document before printing it. It is recommended to preview a document after changing any page or printer parameters. To change the document's page settings click on the ▥ tool bar icon or select the "Page Setup" option from the print menu. Once selected the Page Setup panel will appear:
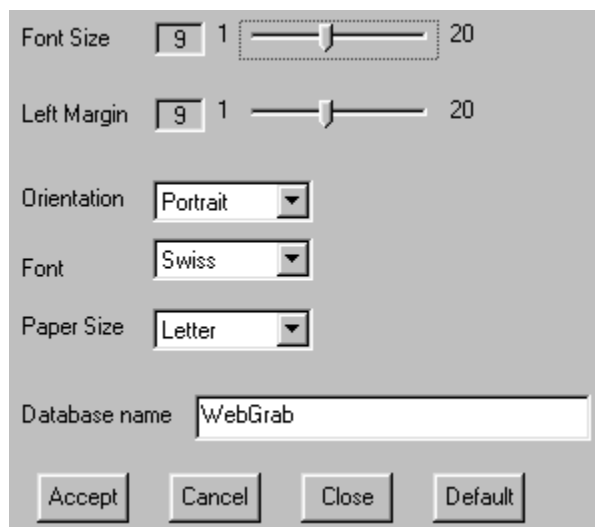
The Page Setup panel allows you to change any of the following page parameters:

- **Font Size** - The font size slider sets the font size from 1(smallest) to 20(largest.)
- **Left Margin** - The left margin slider sets the left margin in points (20 points equal one inch.)
- **Orientation** - The orientation choice box is used to set the page orientation to Landscape or Portrait mode.
- **Font** - The font choice box is used to select a desired font family.
- **Paper Size** - The paper size choice box is used to select a specific paper size.
- **Database Name** - You can change this string to uniquely label this database. The database name will be printed at the top of each page and will also appear on the right hand side of main window's status bar.

After changing or defaulting page parameter settings, you must click on the Accept button to enter the changes. To cancel the changes click on the Cancel button. To revert back to the default printer settings click on the Default button.

Open an Existing Database

To open an existing WebGrab database click on the 📂 tool bar icon or select the Open Database option from the file menu. Once selected a File Selection Box will be displayed.

A list of all the Webgrab database files stored in your Webgrab grab directory will appear. Alternately, you can use the File Selection Box to search for Webgrab database files in other directories or on other drives. **NOTE:** All Webgrab database files will always end with a ".ihd" file extension.

NOTE: A pointer to the last database file opened or created is retained and will be used to load the last database opened each time Webgrab is launched.

Creating a New Database

To create a new Webgrab database click on the ▯ tool bar icon or select the New Database option from the file menu. Once selected a File Selection Box will be displayed.

Enter the name of the file you wish to create in the File name: box. **NOTE:** All Webgrab database files will always end with a ".ihd" file extension. You do not need to include the dot extension when creating a new database. The ".ihd" file extension will automatically be assigned each time you create a new file.

By default all new files will be created in your Webgrab grab directory. You can change the directory path using the Look in: choice box.

NOTE: A pointer to the last database file opened or created is retained and will be used to load the last database opened each time Webgrab is launched.

Renaming an Existing Database

The rename database feature allows you to save the current WebGrab database under a different name making it easier to manage multiple projects. To rename the current database select File from the main menu and click on the Rename Database option. The "Rename the current WebGrab database" dialog will appear. Enter the new name and save the changes by clicking on the Open button.

Importing a Database

The Webgrab import function allows you to import a tab-delimited text file into the database. Database imports are generally used to retrieve data from <u>Webgrab export files</u> or from other spreadsheet and database applications that can generate tab delimited text files. You can manually create import files using the Webgrab <u>Create Template</u> function. **NOTE**: A text based import file must contain at least one string per line.

To import a tab-delimited text file click on the tool bar icon or select the Import option from the file menu. Once selected a File Selection Box will be displayed.

A list of all the text files stored in your Webgrab <u>grab directory</u> will appear. Alternately, you can use the File Selection Box to search for text files in other directories or on other drives. **NOTE:** All import files must end with a ".txt" file extension. After selecting or entering the file name click on the Open button.

The import function parses the text file line by line searching for text strings separated by tabs. If an entry is already in the database, the import function will give you the option to update this entry, skip this entry, or update this entry and all others without prompting. After completion, the import function will report the number of new items that were imported and the number of items that were updated.

Exporting a Database

The Webgrab export function allows you to export the database to a tab-delimited text file suitable for Webgrab imports or other spreadsheet and database applications that will accept tab-delimited text files.

To export the database click on the 🌐 tool bar icon or select the Export option from the file menu. Once selected a File Selection Box will be displayed.

Enter the name of the text file you wish to create in the File name: box. **NOTE:** All Webgrab export files will always end with a ".txt" file extension. You do not need to include the dot extension when exporting the database. The ".txt" file extension will automatically be assigned each time you export the database or the display list.

By default all new export files will be created in your Webgrab grab directory. You can change the directory path using the Look in: choice box.

Creating Import Templates

The Webgrab Create Template function is used to create text template suitable for Webgrab imports.

To create an import template select the Create Template option from the file menu. Once selected a File Selection Box will be displayed.

Enter the name of the text file you wish to create in the File name: box. **NOTE:** All Webgrab template files will always end with a ".txt" file extension. You do not need to include the dot extension when creating a text template. The ".txt" file extension will automatically be assigned each time you use this function.

By default all new template files will be created in your Webgrab grab directory. You can change the directory path using the Look in: choice box.

Merging a Database

The Webgrab merge function allows you to merge the contents of another database with this one. Merges are most frequently used to update the database when other copies of it exist on other systems, such as laptop computers or network drives.

To merge an existing WebGrab database, select the Merge option from the file menu. Once selected a File Selection Box will be displayed.

A list of all the Webgrab database files stored in your Webgrab <u>grab directory</u> will appear. Alternately, you can use the File Selection Box to search for Webgrab database files in other directories or on other drives. **NOTE:** All Webgrab database files will always end with a ".ihd" file extension.

After completion, the merge function will report the total number of new objects that were inserted and the number of objects that were updated.

Backing up a Database

The Webgrab backup function is used to create backup copies of the database. Backup copies of the database should be made regularly in case of disk failures or file corruption.

To create a backup copy of the database, select the Backup option from the file menu. Once selected a File Selection Box will be displayed.

Enter the name of the backup file you wish to create in the File name: box. **NOTE:** All Webgrab database files will always end with a ".ihd" file extension. You do not need to include the dot extension when creating backup copies of the database. The ".ihd" file extension will automatically be assigned each time you perform a backup.

By default all new backup files will be created in your Webgrab grab directory. For archive copies you should always change the directory path to another drive letter, such as A: or a network drive. You can navigate the file system using the Look in: choice box or by entering the destination file in the File name: box.

Troubleshooting

Webgrab has several built-in mechanisms that will automatically repair damage database files. Additional troubleshooting tools are supplied that allow you to gather database statistics and manually repair Webgrab database files. All database transactions and troubleshooting information is logged in the Webgrab status window.

To view your current database statistics select the Database Stats option from the file menu. The current statistics will be displayed in the Webgrab status window.

This statistics function reports the name of the Webgrab data file and index file plus the file status of each file. The file status should always report Open Read/Write unless a read-only copy of the database is opened from a CDROM or other read-only media. The number of index file entries should reflect the exact number of items displayed in the database table. **NOTE**: For an exact count of table items, select the View Row Labels option from the view menu.

If any index file problems are detected select the Compare Index option from the file menu. The results of the test will be displayed in the Webgrab status window:

Comparing the index file to the data file...
The index file checks good.

If the compare index file test fails select the Rebuild Index option from the file menu and note the results displayed in the Webgrab status window:

Rebuilting the index file...
The index file was rebuilt and loaded.
Reloading the database grid...

If you are a registered user and have a damaged or corrupt data file contact gINET software for advanced data recovery procedures.

Using the Toolbar

The Webgrab toolbar is a convenient way to access the most frequently used menu functions. The toolbar is comprised of several different icons, each with a separate meaning:

Click on the New Database icon to create a Webgrab database.

Click on the Open Database icon to open an existing Webgrab Database.

Click on the Print icon to print all the items.

Click on the Preview icon to preview all the items.

Click on the Page Setup icon to modify page parameters before printing or previewing a document.

Click on the Add Object icon to insert a new entry into the database.

Click on the Find Object icon to search for a database entry.

Click on the Delete Object icon to delete an object from the database.

Click on the Import text file icon to import a tab delimited text file.

Click on the Export text file icon to export the database to a tab delimited text file.

Click on the Edit user preferences icon to modify the program configuration.

Click on the Online Help icon to view the on-line help file for this application.

Using the Edit Menu

The Edit Menu is used to perform clipboard operations and set various user preferences:

- **Cut** - Select the "Cut" option to cut a block of text from the WebGrab location box and place it on the Windows or Linux desktop clipboard.
- **Paste** - Select the "Paste" option to paste a block of text from the Windows or Linux desktop clipboard into the WebGrab location box.
- **Preferences** - Select this option to edit your user preferences.
- **Applications** - Select this option to edit the external application associations used to launch a hyperlink, help file, or file manager.
- **File Types** - Select this option to edit the file type associations used to identify Web resources.

Using the View Menu

The view menu allows you to change the program's appearance.

- **View Status Window** - This menu option is used to view the program's status window if it has been closed using the Close Status Windows menu option
- **Close Status Window** - This menu option is used to close the program's status window
- **Clear Status Window** - This menu option is used to clear the program's status window.
- **View Row Labels** - This menu option is used to toggle the database row labels on and off. By default row labels are disabled
- **View Column Labels** - This menu option is used to toggle the database column labels on and off. By default column labels are enabled.
- **View Tool Bar** - This menu option is used to turn the Webgrab toolbar on or off.
- **Move Tool Bar** - This menu option is used to change the orientation of the toolbar.

To retain these program settings from one invocation of the program to the next click on the 🔧 tool bar icon. Once selected the Edit user preferences panel will appear. Check the Remember Frame Properties check box and click on the Accept button to commit the changes. Once the Remember Frame Properties is checked the program setting configured from the view menu will always be retained.

Using the Help Menu

The Webgrab Help menu is used to open online help documents, obtain version information, and open gINET Web resources.

- **Help Topics** - This menu option is used to open the online help document.
- **About Webgrab** - This menu option is used to obtain the Webgrab version number.
- **Contact gINET Software** - This menu option is used to email a message to gINET software using your default mail client.
- **Webgrab Home Page** - This menu option is used to connect you to the Webgrab home page using your default Web browser.
- **WinSock Info** - This menu option is used to view WinSock information. (Windows only)