Thunderstone Webinator
WWW Site Indexer Version 6.1.0

Thunderstone Software

December 22, 2011

Contents

Chapter 1

Document Conventions

Webinator runs on Windows NT, Windows server 2000/2003/2008, and Windows XP. This document refers to all versions of Windows as simply Windows.

Webinator runs on many versions of Unix and Unix-like operating systems. This document refers to all variations as simply Unix.

All filesystem and URL paths are based on the default installation location. `INSTALLDIR` is sometimes used to indicate the directory into which you installed Webinator. The default location for Unix is `/usr/local/morph3`. The default location for Windows is `C:\Program Files\Thunderstone Software\Webinator`.

Examples of command lines and URLs may be broken into multiple lines to fit the printed page. You should not split them when entering them at a command prompt. The split is indicated by ⇝ at the end of the printed line and ↪ at the beginning of the next printed line.

```
http://www.somesite.com/this/a/long/url/with/many/⇝
   ↪subdirectories/that/won't/fit/on/a/line.html
```

If a space is required between the two portions, it is indicated with □.

```
INSTALLDIR/bin/texis profile=PROFILENAME□⇝
   ↪INSTALLDIR/texis/scripts/webinator/dowalk/dispatch.txt
```

Chapter 2

Overview

Webinator is a web walking and indexing package that allows a web site administrator to provide a high quality retrieval interface to collections of HTML and other documents. It is an application of Texis and is written in Texis's Web Script language named Vortex.

It consists primarily of the Texis binary program and two Vortex scripts that are run by the Texis CGI program on your web server and are accessed from a web browser.

One script provides the administrative interface, another provides the site walker and indexer, and the third provides the search function that end users see.

Since these are all scripts, they are easy to modify to provide the look and feel of your site, or to create custom rules for indexing your site.

2.1 Features

Here are some of its features:

- One or more web sites may be indexed into a single database.
- Multiple databases may be maintained.
- It supports cookies.
- There is support for meta data.
- It supports proxy servers.
- Robots.txt and meta robots are respected.
- It provides a totally customizable search interface.
- It provides a totally customizable site walker/indexer.
- A web site may be copied to the local file system.

There are many more features and options to tailor Webinator's behavior to your needs. Almost any option not provided directly by the administrative interface may be achieved by editing the included script(s).

2.2 Obtaining Webinator

Webinator may be obtained from `http://www.thunderstone.com/texis/site/pages/webinator.html`. There you may review the different versions that provide varying size limits and levels of support. Then, you may download the free version or order one of the paid versions.

Follow the instructions on the web site to acquire the package for your operating system. After registering for the free version, you will be given a URL to a compressed tar file for Unix versions or to a setup exe program for the Windows version, and this will contain binaries for your specified operating system.

2.3 Technical Support

Support for Webinator is available via a searchable web message board. It is located at the following URL:

`http://thunderstone.master.com/texis/master/search/msgboard.html`

Anyone may read the discussions. To post a question or comment, you must create an account, which is free, and you must be logged in. Also, once you are signed up, you may "subscribe" to periodic email notifications of new postings to the board. You may select hourly, daily, or weekly notification of new postings.

If you subscribe to periodic notifications, and at some point in the future no longer wish to receive them, you may select "subscribe" again to enter the administrative area where you may delete your subscriptions. Do NOT attempt to get support for free Webinator by any other email or voice channel. Paid users may submit the "Tech Support" form at

`http://www.thunderstone.com/`

Other Webinator resources, such as FAQ, alternate search examples, and such may be found at Webinator's home page `http://www.webinator.com/`.

Chapter 3

Installation

3.1 Unix Download and Installation

For Unix platforms, download the `webinator-6.1.tar.gz` file, from the URL given to you during the registration procedure, to a temporary directory on your machine. (The number `6.1` in the filename may differ, if you are downloading a different version.) Then uncompress it, extract it, and run the install script using the following two commands:

```
gunzip <webinator-6.1.tar.gz | tar xvf -

sh ./install
```

Note: The Webinator install should preferably be run as the user that will actually run the software, not as `root`. The user should be the same one your web server uses to run CGI programs (typically a non-login user); consult your web server config files for details. This user must have permission to place and move files in the install directory and the web server tree. If you must run the install as `root`, it will ask you for the name of a non-`root` user that will be used to run Webinator. **Note:** Once installed, Webinator should **never** be run as `root`.

You will be asked several questions during the installation. For some of these questions, a default answer may appear in square brackets. Eg.:

```
Install dir [ENTER for /usr/local/morph3]:
```

In this case, if you just hit `Enter` without typing a path, the install will use the answer `/usr/local/morph3` as if you'd typed that. **Note:** Just because a default answer is given, does **not** necessarily mean that is the correct or best answer for your particular environment. It is up to you to choose the default or enter your own value based on knowledge of your machine's setup.

You will be asked the following questions:

- **Install directory**

This is the directory where Webinator will place its files and subdirectories. It should be a unique (empty) directory. If it does not exist the install script will ask permission to create it for you. The standard install directory is `/usr/local/morph3`; you should use this if at all possible to avoid potential path issues later. Only enter a different directory if you are specifically unable to install to the standard directory. Whatever directory you choose should be *in*accessible to your web server (ie. outside its server and document directories): the install will place just the public files of Webinator in your web server tree later.

- **CGI directory**
  This is the directory from which your web server runs CGI programs. The install will create a symbolic link to the `texis` executable here. **Note:** Since Webinator runs as a CGI program your web server **must** be configured to run CGI programs. Consult your web server documentation and config files to find out how and where your server places CGI programs. For Apache servers it is typically done with a `ScriptAlias` directive. Note that this is the *file* path to your CGI directory, not the URL entered in a browser.

- **CGI URL prefix**
  This is the URL prefix to the CGI directory you just entered. In other words, it's the URL that you would enter in a browser to access a CGI program in that directory, but without the program name.

  For example, assume you already have a CGI program `findit` installed on this machine, and you access it via the URL `http://www.mysite.com/cgi-bin/findit`. You would enter `/cgi-bin` as your URL prefix. If your site uses virtual hosts, or runs on a non-standard port, you can enter a full URL instead (eg. `http://www.myothersite.com:2001/cgi-bin`).

  If you want to start over with a new CGI directory (previous question), then enter `/newdir` to back up a step.

- **CGI extension**
  This is the filename extension that CGI programs have in the URL. On some web servers, instead of just one directory for CGI programs, any program with a special extension such as `.cgi` at the end signifies a CGI program. If this is true for the CGI URL prefix you've selected, enter it here. For example, if your CGI programs are named `findit.cgi` or `shop.cgi`, then you might enter `.cgi` as the extension. (This may be the case for Apache servers if CGI is set up with an `AddHandler cgi-script` directive instead of `ScriptAlias`.) If your programs do not have an extension in the URL, type `none`.

- **Webinator admin password**
  This is the password for the default Webinator administration account. This password is used to control access to your Webinator walks, so choose a password with care, and ensure that only authorized administrators know it. (Once installed, you can create multiple administration accounts with different passwords if you desire, from the web-based admin interface.) Under some circumstances on some OSes, setting the password from the install may fail. Don't worry: you will be asked to set the password the first time you access the administrative interface.

Once the installation has completed successfully, you can remove the tar and install files, as they are no longer needed:

```
rm -f install webinator-6.1.tar.gz webinator.tar
```

Note: If you move your web server directories around or change your CGI configuration after installing Webinator, you will have to re-install it.

3.2 Windows Download and Installation

The Windows version of Webinator runs on NT 4, Windows 2000/2003/2008, and Windows XP. Download and run the installation program `webinator-6.1.exe` from the URL you were given during the registration procedure. (The number `6.1` in the filename may differ, if you are downloading a different version.)

IIS NOTES:

- A default install of IIS may not include the `scripts` virtual CGI directory. Before proceeding with the install, make sure that the `scripts` virtual directory exists, or that another directory has been created with *Execute Permissions: Scripts and Executables.*
- The URLs to use Webinator will include `/texis.exe`, so if you have installed *URLScan* you will need to allow `.exe` extensions.

During the install you will be prompted for the following choices:

- **Texis ISAPI**
  on the `Select Features` screen, you can choose whether you want to install the IIS ISAPI interface for Webinator. This allows you to use a Unix-style web address for Webinator (no "`.exe`" in the path). See also `Texis ISAPI` (section 6.6, p. 146) for more information.
- **Install directory**
  This is the directory where Webinator will place its files and subdirectories. The directory you choose should be *in*accessible to your web server (ie. outside its server and document directories): the install will place the public files of Webinator in your web server tree later.
- **CGI directory**
  This is the directory from which your web server runs CGI programs. **Note:** Since Webinator runs as a CGI program, your web server **must** be configured to run CGI programs. Consult your web server documentation and config files to find out how and where your server places CGI programs. Under IIS it requires *Execute Permissions: Scripts and Executables* permissions. Note that this is the *file* path to your CGI directory, not the URL entered in a browser. If you are using IIS, the install will attempt to find a suitable directory. A typical default would be `c:\inetpub\scripts`.
- **HTML directory**
  This is the directory from which your web server gets HTML pages, also knows as document root or `DOCUMENT_ROOT`. Consult your web server documentation and configuration to find out how and where your server looks for HTML files. The install will create a directory called `Webinator` and install the publicly visible files, such as the search form and graphics. If you are using IIS, the install will attempt to find this directory automatically. A typical default would be `c:\inetpub\wwwroot`.

- **Webinator admin password**
  This is the password that the default Webinator administration account will have. This password is used to control access to your Webinator walks, so choose a password with care, and ensure that only authorized administrators know it. (Once installed, you can create multiple administration accounts with different passwords if you desire, from the web-based admin interface.)

- If you chose to install Texis ISAPI, then you will see the following:

  – **IIS Version**
    Texis ISAPI needs to know what version of IIS it is working with. Texis ISAPI functionality will be the same on all platforms, but how it operates internally differs greatly. You must choose either `5 or earlier`, or `6 or later`. `5 or earlier` includes IIS 5.1 or any other IIS 5.X versions, and is installed on Windows NT, XP, and 2000 machines. IIS 6 is installed on Windows Server 2003 or later.

  – **ISAPI Destination**
    This is the location that the actual ISAPI program file (`ProxyModule.dll`) will be placed. The default is with the other ISAPI filters in `%SystemDirectory%\inetsrv` (which usually resolves to `C:\windows\system32\inetsrv`).

  – **ISAPI Port**
    This is the port that Texis ISAPI and Webinator will use to talk to each other. Texis ISAPI will be configured to attempt to connect to Webinator with this port, and Webinator will have its Texis Monitor Web Server enabled and listen on the specified port. The default of `10700` should be fine.

3.3 Filesystem Layout

Webinator is installed underneath `/usr/local/morph3` on Unix or `C:\Program Files\Thunderstone Software\Webinator` on Windows by default. It consists of several subdirectories.

This will be the structure on Unix (not all files are listed here):

```
Install Directory
   Readme.txt
   license.key
   conf/texis.ini
   .htaccess
   bin/
      anytotx
      monitor
      texis
   htdocs/
      webinator/
         bar0.gif
         bar1.gif
         common/
            search.css
         factorydefault/
            search.css
         index.html
         xsl/
            default.xsl
   texis/
      default/
         db1/
         db2/
      monitor.log
      scripts/
         errorscript.vs
         webinator/
            dowalk.vs
            search.vs
      testdb/
      vortex.log

HTML Directory
   webinator/
      (same as Install Directory/htdocs/webinator/ tree)

CGI Directory
```

```
texis
```

Note: In versions prior to 6, the configuration file was called `texis.cnf` instead of `conf/texis.ini`. Version 6 will try to load it from the old location if it cannot be found at the new location.

The `webinator` directory contains the search interface scripts, several GIF files used by the search interfaces, and an `index.html` that contains a hyperlink to the administrative interface, as well as the online documentation.

All of the directories that should not be referenced by web browsers contain a `.htaccess` file that denies all access in the event that you chose an install dir under your web server's document root (not recommended). If you did install under your document root and your web server does not respect `.htaccess` style protection you should block web access to those directories by whatever means your web server provides.

This will be the structure on Windows (not all files are listed here):

```
Install Directory
   Readme.txt
   license.key
   conf/texis.ini
   anytotx.exe
   monitor.exe
   texis.exe
   htdocs\
      webinator\
         bar0.gif
         bar1.gif
         common\
            search.css
         factorydefault\
            search.css
         index.htm
         xsl\
            default.xsl
   texis\
      default\
         db1\
         db2\
      monitor.log
      scripts\
         errorscript.vs
         webinator\
            dowalk.vs
            search.vs
      testdb\
      vortex.log

CGI Directory
   texis.exe
```

The `bin` or install directory contains the `texis` program and other related utility programs.

The `texis` directory contains the databases and Texis log files.

3.4 File Permissions and OS Specific Notes

- **Windows**
  IIS will typically run `texis.exe` as the anonymous user `IUSR_machine`. If you want searches to automatically recompile scripts for you, then this user will need write permission on the directories containing the scripts: `texis/scripts/webinator`.

Another option is to test and compile the scripts in a staging area, and when you are satisfied with the results, simply move the compiled `.vtx` file into place.

Texis requires that its monitor process is running. It will attempt to start it if it's not already running. When Texis is running under the web server, there might not be permission available for it to run properly. As administrator, you can register the Texis monitor as a service to run in background and when the system starts up. The install will do this if run as an administrator. You can do this manually from a command prompt when logged in as administrator:

```
monitor -R
```

This will start the monitor service immediately, so there's no need to reboot to activate it.

If you ever wish to unregister the Texis monitor as a service, do this from a command prompt when logged in as administrator:

```
monitor -U
```

- **Unix**
  It is important that texis and its related utility programs always run as the same userid, and that that userid is the owner of the databases. Web servers generally run CGI programs as some user with little or no permission. The installation attempts to get around this problem by making the programs `setuid` to the correct user. If it is not able, you will receive a warning. It is up to you to ensure that `texis` is always run as the same userid.

  The standard Unix commands for making a program `setuid` to some user, `myself` for example, are:

  ```
  chown myself texis
  chmod u+s texis
  ```

  The above commands may only be run by the `root` user on some systems.

3.5 Customizing Webinator's Appearance

You may make common changes to Webinator's search appearance by using `Search Settings` from the administrative interface main menu. You may select color, font, size, result style and order, as well as setting boilerplate HTML to wrap around the search form and results.

But you are not limited to these features. You may change any and all aspects of the search program's appearance and behavior by modifying the supplied `search` script or writing a completely new one.

See

`http://www.thunderstone.com/texis/site/pages/webinator5_scripts.html`

for some examples of custom scripts.

For details on programming with Texis Web Script (Vortex), see the Vortex manual at the Thunderstone web site, `http://www.thunderstone.com/site/vortexman/`.

See also `Customizing the Search` (section 6.4, p. 143) for some insight into the inner workings of the default search script.

Chapter 4

Operation

4.1 Running the Administrative Interface

Webinator's administrative interface is a web application that you access using your web browser. Access it using the URL that was given to you during installation. It will be something like:

- On Unix:
  `http://YOURSERVER/cgi-bin/texis/webinator/dowalk`
- On Windows using CGI:
  `http://YOURSERVER/scripts/texis.exe/webinator/dowalk`
- On Windows using ISAPI:
  `http://YOURSERVER/texis/webinator/dowalk`

Where `YOURSERVER` is the hostname, and possibly the port number, used to access the web server where Webinator is installed.

The `cgi-bin` and `scripts` portions refer to the CGI directory you specified during installation. The examples given above are the most common. Your path could be different.

`texis` and `texis.exe` are the names of the Texis Web Script interpreter and is a program that resides in your CGI directory. It is not a directory.

The portion after `texis`, `/webinator/dowalk`, is a "virtual" path indicating the location of the administrative script relative to your installation's `ScriptRoot` directory. `ScriptRoot` is the `texis/scripts` subdir of your install, so `/webinator/dowalk` in the URL is referring to the file `texis/scripts/webinator/dowalk` under your install dir. This is the administrative script that controls Webinator.

When you run the administrative interface you will be asked for the login and password. By default there is one login name. It is `webinator` in all lowercase. If no other accounts have been added, you will not have to enter the name. It will be filled in for you. Your login will be remembered in a cookie until you logout. This way, you don't need to enter the password every time you enter.

Note: If you share your computer with others, or it is available to people who should not be administering Webinator, then you should logout when you are finished. This will help prevent unauthorized configuration.

The Webinator administrative interface uses JavaScript to enhance its functionality and make it easy to use, but the interface will also work well without JavaScript. No functionality of Webinator will be lost if JavaScript is turned off in your browser (eg. to prevent pop-ups on other sites). In this document, the user interface description assumes that JavaScript is enabled.

4.2 First Time Run: Quick Start

Step 1: Create an Account

During installation you were asked for a password for the default administration account (`webinator`), which you should now enter at the prompt. If for some reason this step did not happen, the first time you run the administrative interface you will be asked to create and enter a password. You should choose a password that is easy for you to remember but hard for someone else to guess, as this is an account that will control administrative access to Webinator (additional accounts may be created later as needed). You will need to enter the same password twice (two input boxes will be provided) to help check for typing mistakes. Passwords are case sensitive. Once the password is created and `Change` is pressed, you will automatically be logged in and taken to the `Profiles` page to create a profile.

Step 2: Create a Profile

A *profile* is a collection of data (URLs/documents) to be searched, plus the settings that control that search; a profile must be created and walked before searches can occur.

On the `Profiles` page, a default profile name and data directory will be filled in for you to create. You may change either of these if desired, then hit the `Create Profile` button.

A new profile will be created but a site walk/index will not be started yet. You are then presented with the main walk settings page. The `Base URL` will be automatically filled in with the name (or IP address) of your web server. If you wish to walk a different site you may change the `Base URL` at this point.

Step 3: Walk the Profile

Once you're satisfied with the URL and extension settings, you may hit the `GO` or `Update and GO` button to begin a walk of your site. A walk will be started in the background and you will be taken to the `Walk Status` page. This page will show you the status of the walk in progress and indicate when the walk is complete. This page will automatically refresh every 10 seconds with the latest progress information until the walk is complete. When the walk is complete you will see a summary of errors.

Last Step: Search

Once the walk is complete, you may click `Live Search` on the menu at the top of the page. This will take you to the search that users will use. It is also the URL you can place on your web page(s) to send users to the search.

You now have a site index that you can use. There are many options to control the site walk as well as the search interface appearance. They are described in detail elsewhere in this manual. Use the `All Walk Settings` button on the administration script's menu to see all of the options. Click the question mark (`?`) next to an item to get help for that item.

Since the walker, administrative interface, and search are all scripts with source code provided, you are not limited to the settings available in the administrative interface. Any or all of the scripts may be modified to take on new behaviors.

4.3 Administrative Interface Overview

Webinator's administrative menu has the structure given below. Each item is described on the pages that follow.

```
Entry
   Basic Walk Settings
      Update
      GO, Update and GO
      STOP
   All Walk Settings
      Update
      GO, Update and GO
      STOP
   Search Settings
      Update
   Profile Tools
      List/Edit URLs
      List Duplicates
      SOAP Tools
      Test Fetch
      Best Bet Groups
   Walk Status
      Refresh
      STOP Walk
   Query Log
   Test Search
   Live Search
   Profiles
      Create Profile
      Select a Profile
      Delete a Profile
   Accounts
      Add a User
      Change Password
      Delete
   User Groups
   Access Control
   Maintenance
   Documentation
   Webinator Home
   Logout
```

4.3.1 Entry

Upon entry to Webinator's administration interface you are prompted for user name and password. If you have logged in previously and still have the cookie and have not logged out, the login page is bypassed and you are taken directly to `Profiles` (see section 4.3.10, p. 24).

Your login is remembered in a cookie until you logout. This way you don't need to enter the password every time you enter. If you share your computer or it is otherwise available to people who should not be administering the Webinator, you should logout when you are finished.

4.3.2 Basic Walk Settings

This is the central area for configuring a walk. The most commonly used walk related options and their settings are presented and they may be changed here. The Basic Walk Settings are a subset of the All Walk Settings. Next to each option is a question mark (`?`) which, if clicked, takes you to help for that option. The options are documented individually later in this manual in section 4.4.

At the bottom of the page is a set of three buttons. Pressing any of the buttons affects all options on the entire page.

- `Update`
  This button causes all changes on the form to be saved. No walk is started.

  If the **Rewalk Schedule** has been changed, the new schedule will go into effect immediately.

  If **Categories** have been changed, the walk database will be updated to reflect the new categories. The search interface will reflect the new categories.

  If **Single Page**, **Page File**, or **Page URL** has been changed, the listed individual pages will be fetched into the live search database and made available for searching.

  If the **Word Definition** or **Text Search Mode** is changed, the search index on the live database will be dropped and recreated. Searches might not work while the index is being rebuilt.

- `GO` or `Update and GO`
  The `GO` button will change to `Update and GO` after you make a change to any setting on the form. The ultimate behavior for either is the same.

  The current settings from the form will be saved as is done when you click `Update`. Then a new walk will be started. The new walk will be performed to either a temporary database or the live database, depending on the setting of Rewalk Type (Section 4.4.14). Then you will be shown the walk status page where you may monitor the progress of the walk.

  Changes to **Categories** or **Word Definition** will not be reflected until the walk finishes.

- `STOP`
  When a walk is in progress the `GO` button is replaced by the `STOP` button. This button terminates the running walk and abandon the work that it has done so far.

- `Reset`
  This button reverts all settings on the page to what they were when the page was first loaded.

4.3.3 All Walk Settings

This is the central area for configuring a walk. This is similar to `Basic Walk Settings` except that all walk related options and their settings are enumerated and may be changed here.

4.3.4 Search Settings

This page contains all of the settings related to the search interface that end users see when performing searches.

All search options and their settings are enumerated and may be changed here. Next to each option is a question mark (`?`) which, if clicked, opens help for that option. The options are documented individually later in this manual in section 4.6.

At the bottom of the page is a set of buttons. Pressing any of the buttons affects all options on the entire page.

- `Update Test`
  This button causes all changes on the form to be saved in the set of test settings, which can be tested via the `Test Search` link on the left side of the interface. It does **not** modify the `Live Search` settings.

  This allows you to "try out" settings before applying the changes to your live search users' interface.
- `Update Live and Test`
  This button updates both the `Live Search` and `Test Search` settings. Use this either after testing out the settings via `Update Test`, or for small changes that you don't feel the need to test out and immediately want to make live.
- `Copy Live to Test`
  If you try out changes via `Test Search` and you decide you don't want them, you can use `Copy Live to Test` to discard the test changes you've made and revert back to the current `Live Search` settings.
- `Reset`
  This button reverts all settings on the page to what they were when the page was first loaded.

4.3.5 Profile Tools

The Profile Tools contain multiple tools for working with your profile.

List/Edit URLs

On this page, you may list or delete all or selected URLs from the database. You should always list before you delete, so you know that you are deleting the correct ones. While listing URLs, you may display all known information about a given page. You may also create categories for selected sets of URLs from this interface.

If a walk is in progress, delete is disabled and you are given the choice of listing URLs from the live search database or the new database being built by the walk.

Select `List` or `Delete` from the drop down list. The default is always `List` for safety.

In the pattern box, enter the URL or pattern for URLs for which you want information. This may be an exact URL or a wildcard pattern, which lists all URLs matching the wildcard pattern. For a wildcard pattern, use asterisk (`*`) to match anything and question mark (`?`) to match any single character. You may enter up to 10 different URLs or patterns in the box to find them all at once. Put a space between patterns when entering multiples. Leaving the pattern box blank implies `*`, and this will cause every URL in the database to be listed. Deletion will be denied if the pattern is blank or `*`.

Select the order in which you wish to see the list:

Depth	URLs encountered first in the walk will be listed first
URL	URLs are ordered alphabetically
Newest first	URLs are ordered by modification date with newest ones first
Oldest first	URLs are ordered by modification date with oldest ones first
Largest first	URLs are ordered by download size with largest ones first
Smallest first	URLs are ordered by download size with smallest ones first

Then `Submit`.

All matching URLs will be listed. Clicking on a listed URL opens a page of details about that URL. On that detail page, everything the database knows about that URL is presented. You can also see what pages refer to the selected page by clicking `Parents` and what pages the selected page refers to by clicking `Children`. The `test` link next to the URL can be used to do a live test fetch of the page to find out how Webinator processes it. See `Maintenance->Test Network and Servers` 4.12.3.

If your pattern matches less than the entire database, you will be given a form from which you can create a category using the same pattern(s). Simply enter the name of the category to create and click `Submit`. The name is the name that users will see on the search form. This new category will also appear on the main settings page along with the other categories. It will also be immediately available to search users.

If the profile is a meta search, then the profile has no URLs of its own to list. The List/Edit URLs page will instead display links to the list/edit URL pages for each of its target profiles.

Live Search and New Walking Database

These options are presented on the `List/Edit URLs` page (see 4.3.5) if a walk is active. They allow you to choose which database to query. The “Live” database is the one from a previous successful walk that is what search users see. The “New” database is the database currently being built by the new walk. It is not visible to search users.

List Duplicates

This section allows you to list all the duplicates of a given page. The URL entered may be the URL that was kept in the walk, or any of the pages that were excluded as a duplicate of pages already in the walk.

If `Keep Refs` was used in the walk, then all the pages that linked to the duplicate pages will also be listed.

Test Fetch

Test how Webinator will handle specific URLs. See `Maintenance -> Test Network and Servers`, p. 94.

Best Bet Groups

The Best Bets are grouped together. This allows different groups to be shown in different places, and easily rotated in or out. For example, you might have one group of links that you have determined to be the most probable results for a user's query, and another group that includes links you want to promote.

The Group Name is how the group will be identified elsewhere in the administrative interface. This should be chosen to readily remind you of the purpose behind the group.

The Result Type indicates which fields will be shown on the results page. The title and description are entered by the administrator, rather than always being taken from the page.

4.3.6 Walk Status

This page shows the status of the latest walk for the current profile. If a walk is in progress, it is the one reported.

During an active walk, it indicates a summary of how many pages are to be walked in the next hour, how many were walked in the last hour, and the total number of pages. There is a list of the most-recent URLs fetched, with number of errors and duplicates found, followed by a list of the next URLs to be walked. Below that is summary information about the walk itself, including walk start time, starting URLs, and some profile settings. The Walk Status page updates automatically every 10 seconds until the walk is complete or another page is selected. (After 10 minutes of user inactivity it will refresh once a minute to save traffic.)

When no walk is in progress, the report also includes a list of errors and duplicates encountered. If the last walk was abandoned, the report includes information about how far it went, as well as the report from the last complete walk.

Now button

During the walk the **Refresh display:** `Now` button may be selected to force a Walk Status display refresh before the 10 second automatic refresh. Note that this only affects the display, not the walk itself.

Pause/Auto button

The **Refresh display:** `Pause` button pauses the Walk Status display (prevent the browser from refreshing the display every 10 seconds): this changes the button to `Auto` which will have the opposite effect (resume the auto-refresh). This is useful when examining the status page in detail, and avoiding being interrupted by the browser auto-refresh. Note that both buttons only affect the display, not the walk itself.

STOP walk button

The **Current run:** `STOP walk` button on the Walk Status page stops the current walk. If the walk type is `New`, the walk will be abandoned (current live search is left intact and not updated). If the walk type is `Refresh`, the new pages are always live (since refresh uses one database), but the search indexes are not updated.

Pause walk and Make live button

The **Current run:** `Pause walk and Make live` button pauses the current walk, updates its search indexes for speed, and makes the walk live (ie. deletes the current live database and replaces it with the current walk). This can be useful if you ran out of disk space while indexing and subsequently freed up some space, or if a long running walk was stopped and you want to use the incomplete walk. If the walk was abandoned due to an error, make sure you resolve the problem before trying to make the new database live.

4.3.7 Query Log

The query log pages provide detailed and summary information about queries. Query logging must be turned on to generate information on the query log pages. If query logging has never been turned on for the current profile, there will be nothing to see. The query log is erased each time the database is rewalked.

The pages are as follows:

- Query Report
- Top Query Words
- Top Queries
- No Hits
- Best Bet Clicks

The query log lists the time that each search occurred, the IP address of the web user performing the search, the number of hits for the search, and the user's query. For URL clickovers, it displays the query instead of the number of hits and the actual URL instead of the query.

Selecting the Date/Time for a listed query will display a page with complete information about the search. This page includes everything from the summary list, and any non-default parameter settings from the search. A hyperlink is provided so that you may perform the same query as the user.

4.3.8 Test Search

This hyperlink opens the search interface. It forces the interface to use the search settings listed on the `Search Settings` page, whether they have been applied or not. This allows you to test search settings without affecting end users until you are satisfied with the new settings.

This mode also places two extra hyperlinks at the top of the search pages. `Back to Administration` allows you to return to the Webinator administration interface. `Make this appearance live` does that too, but it additionally makes the search settings you are testing "live", so that end users also see the search setting effects.

4.3.9 Live Search

This hyperlink opens the Webinator search interface as end users see it.

4.3.10 Profiles

This page presents a list of existing profiles. A profile contains the walk and search settings for a collection of pages. The profiles are listed in the order of creation by default; clicking on `Name` will re-order by profile name. You can click on a profile's name to see and/or change its settings and status or to start a walk.

You can click on `Delete` next to a profile to delete that profile. You will be asked whether you really want to delete the profile or not.

When a profile is deleted, all of its settings are lost and any walk database it has created is deleted. There is no way to get back any of these items after the profile is deleted. **Note:** Under Windows it is possible that the walk database will not be completely deleted if there are currently searches being performed on the database. You should not delete a database that is being actively searched. If you do this, you will need to delete the remnants of the database by hand.

You may also create a new profile by entering a new name and data directory. You may not use a data directory that is in use by another profile. You generally specify a new data directory. The directory will be created if it does not already exist.

You can copy settings from an existing profile to your new profile by selecting its name from the drop down list. This allows you to set up another site similar to an existing one. It allows you to experiment with the walk settings for an existing site, without potentially harming the good walk that is being searched by your users.

Here are notes about the import process and differences between versions 2 and 4.

- **-[no]unique**: Databases are unique by default ("Prevent Duplicates" 4.5.61).
- **-j**: It is automatically implied ("Stay Under" 4.5.60).
- **-k**: The default expressions are broader ("Word Definition" 4.5.46). The import will replace the defaults with what is in your old profile.
- **-n**: All known plugins are predefined in `dowalk` function `doplugin`. You may need to add extensions to the "Allow Extensions" list (p. 29) and/or MIME types to the "Mime Types" list (p. 59) though. Import will do this for you.
- **-r**: Robots META tags are also supported. Import will apply your old setting to both robots.txt and meta robots.

- **-b**: Permanently on (walks are always breadth first).
- **-L**: Permanently on (virtual hosting demands it).
- **-l**: Not changeable. Log files contain different information.
- **-q**: Quit time is not supported.
- **-c, -A**: Site copying is not supported.
- **-[no]dnscache**: DNS caching is not supported.

4.3.11 Accounts

This section provides information to maintain multiple login accounts for access to Webinator administration. All users are listed on this page. You may add users, delete users, and change individual user passwords. The default user, called `webinator`, may not be deleted.

The Accounts page also allows you to create multiple administrative users. There is no distinction among them after they are created. All users have full administrative permissions, and they may create and delete any user or change any user's password. This is a basic security mechanism meant to keep unauthorized persons from using the web based administrative interface. The purpose of supporting multiple administrative users is that you can create distinct passwords, which you can revoke in the future without needing to change a single global password that all administrators know.

User names and passwords are stored in the `SYSUSERS` table of the default database. This is only a holding place for them. No Texis permissions are granted or revoked for these users. A benefit of storing the users in `SYSUSERS` is that any users that you might create in the default database by other means than the Webinator interface will also automatically become Webinator administrators.

The passwords are one-way (forward) encrypted. This means that a forgotten password may not be discovered. The only way to deal with a forgotten password is to change the password. In the event that all passwords are forgotten you can delete the `webinator` user from SYSUSERS using `texis -s` from a command prompt, and then enter an appropriate SQL delete statement. The administrative script will then create the `webinator` user anew and ask you for a new password.

Add a User

To add an administrative user, enter the new user's login name and password. You will need to enter the new password a second time into the `Confirm` box to protect against typing mistakes (since you can't see the password you are typing).

Names and passwords are case sensitive. "Joe" is different than "joe". You should choose passwords that are easy to remember, but difficult for someone else to guess.

Change Password

Here you may change the password for the selected user. You will need to enter the new password twice to protect against typing mistakes (since you can't see the password you are typing). Enter the password once the `Password` box and again into the `Confirm` box

Passwords are case sensitive. "Joe" is different than "joe". You should choose passwords that are easy to remember, but difficult for someone else to guess.

Delete

This will delete the selected user. You will be prompted to confirm whether the user should really be deleted or not. Once a user is deleted, there is no way to get it back except to re-add it.

The default user, "`webinator`", may not be deleted.

4.3.12 User Groups

User groups may be created on this page, by clicking the `Add a Group` link. Existing groups may be edited or deleted with the appropriate links. User groups are used to associate administrative users into similar-privilege groups for easier access control maintenance. See the User Groups section for more details (p. 85).

User groups are supported in the full Texis product, but not Webinator-only.

4.3.13 Access Control

The Access Control page allows configuration of administrative users' access to administrative actions (creating profiles, starting walks etc.). In conjunction with user groups, access control can be used to restrict certain users to only certain actions, instead of allowing all users access to all administrative functions. See the Access Control section for more details (p. 85).

Access Control is supported in the full Texis product, but not Webinator-only.

4.3.14 Maintenance

The Maintenance page contains various links for maintaining and editing operating-system and overall settings.

See also `Maintenance` 4.12.

4.3.15 Documentation

This provides a hyperlink to the online version of this document.
`http://www.thunderstone.com/site/webinator5man/`

4.3.16 Webinator Home

This provides a hyperlink to the online home of Webinator.

`http://www.thunderstone.com/texis/site/pages/webinator.html`

4.3.17 Logout

This will log you out of the administrative interface and clear your login cookie. It then takes you back to the login page.

4.4 Basic Walk Settings

This page contains the settings that are used most commonly. They are available in `Basic Walk Settings`.

The settings on the `Basic Walk Settings` page are a subset of the settings on the `All Settings` page. Use the page that is most convenient for your current task.

4.4.1 Database

Syntax: the full path to the database directory on the server's disk

This indicates what database is being used by the currently selected profile. The database is only settable when creating a profile. A new profile must be created to use a new database.

4.4.2 Walk Summary

This is informational only. It contains summary information about the most recent walk and recategorizations. The information includes the date and time of the walk, whether the walk was successful, how many pages were indexed, and the number of duplicate pages.

4.4.3 Notes

This is a scratch pad area for the administrator of the profile. It in no way affects the walk or search.

4.4.4 Base URL

Syntax: one or more URLs, one per line

This is the address where the web crawler will start walking your site. If the whole site is to be searched, simply enter your web address, for example "`http://www.mysite.com`". If the search is to be limited, specify the address to start the search or create a page listing the URLs to search. The search will only

return information from your web site - no off-site searching will be done. Directory URLs should include a final forward slash "/". Example - "http://www.somehost.com/mysite/". If you have a virtual domain that just redirects to another URL, enter the destination URL as your Base URL instead of your virtual domain name.

You may specify multiple base URLs to index multiple sites; Webinator's idea of a "site" is a single host as identified by the hostname portion of a URL. Therefore http://www.mysite.com, http://www2.mysite.com, and http://mysite.com would all be considered different sites.

In version 4.02.1046373961 Feb 27 2003 and later, the special "protocol" http-post or https-post may be used for a Base URL. This uses the POST method instead of the GET method to fetch the URL, using the query string as POST data (it must be URL-encoded). This can be used to start walking at a login page form that requires POST instead of GET. Note that the URL stored in the html table will have the -post and query string removed for security. During a Refresh walk, when a URL is about to be refreshed, the probable Base URL that led to it (ie. the one with the longest prefix) will also be fetched. This helps ensure that login cookies are properly restored to allow Webinator access during the refresh. Example:

"http-post://www.somehost.com/login.asp?user=bigbird&pass=open-sesame"

In version 5, a username and password may be given in the Base URL. Normally, if only one login is required to access the site to be walked, the username and password should be given in the Login Info walk setting. However, if several different logins are required, the additional logins can be specified as user:password@ prefixed to the hostname in the Base URL. Note that the user/pass is for WWW Basic Authentication. If your site uses a custom or form-based login, use http-post instead. Example:

"http://MyName:MyPassword@www.myhost.com/login.asp"

See also URL file 4.5.6, URL URL 4.5.7, Single page 4.5.8, Page file 4.5.9, and Page URL 4.5.10 for more ways to specify URLs.

4.4.5 Enterprise

Syntax: a single domain name

The name of your company's domain. This is useful if your company's web presence consists of multiple hosts within its domain, and you want them all indexed together as a unit.

This allows you to walk any URLs encountered during the walk of the base site(s) that are within the given domain. Webinator will attempt to guess this value for you, but you may set it to whatever you wish. Check the Yes box to enable this feature.

See also Extra domains 4.5.13 which is the same but allows more than one domain. These options may be used together.

4.4.6 Robots

Syntax: select Yes or No buttons

robots.txt

With this set to Yes, Webinator will initially get `/robots.txt` from any site being indexed and respect its settings for what prefixes to ignore. Ignoring robots.txt is not generally recommended.

See also `Robots.txt` 5.4.

Meta

Respect the meta tag called `robots`. With this set to Yes Webinator will process and respect the robot control information within each retrieved HTML page.

See also `Robots.txt` 5.4.

4.4.7 Allow Extensions

Syntax: one or more file extensions separated by space

A list of the URL extensions that the crawler will accept. The default list is empty, i.e. all extensions are allowed.

To search MS-Word documents, use `.doc`. For Shockwave/Flash use `.swf`. For WordPerfect documents specify whatever extension you use and ensure that the web server returns the MIME type `application/wordperfect` as there is no consistent extension for WordPerfect documents. Any extensions not listed here will not be searched or walked.

A few other extensions you may find useful are

```
.asp
.cfm
.jsp
.shtml
.jhtml
.phtml
```

4.4.8 All Extensions

Syntax: select Yes or No button

Retrieve all files instead of only those listed in `Allow Extensions`. This turns off checking of URL extensions. All URLs will be retrieved regardless of the extension (including images and such files).

4.4.9 Exclude Extensions

Syntax: list of extensions

A list of URL extensions that the crawler will reject. The default is empty, i.e. no extensions will be rejected.

4.4.10 Exclusions

Syntax: zero or more strings, each on a separate line

Excludes URLs containing any of the specified literal strings anywhere in the URL (hostname, path, or query).

See also `Exclusion REX` 4.5.16 and `Exclusion prefix` 4.5.17 for more ways to exclude URLs.

4.4.11 Crawl Delay

Syntax: a decimal number from 0 to 10

Causes Webinator to wait the specified number of seconds between page fetches. Normally set this to 0, and Webinator will fetch and process pages as quickly as it can. Increase the Crawl Delay if the web server cannot handle being hit rapidly. Increasing this value forces the walk to take at least the following number of seconds to complete: the Crawl Delay number times the number of pages on the site.

Decimal numbers may be specified - `0.1` will cause it to walk no more than 10 pages per second, etc.

Note: Using a delay larger than 0 forces `Threads`(4.4.12) to 1. A delay defeats the advantage of multiple threads and large delays could cause unexpected page fetch timeouts.

4.4.12 Parallelism

Syntax: whole numbers from 1 up

Threads

This is the maximum number of simultaneous page fetching threads to allow against each site. Setting Threads higher than 5 is probably not very helpful, unless you have many “Single Pages” that are on various hosts.

Servers

This is the maximum number of different web servers to walk simultaneously. Setting this too high can stress your memory, cpu, and network.

4.4.13 Verbosity

Syntax: whole number from 0 through 4

Sets how much information the walker should provide about what it’s doing. The default verbosity level is 2. The values are described in the following table.

The levels are cumulative. In other words, each level includes the previous levels.

Warning: at Verbosity 4, full Primer URLs will be printed to the Walk Status Log. If you use Primer URLs that contain credentials that you don’t want other Webinator administrators to see, you will need to restrict access to the Walk Status, in addition to the Primer URL, when using Verbosity 4.

Table 4.1: Verbosity Levels

Level	Description
0	Issue no messages except errors
1	Display starting point URLs
2	Display selected setting info
3	List URLs found in URL files
4	Indicate why URLs are rejected

4.4.14 Rewalk Type

Syntax: select from drop down box

This determines how rewalks are performed.

New

The type `New` creates a new database and does a complete walk of everything, starting with the Base URLs. A `New` walk does not disturb the existing database.

Refresh

The default rewalk type `Refresh` updates the existing database, and only downloads files that have been modified or created since the last walk. Pages that are no longer present on the server are removed from the database.

Here are other considerations for using `Refresh`. Pages that were referenced but were missing in the initial walk (the walk prior to the `Refresh`), but were added after the initial walk, will be missed by `Refresh` if their parent page has not been modified. If you change your settings to be more inclusive (ie add extensions, ignore robots, add domains, etc.), you should do a `New` walk once, because a `Refresh` is not likely to find the newly allowed data, unless all of the pages leading to this data have been modified.

If more than 30%-50% of your site changes between walks you may be better off using a `New` walk instead of `Refresh`. Also, many dynamic content generators do not give modified dates which will cause every page to be rewalked. In that case you should use `New` instead of `Refresh`.

Refresh in version 5 vs. 4

In Webinator version 4 and earlier, the refresh walk checked every page in the database to determine whether it needed updating. Since only changed pages need updating, and those are typically a small percentage of the site, checking for changed pages is faster than doing a complete new walk. However, it is still time-consuming, because the web server must be accessed for every page on the site, and only the web server can inform Webinator whether the page has changed.

In Webinator version 5 and later, there is an improved refresh process. The walk is adapted to focus on the

small but important group of changing pages. As each page is walked, a refresh period is calculated for that individual page. The calculation is based on whether the page has changed since the last time it was fetched, and how long ago that fetch was. This refresh information is used to determine when the page should be checked again. In this way, the walk prioritizes the walking of pages that change often or are new, and it delays the fetch of pages that seldom change.

Thus, when a walk (scheduled or manual) takes place, only the pages that need to be refreshed now are actually fetched – not the entire database. The result is a database that is updated by a process that consumes fewer server resources.

Rewalk Type Summary Table

The following table summarizes the trade-offs for the new and refresh rewalk types.

Method	Advantages	Disadvantages
`New`	Guarantees most accurate representation of current site. Does not disturb live search database.	Uses more bandwidth and temporary disk space. Longer time before site changes are reflected in live search.
`Refresh`	Faster. Uses less bandwidth and temporary disk space. Site changes are reflected in live search much sooner.	Could get out of sync with actual site under rare circumstances. A lot of changed pages could substantially slow searches during the walk. Requires `If-Modified-Since` support on walked web server.

4.4.15 Rewalk Schedule

Syntax: select from drop down boxes

This performs a rewalk on the schedule specified. The rewalk action is the same as the one that can be started manually by clicking the `GO` button. The `Frequency` defines how often to automatically rewalk. The `Hour` defines which hour to start the rewalk for daily or weekly runs.

You can define multiple walk schedules for the same profile by clicking the `Add More Schedules` link. This gives you more granular control in setting schedules. For example, instead of choosing between `once a day` and `once an hour`, you can have a crawl launch 3 times a day by making the 3 schedules

- `Daily at 8AM`
- `Daily at 12PM`
- `Daily at 4PM`

To remove a schedule, set its `Frequency` to `-None-`.

See also `Notify` 4.5.2. If you are using "`On Change`" see also `Watch URL` 4.5.1.

4.4.16 Action Buttons

These buttons tell Webinator to do something now. Only the buttons applicable to the current status are displayed. The buttons are as follows:

- `Update`: Save the current settings for future use but don't begin a walk.
- `GO`: Begin a walk using the current settings.
- `Update and GO`: Save the current settings then begin a walk using those settings.
- `STOP`: Stop and abandon the walk that is currently running.

See the Walk Settings section (4.3.2) for details about the operation of these buttons.

4.5 Advanced Walk Settings

These are the advanced settings that are used less commonly than the settings available in `Basic Settings`. The advanced settings are available in `All Walk Settings`. You are not limited to the features listed here. You may modify the `dowalk` script to create additional features and to make the walk behave however you want it to behave.

See also `Customizing the Walker` 6.5 for information about the inner workings of the `dowalk` script.

4.5.1 Watch URL

Syntax: an HTTP URL

The URL specified here will be refreshed every time that Webinator starts a refresh walk. This can be used if you have a page that lists new documents that are added to the site as it will ensure that the links are found as soon as possible.

4.5.2 Notify

Syntax: an email address

If this is set, a summary report will be sent to the supplied email address when a scheduled rewalk occurs.

4.5.3 Attach Logs

This selects the log files to attach to the walk notification. The log files and walk errors are for the period of the refresh walk, and are sent as tab separated files that can be opened with programs such as Excel for further processing.

If the query log is attached it will be cleared after being emailed. This is an alternative to separate query log rotation and emailing and is particularly useful when using mode new for rewalks and you don't want to lose the query log. The query log is compressed for delivery with "zip" if present. If you want to use another program or your zip executable is not where dowalk expects you can modify dowalk and set `$zipexe` to the full path of your zip program. If your program uses different command line options than zip you'll also need to adjust the `<exec>` lines where `$zipexe` is used to accomodate your program. If `$zipexe` doesn't exist the log will be emailed uncompressed so not having zip won't preclude receiving the logs, though they may be large and be rejected by some email systems due to size. See also `Rotate Schedule` (section 4.6.3).

4.5.4 Categories

Syntax: textual name and URL pattern pairs, additional input boxes will appear as you fill the ones provided

Webinator can create searchable sub-categories that will appear in a drop down box on the Search page. Enter the name of the category on the left, and its corresponding URL pattern on the right. URL patterns may contain asterisk(`*`) to indicate "anything" and question mark(`?`) to indicate any single character. There may be more than one pattern for each category. Separate multiple patterns with space. The following table provides an example.

Table 4.2: Example Categories

Category	URL Pattern
`Demonstrations`	`http://SERVER/demos/*`
`Manuals`	`http://SERVER/manual/*`
`Books`	`http://SERVER/a1/* http://SERVER/b3/*`

This example would create a category named "`Demonstrations`" which would only search the URL "`http://www.mysite.com/demos/`" and any files under this directory, thereby creating a more concise match to the user's search. The same is true for "`Manuals`". However, the "`Books`" category would include pages from both the "`a1`" and "`b3`" directories. The user would now have the option to search within just these categories or the entire database. The pattern should *not* be a single page unless you want a category with a single page in it (e.g. "`http://www.mysite.com/manual/index.html`" would be incorrect). It should typically be a prefix for a directory that has multiple pages within it followed by an asterisk (`*`).

For best search performance, categories that overlap one another – ie. contain pages in common – should be listed most-commonly-searched first.

4.5.5 Categories Type

Syntax: radio button choice

The **Categories Type** setting sets what type of categories are being used, and how to optimize category searches. It set to one of:

- `Auto-detect`

Automatically detect what kind of category is being used on a per-category basis, and optimize searches accordingly. This lets non-overlapping categories (ie. those whose pages do not occur in any other category) be searched fastest, while still supporting overlapping categories as fast as possible. This is the default mode.

- `Overlapping`
  Assume that any category might overlap another. Category searches will be slower than with the other modes. This mode was used before the **Categories Type** setting existed. It can be set as a fallback if the cached overlap data is believed to be incorrect for some reason, eg. category searches are wrong.

- `Non-overlapping`
  Assume that no category overlaps another. All category searches will be as fast as the fastest `Auto-detect` mode search, but searches for overlapping categories may not show all results. This mode can be set to force higher-performance searches at the potential expense of accuracy.

For best search performance, categories that overlap one another – ie. contain pages in common – should be listed most-commonly-searched first in the **Categories** setting list. Also, the `CatnoLowest` field should be selected as one of the **Compound Index Fields** (p. 50. These guidelines will allow the `Auto-detect` mode to optimize the most searches to the fastest possible speed.

4.5.6 URL File

Syntax: the full path to a file on the web server's disk

This allows you to specify a file containing a list of site URLs to walk. This is an additional way of specifying more Base URLs 4.4.4. This file will be reread each time a rewalk is started. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

4.5.7 URL URL

Syntax: an HTTP URL to a plain text file (NOT HTML)

This allows you to specify the URL of a plain text file containing a list of site URLs to walk. This is an additional way of specifying more Base URLs 4.4.4. This URL will be refetched each time a rewalk is started. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

Warning: Due to the nature of `Stay Under`, a large number of URL URLs (1000+) in different directories will cause the crawl to progress very slowly, as all URLs encountered will need to be checked against every one of those directories. In such a situation, we recommend turning off `Stay Under` and instead writing your own `Required Prefix`/`Required REX` expressions, which will be more efficient.

4.5.8 Single Page

Syntax: one or more HTTP URLs, one per line

Here you may specify URLs for individual pages to include in the index. These pages are fetched and stored in the database like others but the hyperlinks on them are not followed during a walk.

If you change this and click "`Update`" instead of "`GO`" the added pages will be fetched immediately and added to the existing database. Pages removed from the list will NOT be removed from the database until the next rewalk.

4.5.9 Page File

Syntax: the full path to a file on the web server's disk

This may be used to specify a file containing URLs for individual pages.

If you change this and click "`Update`" instead of "`GO`" the added pages will be fetched immediately and added to the existing database. The file itself is not checked for changes, and pages removed from the file will NOT be removed from the database until the next rewalk. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

See also `Single page` 4.5.8.

4.5.10 Page URL

Syntax: an HTTP URL to a plain text file (NOT HTML)

This may be used to specify the URL for a plain text file containing URLs for individual pages. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

If you change this and click "`Update`" instead of "`GO`" the added pages will be fetched immediately and added to the existing database. The file itself is not checked for changes, and pages removed from the file will NOT be removed from the database until the next rewalk.

See also `Single page` 4.5.8.

4.5.11 Strip Queries

Syntax: select Yes or No button

Strip query strings from all URLs. Some URLs have query strings on the end indicated by a question mark (`?`). With this option set to Yes, all query strings are removed from URLs before they are processed or retrieved.

4.5.12 Ignore Case

Syntax: select Yes or No button

This tells Webinator whether to ignore case in URLs or not. The case of hostnames is always ignored but the case of paths and filenames is respected. Some web servers don't respect case and people use various

random capitalizations within filenames making the same file look like different URLs.

4.5.13 Extra Domains

Syntax: one or more domain names separated by space or line break

Allow walk to fetch pages from any host in the specified domain(s). Any URL with a hostname ending in any of the specified domains will be accepted.

e.g.: Given a base URL of `http://www.mysite.com/` and extra domain `othersite.com` Webinator will walk all of `www.mysite.com` and any URLs referring to any machine in `othersite.com`.

This option is not a "restricter" but an "enabler". All hosts specified will be walked and any others that match the given domain(s) will also be walked.

Note: This option does NOT direct the walk to completely index every web server in the specified domain. It simply allows walking them if a reference to them is encountered.

4.5.14 Extra Networks

Syntax: one or more IP address prefixes separated by space or line break

Allow walk to fetch pages from any host within the network specified by the numeric IP address(es).

e.g.: Given a base URL of `http://www.mysite.com/` and extra network `192.0.2` Webinator will walk all of `www.mysite.com` and any URLs referring to any machine having an IP address prefix matching `192.0.2`.

Note: This option does NOT direct the walk to completely index every web server in the specified network. It simply allows walking them if a reference to them is encountered.

Note: Using this option has the potential to slow the walk, because every URL's hostname must be looked up. If there are many different off-site hosts, or your DNS is slow, the walk may be slowed substantially.

4.5.15 Extra URLs REX

Syntax: zero or more regular expressions (REX), separated by space or line break

Restricts walks to fetch URLs only matching any of the specified regular expressions anywhere in the URL (hostname, path, or query) when the Base URL matches.

If a Base URL is matched by an Extra URLs REX, then the only URLs that match the Extra URLs REX will be crawled on that host. If a Base URL does not match an Extra URLs REX, then it is walked as normal.

It is a rarely used setting, most commonly used in conjunction with a hostname to fetch matching URLs on an additional host. Links still need to be found to those pages for them to be indexed.

For example, with the following Extra URLs REX:

```
>>=http://products\.mysite\.com=!supplierid+supplierid\=BigCo
```

(which matches a URL that begins with `products.mysite.com` and contains `supplierid=BigCo`), and using the following Base URLs:

```
http://products.mysite.com/listProducts.aspx?supplierid=BigCo
http://help.mysite.com/index.aspx
```

The Extra URLs REX matches the `products.mysite.com` URL, so only pages with `supplier=BigCo` will be walked, while all of `help.mysite.com` will be walked (following other inclusion/exclusion rules).

Available from version `4.3.9`.

See also `Extra Domains`, p. 37.

4.5.16 Exclusion REX

Syntax: zero or more regular expressions (REX), each on a separate line

Excludes URLs matching any of the specified regular expressions anywhere in the URL (hostname, path, or query).

Table 4.3: Exclusion REX examples

REX	Matches
`/scratch[0-9]/`	a subdirectory named `scratch` followed by a single digit
`[^\alnum]test[^\alnum]`	the word `test` (but not `retest` or `tester` etc.)

See also `Exclusions` 4.4.10, `Exclusion prefix` 4.5.17 and `Exclude by Field` 4.5.18.

4.5.17 Exclusion Prefix

Syntax: zero or more URL prefixes, each on a separate line

Excludes URLs beginning with any of the specified prefixes. The entire URL (hostname, path, and query) is used for comparison.

Examples:

```
http://www.mysite.com/scratch0/
http://www.mysite.com/scratch1/
http://www.mysite.com/books/t
```

See also `Exclusions` 4.4.10, `Exclusion REX` 4.5.16 and `Exclude by Field` 4.5.18.

4.5.18 Exclude by Field

Syntax: Metamorph query, field to search, what to exclude

This provides more flexible control of what to exclude and how to exclude it. One exclusion per row of controls may be entered; new blank rows will be provided as rows are used. The `Metamorph Query` column is where a Metamorph query (ie. a typical search on Webinator) is entered: eg. several keywords or a regular expression. The `Field` and `Meta Field` columns determine what the `Metamorph Query` searches: if `Meta Field` is non-blank, that named meta field is searched, otherwise the field selected in `Field` is searched. The `Exclude` column controls the action for pages that match the query: `Pages and links` indicates that both the matching page and its links are to be excluded; `Pages only` indicates that the matching page is to be excluded but its links are still followed – this is useful for excluding navigation-only pages; `Links only` indicates that the page is still included but its links are excluded.

See also `Exclusions` 4.4.10 and `Exclusion REX` 4.5.16.

4.5.19 Additional Fields

Syntax: Name, Type, Searchable, Sortable, Output

The additional fields allow you to add up to three additional fields to the index. These fields can be included in the output if you use the XSL or XML output, sorted on, and searched on. They are populated with the `Data from Field` (p. 39) settings.

Additional Fields are supported in the full Texis product, but not Webinator-only.

Name - specifies the name of the additional field. It also specifies element that will hold the field contents if it is output in XML. The name must be a valid XML element name (may contain only `alnum` or `-_` and must start with a letter or _)

Type - specifies the internal storage type for the additional field. Anything can be stored as `Text`, but if you want to do numeric or date comparisons (such as sorting), you have to use an appropriate data type.

Searchable - specifies whether this additional field is directly searchable. This is done with an additional URL parameter that is separate from the normal query. Please see the `Additional Fields` section of `Procedures and Examples`, p. 120, for more details.

Sortable - specifies whether you allow sorting by this additional field. This is done with the `order` search parameter. Please see the `Additional Fields` section of `Procedures and Examples`, p. 120, for more details.

Output - specifies whether this field should be included with the output for XML results. Note that this ONLY refers to XML output, none of the 'stock' result styles will include additional fields. If you want an additional field to show up in your search results, you must set `Output` to `Y` for the field, use `XSL Stylesheet` result style, and customize the stylesheet to display the element for the Additional Field.

4.5.20 Data from Field

Syntax: REX expression, Replace expression, field to search, where to store it

This provides alternate means of setting both the HTML fields (`Modify Date`, `Title`, `Description`) and any Additional Fields. It allows getting page information from non-default places by searching and

optionally replacing the data. New blank rows will be provided as rows are used. See below for examples.

REX Search - Allows you to specify a REX expression to narrow down what contents of the `From Field` will be used. Leave it empty to use the entire field.

Note that a `REX Search` must be specified for the following `From` field types:

- `HTML`
- `Text`

You can specify they entire field for these by using `.*` as the `REX Search`.

Replace - `Replace` can be used to specify a subset of the value to be stored in the `To` field (or subset of the match, if you're using `REX Search`. It uses `sandr` replacement string syntax.

From Field - specifies what the source field is for the data.

- `HTML` - the raw HTML source of the page.
- `Text` - the text of the page, after HTML rendering has been applied.
- `Title` - the HTML title of the page
- `All Meta` - the contents of all `meta` headers specified in the HTML page.
- `Meta Field ->` - the contents of a specific `meta` field, specified in the next input box, **From Meta Field**.
- `Keywords` - the contents of the `keywords` meta header.
- `Description` - the contents of the `description` meta header.
- `Mime Type` the MIME type of the page. This may have been derived from the `Content-Type` header, a `<META HTTP-EQUIV>` tag, or the URL extension, depending on what is available.
- `URL` - the URL of the page.
- `URL Decoded` - the decoded version of the URL. Any `%XX` 'URL-safe' sequences in the URL are replaced with their real characters. E.g. `Pre%20%2D%20Expense%20Report.doc` is decoded into `Pre - Expense Report.doc`.
- `URL Protocol` - the URL's protocol, e.g. `http`.
- `URL Host` - the host (without port number) from the URL.
- `URL Host and Port` - the host (and port number if given) from the URL.
- `URL Path` - the file path from the URL.
- `URL Path Decoded` - the file path from the URL, URL-decoded.
- `URL Anchor` - the anchor from the URL (if any), i.e. the part after the # (pound sign). May not be available if already stripped.

- `URL Query` - the query string from the URL (if any), i.e. the part after the `?` (question mark).
- `URL Query Var ->` - the value of the URL query-string variable named in **From Meta Field**, URL-decoded.

From Meta Field - If `Meta Field ->` or `URL Query Var ->` is given as the **From Field**, this field is used to specify which meta field's or query var's contents to use as data. Leave blank otherwise.

Entering text in this field will force the use of `Meta Field ->`, regardless of the **From Field** setting.

To Field - specifies where information should be stored. `Modified Date`, `Title`, and 'verb'Description' are the standard HTML fields. If you've defined any Additional Fields, they will also be listed as selections here.

If you just added an Additional Field, you will need to hit `Update` for the Additional Field to appear in the `To Field` list.

Data From Field Example - Using Description for Title

If there's a site that uses the same HTML title for every page but has a nice description, you can use the following settings to store the description in the `title` field (in addition to the `description` field).

REX Search - *(Empty)*
Replace - *(Empty)*
From Field - `Description`
From Meta Field - *(Empty)*
To Field - `Title`

Data From Field Example - using PublishDate for Modified Date

If you're crawling a site of articles that specify a `PublishDate` meta field for every page, you can use that field's value instead of the normal Modified Date. **REX Search** - *(Empty)*
Replace - *(Empty)*
From Field - `Meta Field ->`
From Meta Field - PublishDate
To Field - `Modified Date`

Data From Field Example - grabbing Price from meta

If the site your crawling defines a meta header on each page containing a price, it's possible to store that numeric data in an Additional Field for searching. Assuming you've already defined an Additional Field called `Price`, the following settings would save that meta field in the Additional Field.

REX Search - *(Empty)*
Replace - *(Empty)*
From Field - `Meta Field ->`
From Meta Field - Price

To Field - `Price`

Data From Field Example - grabbing Price from Text

The target site might not be organized enough to stick the Price value in a meta header. If every page contains text in the format `Price: $19.95`, Data From Field can key in on that.

REX Search - `Price:=\space+\$\P=[0-9\.]+`
Replace - *(Empty)*
From Field - `Text`
From Meta Field - *(Empty)*
To Field - `Price`

Notice that we use the field `Text` as the source, not `HTML`. By operating on the formatted text instead of the raw HTML source, it allows proper operation even if the HTML source uses things like `Price: <b>$19.95</b>` or `<td>Price:</td><td>$19.95</td>`.

4.5.21 Required REX

Syntax: zero or more REX expressions, separated by whitespace

If specified, *all* URLs walked by Webinator must match at least one of these expressions. Opposite of `Exclusion REX`.

4.5.22 Required Prefix

Syntax: zero or more URL prefixes, separated by whitespace

If specified, *all* URLs walked by Webinator must match at least one of these prefixes.

4.5.23 Max Page Size

Syntax: a whole number from 1 up

Sets retrieved page size limit to the specified number of bytes. Pages larger than the limit will be truncated - not discarded.

Note: PDF files tend to be very large for the amount of text contained within them. Truncated PDF files are not processable due to their design. Make sure this setting is large enough to handle the largest PDF file you want to index.

4.5.24 Max Pages

Syntax: a whole number from -1 up

Limits the number of pages retrieved in a run to the specified number. Use `-1` for no limit.

4.5.25 Max Bytes

Syntax: a whole number from -1 up

Limits the number of bytes retrieved in a walk to the specified number. Use `-1` for no limit. The actual limit is rounded up to include the size of the last page so that it does not get truncated.

4.5.26 Max Depth

Syntax: a whole number from -1 up

Limits the depth of page retrieval to the specified number. Use `-1` for no limit. Depth is determined by counting how many links were traversed to reach a particular page. The base URLs are all at depth 0. URLs referred to by the base URL are depth 1, and so on.

4.5.27 Max URL Size

Syntax: an integer from 1 through 2033

Limits the size of URLs crawled. URLs longer than this will be skipped. Should not exceed 2033. The default is 1024.

4.5.28 Max Requests

Syntax: an integer greater than 0

This gives the maximum number of server requests (page fetches) to make on a single server connection (ie. Keep-Alive requests), if the server and protocol support multiple requests. Multiple requests per connection increases crawl speed, and is needed for Windows/NTLM-protected pages. The default is `100`.

4.5.29 Max Connection Lifetime

Syntax: an integer greater than 0

This gives the maximum lifetime (in seconds) for a connection to a server. Multiple requests per connection may be made (if the server and protocol support it) until the connection is this old. The default is `600` (i.e. ten minutes).

4.5.30 Page Timeout

Syntax: a whole number from 1 up

Causes Webinator to timeout after the specified number of seconds during each page fetch. This includes the time to lookup the IP address of the host, make the connection to the server, and download a single page. A timeout does not cause the entire process to quit. That page is just skipped and considered unavailable.

4.5.31 Meta Tags

Syntax: zero or more meta tag names, each on a separate line

This option tells Webinator to look for the specified meta data in fetched documents and store it in the database. Then, this data is included in text searches. The meta tags “Description” and “Keywords” do not need to be specified here because they will be indexed by default. See below.

4.5.32 Standard Meta

Syntax: select Yes or No button

This option indicates whether to automatically extract the standard meta tags “Description” and “Keywords” from HTML documents. If “Yes”, description and keywords meta data will be extracted and stored in their own fields within the database, unlike other meta data which will be collected and placed together into a single meta field in the database. These meta tags will be included in the search with a higher precedence than other meta tags.

4.5.33 All Meta

Syntax: select Yes or No button

Extract all meta data from HTML documents and place this data into the meta field for searching. This eliminates the need to know the name of all possible meta tags, but it also opens the possibility of recording all manner of nonsensical meta data.

4.5.34 Storage Charset

Syntax: standard IANA character set (charset) name

This sets the charset for storing page text in the database during walks. Pages will be translated to this charset when inserted. If a page cannot be translated, it is stored and labeled with its source charset (if known). If left empty (the default) it is `UTF-8`. This charset should be a superset of `US-ASCII` (same 7-bit sequences), and translatable by Webinator from all walked pages’ source charsets.

Note that this is *not* necessarily the charset that search results will be displayed in: see Display Charset under Search Settings. This setting is the default value for Display Charset; see notes under Display Charset.

4.5.35 Source Default Charset

Syntax: a standard IANA character set (charset) name

If the source charset for a walked URL is not labeled and cannot be determined, assume it is this character set. Default is `ISO-8859-1`. This should only be changed if a large number of walk pages are in an unlabeled different charset, eg. a Windows charset.

4.5.36 XML UTF-8

Syntax: select Yes or No button

Whether to attempt to clean up UTF-8 data for XML output: remove invalid sequences and characters. Should be `Yes` if XML output (eg. result style 8) is used (and Storage Charset should be empty). This helps avoid browser errors with XML pages. *Note:* if XML output is *not* being used, this should be set to No, as certain characters that are HTML-safe but not XML-safe will be removed if enabled.

4.5.37 Keep HTML

Syntax: select Yes or No buttons

Specifies whether to include the named type of text in the database.

ALT text

`ALT` text from `IMG` or `AREA` tags.

<STRIKE>

Text between `<STRIKE>` and `</STRIKE>` tags.

<DEL>

Text between `<DEL>` and `</DEL>` tags.

<FORM>

Text of form elements, such as `<input>` tags, `<select>` boxes, and `<textarea>` elements.

4.5.38 Keep Links

Syntax: select Yes or No buttons

Specifies whether to follow the named type of links when crawling.

Stylesheet

Links from `<LINK HREF=... REL=stylesheet>` tags. Note that non-`stylesheet` `<LINK>` tags will still be followed. The default is N.

<FORM>

Links from `<FORM ACTION=...>` tags. Without the rest of the form properly filled out, such links can often produce nuisance error pages from database-driven sites. The default is N.

4.5.39 Remove Common

Syntax: select Yes or No button

This causes common leading and trailing text from pages to be removed from the database. This is good for eliminating navigation menus and other static boilerplate text at the beginning and/or end of each page.

4.5.40 Ignore Tags

Syntax: one or more pairs of strings, more input boxes are added as you fill string pairs

All data between the specified begin and end will be stripped from the HTML before the text is extracted. These are simple strings, not patterns or REX's and the case is ignored. This is useful for excluding boilerplate or otherwise unwanted portions of HTML documents.

4.5.41 Keep Tags

Syntax: one or more pairs of strings, more input boxes will be added as you fill string pairs

All data NOT between the specified begin and end will be stripped from the HTML before the text is extracted. These are simple strings, not patterns or REX's, and the case is ignored. This is useful for extracting prime interest areas of HTML pages without the surrounding boilerplate.

4.5.42 Ignore Characters

Syntax: List of characters

List characters here which should be removed from the text and query. These can be punctuation that is optional. Examples are part numbers, phone numbers, etc. Take care to avoid removing important characters, which you may want to delimit words. Eg. with the setting "`-@`", the text "`part 123-45@6`" would be stored (and searchable as) "`part 123456`" instead.

4.5.43 Plugin Split

A group of settings that control whether and how to split `anytotx` plugin output into multiple sub-URLs in the table. Non-text files, such as PDFs, that `anytotx` processes are often very large or composed of sub-files. The Plugin Split setting allows these files to be split up for finer-grain searching. Split files will cause more than one URL to be entered in the `html` table (and thus also in potential search results) for the original URL. Such subsequent URLs will have an anchor appended to distinguish them from each other; usually this is the sub-file name, but it may be generic eg. "`#part5`" if there are no sub-files. *Note:* adjusting any of these settings can affect the ability of Refresh-type rewalks to complete successfully (New walks operate as usual).

Depth The Depth setting controls at what depth to split `anytotx` output. Each time a multi-file archive is unpacked by `anytotx`, the depth increases. Depth 0 (the default) means split at the top level (ie. do not split). Depth 1 would therefore insert each file of a ZIP file as a separate URL in the table.

Bytes The Bytes setting controls how many bytes each part will be after the file has been split. The default of 0 indicates do not split. This is useful for large monolithic files that have no detectable sub-file or page structure. If both Pages and Bytes are set, the first limit reached is used for each part.

AtPage The AtPage setting controls whether to force the Bytes-controlled splitting to occur at a page boundary (a Ctrl-L). Checking this may make each part arbitrarily larger than the Bytes setting, because a part may extend to the next page break. With this setting unchecked, a part may be up to 50% larger than the Bytes setting, because the page-break check will only go that far over the limit.

Pages The Pages setting controls how many pages to group in a part. The default of 0 does not split at all. If both Pages and Bytes are set, the first limit reached is used for each part. For example, setting Pages to 10 and Bytes to 100000 would break at 10 pages or 100KB, whichever comes first. This is useful to catch page-bounded documents like PDFs, and simultaneously avoid generating huge text for non-paged documents.

Plugin Split was added in version 4.03.1049838346 Apr 8 2003.

4.5.44 Language Analysis

If **Enable** is set to `Y`, pages walked are processed through the Language Analysis Module (LAM), obtained and installed separately. This module helps support searching in languages such as Chinese, Japanese and Korean, where there is often no whitespace to delineate one "word" (logogram, or group of characters) from another, making searching difficult. The Language Analysis Module inserts spaces between words in the text of such pages, enabling ordinary non-wildcard searches to match better. At search time, users' queries are also passed through the module, so that they can match the processed pages' text.

Language

A two-letter ISO 639 language code "hint" for the LAM. If all or a majority of the crawled data is a single language, entering that language's code here will help the LAM process data better. The default is empty (no hint). Added in Texis version 6.00.1294975881 20110113.

Preserve 7-bit

Whether to preserve the separation of all-7-bit tokens. Sometimes the LAM will separate alphanumeric tokens that are not language words, e.g. part numbers, causing search problems. Setting this to `Y` will attempt to preserve the separation (or lack thereof) of all-7-bit tokens in the crawled text.

4.5.45 CJK Mode

Syntax: select Yes or No

`CJK Mode` modifies the crawl and the search for better handling many Chinese, Japanese, and Korean queries.

At index time, multi-byte UTF-8 characters are indexed as individual words. At search time, multi-byte UTF-8 characters in the query are separated by spaces, and quotes surround the sequence to make it a phrase.

This allows the query to match where spacing may cause it to otherwise not match.

4.5.46 Word Definition

Syntax: one or more regular expressions (REX), each on a separate line

Sets the word matching expression(s). Each line is a regular expression defining what is considered a word within the textual content of the retrieved documents during the index process. The default expressions index normal words and some special items such as domain names.

You may supply multiple expressions, one per line, if you can't define your idea of all possible words in one expression.

For example, `>>\alpha=\alnum{1,20}` will index "words" beginning with an alphabetic character followed by 1 to 20 alphabetic or numeric characters.

If **Word Definition** is changed, the **Language Characters** setting (p. 78) should generally be updated to reflect any new characters added.

Changing the word definition with `Update` instead of `Update and GO` will cause the existing search index on the data to be dropped and rebuilt. The database will not be searchable during the time that the index is being rebuilt; this may take several minutes or more for large profiles.

4.5.47 Text Search Mode

Syntax: select from options or enter custom mode

(Note: In earlier releases this setting was known as **Character Match Mode**.)

Sets the character-matching mode for text (keyword) searches. This controls aspects like case-sensitivity, ignoring accents, etc. The selectable values are:

- `Loose` - Ignore case, ignore diacritics (accents), expand ligatures, ignore width differences. **Storage Charset** should be empty or `UTF-8`, though `ISO-8859-1` may sometimes work. With this mode, not only will a lower-case "`e`" match an upper-case "`E`" and vice-versa (ignore case), but "`e`" will match "`é`" (Unicode U+00E9), "`oe`" will match "`œ`" (U+0153), and full-width will match half-width characters (for ASCII and katakana).
- `Strict` - Ignore case only. "`e`" will match "`E`", but not "`é`". **Storage Charset** should be empty or `UTF-8`, though `ISO-8859-1` may sometimes work.
- `Strict ISO-8859-1` - Ignore case only, and assume **Storage Charset** is `ISO-8859-1`. For back-compatibility. Available only for **Text Search Mode**.
- `Exact` - Match characters exactly, respecting case, diacritics, width etc. Available only for **Attribute Compare Mode**.
- `Custom ->` - Use the custom mode entered in the **Custom Mode** box. This is a comma-separated list composed from the following tokens; consult Thunderstone tech support for advice:
  - `iso-8859-1` - Assume text is ISO-8859-1 encoded. Should only be used if **Storage Charset** is also `ISO-8859-1`. If this flag is not set, text is assumed to be UTF-8, though occasional ISO-8859-1 characters will usually be able to match their UTF-8 equivalents.

- `ignorediacritics` - Ignore diacritic marks (accents, umlauts, etc.). Eg. "`e`" will match "`é`" (U+00E9) and vice-versa.
- `expandligatures` - Expand ligature characters. Eg. "`oe`" will match "`œ`" (U+0153) and vice-versa. Note that with this flag off, certain ligatures may still be expanded if necessary for case-folding with `ignorecase`.
- `ignorewidth` - Ignore half-/full-width differences, eg. for ASCII and katakana characters.
- `ignorecase` - Ignore case differences, eg. "`e`" matches "`E`" and vice-versa; this is the default. The alternative is `respsectcase`.
- `respectcase` - Case-sensitive search, eg. "`e`" does *not* match "`E`". The alternative is `ignorecase`.
- `unicodemulti` - Use Unicode case-compare tables, with multi-character expansions where needed (eg. for ligatures). The alternative is `ctype` or `unicodemono`.
- `unicodemono` - Use Unicode case-compare tables, but do not expand characters. The alternative is `ctype` or `unicodemulti`.
- `ctype` - Use the operating system's `ctype.h` case-compare tables. Only codepoints U+0001 through U+00FF (ie. single-byte or ISO-8859-1 range) are supported, though the actual encoding may be ISO-8859-1 or UTF-8 depending on the `iso-8859-1` flag. The alternative is `unicodemulti` or `unicodemono`.

Note: Changing the **Text Search Mode** setting will cause text search indexes to be rebuilt, which may take several minutes or more for large profiles.

4.5.48 Attribute Compare Mode

Syntax: select from options or enter custom mode

Sets the character-matching mode for attribute comparison searches, e.g. equals, less-than, order-by, `IN`. This controls aspects like case-sensitivity, ignoring accents, etc. See **Text Search Mode** (p. 48) for details on what the setting values mean. The default is `Exact`. Note that searches on `Enum` fields are unaffected by this setting, as the `Enum` type is defined to be case-insensitive.

Note: Changing the **Attribute Compare Mode** setting will cause **Extra Indexes** (if any) to be rebuilt. This may take a few minutes on large profiles, and may prevent crawls from proceeding until the indexes finish.

4.5.49 Index Fields

Syntax: list of fields ordered by desired weight

These fields will be searched by the user's text query. Fields listed higher will be weighted higher in search results, according to the Position in Text search setting.

Note that changing these fields will cause indexes to be rebuilt, which may take several minutes or more for large-data profiles. The old setting will be used until the index rebuild is complete.

4.5.50 Compound Index Fields

Syntax: list of field(s) from select boxes, any order

These fields will be indexed along with Index Fields, but in the compound portion of the main search index. They are not searched by the text query, but are used to improve accuracy and performance for certain ancillary queries performed in *addition* to the main text search, such as when ordering results by last-modified date, or searching by Depth. The default values are `Visited`, `Modified`, `Depth` and `Pop`.

The selected fields may be in any order; they are used only when needed, unlike Index Fields, all of which are always searched by the user's text query. However, note that adding a field to Compound Index Fields will not help search performance if there is no text query also.

Note that as this is the same overall index as Index Fields, changing any of these fields will cause indexes to be rebuilt, which may take several minutes or more for large-data profiles. The old setting(s) will be used until the index rebuild is complete.

4.5.51 Extra Indexes

Syntax: select-box for index type and table, text box to enter index name and fields

Extra Indexes may be created to improve search performance and accuracy in situations where the main text index (Index Fields) and/or its Compound Index Fields are not sufficient. They are not generally created unless suggested by Thunderstone tech support for certain queries.

Note that creating an Extra Index on a large-data profile may take several minutes or more. If the index Type is not Metamorph nor Metamorph Inverted, creating the index may also impede crawls or other database modifications. Non-Metamorph/Metamorph-Inverted indexes should therefore be created *before* the profile is crawled or populated with data to avoid this issue, if possible. Extra Indexes should only be created when the profile is not actively crawling, to minimize load and potential crawl impediments.

4.5.52 Spell-check Dictionaries

Syntax: select-box choice

This setting controls what dictionaries to create for spell checking. The default (`Create all`) is to create all needed dictionaries. However, this can consume significant time and memory for some large-data profiles, so to conserve system resources, only the multi-word-occurrence dictionary may be created (`Create multi-word only`). This may reduce spell-check suggestions at search time however. To further conserve system resources, no dictionaries at all may be created (`None`). This will disable spell checking at search time.

4.5.53 Primer Type

"Primer URLs" are URLs that are fetched before actually starting a crawl. They are not stored in the search database, but instead are used to "prime" Webinator with any necessary credentials (eg. login cookies) for accessing the rest of the site. By default, the Base URL is used, in case any session/ASP cookies are needed.

The **Primer Type** setting specifies which (if any) urls are used to prime the profile:

- `None` - No primer URL is used. The Base URLs are crawled as normal.
- `Base URL` - the Base URLs are used to prime the walk. This differs from `None` in that the base URLs are submitted once and the results discarded, and then submitted again for crawling.

  This is useful in situations where the Base URL contains login information, and the page returns "thank you for logging in" with no other content until the page is requested again.
- `Custom` - The URLs listed in the **Custom Primer URLs** setting are used, as described below.

For HTTP Basic or NTLM protected web sites, the **Login Info** setting should be used instead.

4.5.54 Primer URLs

Syntax: URL, optional variables, optional bad-login query, optional URL query

When the **Primer Type** setting is set to `Custom`, the **Primer URLs** setting values take effect. There are two ways to use a custom primer URL - submitting the form directly, and filling out the form.

Submitting the Form Directly: Custom Primer URL

If a form-based login must be filled out before accessing a site, the **Custom Primer URL** can be set to the `<FORM ACTION>` URL of the login (fully-qualified), with any form variables (eg. user/pass) filled out in the query string. If the `<FORM METHOD>` must be `POST` instead of `GET`, the URL protocol may be changed to the pseudo-protocol "`http-post`". Eg.:

```
http-post://login.acme.com/checkLogin.asp?User=Admin&Pass=open-sesame
```

would be submitted using the `POST` method, with the given query-string variables sent as the content. Note that the query-string variables and values should be URL-encoded.

Filling Out the Form: Custom Primer Variables

Sometimes submitting the form directly is not sufficient. Forms on web pages can contain dynamic hidden variables, such as a `viewstate` for session tracking. This means the form must be opened, filled out, and submitted, instead of simply submitting a pre-defined action URL.

This is achievable with the **Custom Primer Variables** setting. Instead of setting **Custom Primer URL** to the action of the login form, you set it to the URL of the page that contains the form. **Custom Primer Variables** is a URL-encoded list of name/value pairs to set on the **Custom Primer URL** page.

When **Custom Primer Variables** is set, the **Custom Primer URL** is fetched, and then the variables specified in **Custom Primer Variables** are used on the form, and then *that* form is submitted.

For example, let's say there's a `pleaseLogin.asp` page that submits to `checkLogin.asp`, and the form contains a dynamic state that has to be included or `checkLogin.asp` will reject the login. If you set **Custom Primer URL** to

```
http://login.acme.com/pleaseLogin.asp
```

and set **Custom Primer Variables** to

```
User=Admin&Pass=open%26close
```

The `pleaseLogin.asp` page will be fetched, the form field `User` will be set to `Admin` and `Pass` will be set to `open&close` (note the URL-encoding), and then form on the `pleaseLogin.asp` page will be submitted, going to `checkLogin.asp`.

This means that if the form on `pleaseLogin.asp` contains

```
<input type="hidden" name="sessionstate" value="abc123xyz"/>
```

then that hidden variable will be submitted along with the rest of the form.

Checking for Bad Logins: Bad Login MM Query

Sometimes, the primer URL login may fail, eg. bad login. However, since the only error indication may be a "Login failure"-type message and not a true HTTP error code, Webinator may not be able to detect this and might continue walking useless (permission-denied or "Please log in first") pages.

To help detect such a primer URL failure, a **Bad Login MM Query** may be entered. If non-empty, this is a Metamorph query to run against the HTML returned from the associated primer URL. If it matches, the primer URL is considered a failure, and the crawl is stopped for that particular site (other Base URLs will continue).

Multiple Primers: Base URL MM Query

If multiple custom primer URLs are being used, you can control which ones are used for which Base URLs via Base URL MM Query.

By default, primer URLs are only used on Base URLs that have a matching protocol and hostname. If **Base URL MM Query** is non-empty, then this Metamorph query will be run against the Base URL being crawled. The associated primer URL will only be fetched if it matches.

4.5.55 Login Info

Syntax: name and password

Specify a username and password for sites that require a login to view certain pages. These are used with HTTP Basic, Windows NTLM, and FTP authentication. Other authentication methods are not supported currently. Without proper login, protected pages will be skipped.

If this is a domain account, enter both in the Username field, separated by a forward slash (`/`), i.e. `MY_DOMAIN/myuser`.

If you are trying to walk a site where a login form is provided on a web page, you may be able to walk it by using the action URL from the form with the form variables encoded onto the end as your base URL. For example if the form variable names were Uname and Upass and the action URL was `http://www.mysite.com/login.asp` you may be able to use a URL like

```
http://SERVER/login.asp?Uname=YOURNAME&Upass=YOURPASSWORD
```

Note: The search interface displays hit context and has an option to view the entire text of the page. This allows search users to view "protected" pages without entering a password.

4.5.56 Proxy

Syntax: the full URL to a web proxy server

This specifies the URL (not just hostname) of a proxy web server through which to pass page fetch requests. Blank means don't use a proxy.

4.5.57 Proxy Login Info

Sets the user name and password to authenticate to proxy servers, using the `Proxy-Authenticate` header and Basic Authentication. Used if the Proxy URL is filled in. Added in version 4.01.1031600000 Sep 9 2002.

4.5.58 Cookie Source Path

File path to a Netscape or Microsoft Internet Explorer format cookie file to read at start up. This allows persistent cookies saved by a browser to be read by Webinator, so it can inherit the browser's state. To easily walk a site that requires a custom login (ie. not HTTP Basic authentication), and that uses persistent cookies, just login normally using a browser run *on* the Webinator machine itself. Then, enter that browser's cookie file in the Cookie Source Path setting (this is typically `%USERPROFILE%\Cookies` for Explorer on Windows). Then, Webinator will automatically inherit the browser's permissions. Added in version 4.02.1042043803 Jan 8 2003.

4.5.59 Off-Site Pages

Syntax: select Yes or No button

Allow retrieval of individual off-site pages. By default Webinator will not retrieve pages that are not on the same host as the base URL(s). Using this option, pages not on the same machine will be retrieved, but none

of the pages that they reference will be walked. This option also allows off-site redirects, frames, and iframes to be fetched.

4.5.60 Stay Under

Syntax: select Yes or No button

When this flag is Yes, walks will stay under the directory specified in the base URL(s). When this is No, if a hyperlink to another location on the same site is encountered, the will follow the link. In neither case will the walk go to other sites unless they are in the list of walk URLs or allowed domains or networks.

4.5.61 Prevent Duplicates

Syntax: select Yes or No button

This option enables extra checking for duplicate documents. Documents with the same content are only be stored once, even if their URLs are different. This is accomplished by hashing the textual content of the page and not storing any page with a hash code that is already in the database.

4.5.62 Duplicate Check Fields

Syntax: checkboxes to choose fields

These are the fields which will be checked for duplicate prevention (if `Prevent Duplicates` is enabled). The concatenation of these fields is hashed for each incoming document, and if the hash is the same as an existing document, the incoming document will be discarded as a duplicate.

By default, only `Body` is checked, as the body is the majority of search content for a document, and thus another document that has an identical body should be considered a duplicate even if it has a slightly different title or description.

However, sometimes errors in processing (eg. `anytotx`) can cause the bodies of large numbers of documents to become empty and thus be considered duplicates of each other and removed. In this case it may be desirable to either turn off `Prevent Duplicates` or check more fields in `Duplicate Check Fields`.

Note: Changing `Duplicate Check Fields` after a walk has completed (ie. before a later `Refresh` type walk) may cause new documents to not be removed as duplicates as expected, since the pre-existing documents' hashes are now for a different set of fields. This will not cause errors or corruption; it just might leave some newly-duplicate documents in the database.

4.5.63 Store Refs

Syntax: select Yes or No button

Controls whether URLs referenced by retrieved pages are added to the refs table. This can save some time

during the walk, as well as, disk space if it's turned off. But turning it off prevents the "`Show Parents`" option in the search from working. It also reduces the detail available from walk error reports.

4.5.64 Inline Iframes

Syntax: select Yes or No button

This indicates whether to treat iframes as a part of the page they are on or as separate stand alone pages. Selecting Yes will make them part of the page. Selecting no will make them separate.

4.5.65 Max Frames

Syntax: a whole number from 0 up

This indicates the maximum number of frames allowed on a page. Pages with more frames than this are discarded. If this is set to 0, the frames of framed documents are treated as independent, stand-alone pages.

4.5.66 Execute JavaScript

Syntax: select Yes or No button

Execute JavaScript that is contained on fetched pages and that might alter or generate the page content and URLs.

4.5.67 Fetch JavaScript

Syntax: select Yes or No button

Fetch JavaScript that resides at a separate URL instead of being inline on the page (eg. `<SCRIPT SRC>` tags).

4.5.68 JavaScript String Links

Syntax: select appropriate checkboxes

Sets which additional sources of potential JavaScript links to check. Some JavaScript links may not be found when scripts on a walked page are executed, so the internal list of all JavaScript string objects is scanned for potential URLs according to the checked boxes. `Menu` will look for common JavaScript menu navigation system links; `Protocol` will look for strings that look like valid fully-qualified Web links; `File` will look for probable file strings.

Note that any of these sources may potentially find incorrect links, especially the `File` type. Checking `File` is generally used only as a last-ditch effort to find some JavaScript links.

4.5.69 Debug JavaScript

Syntax: select Yes or No button

Print additional debugging messages for JavaScript errors.

4.5.70 JavaScript Memory

Syntax: numeric memory size eg. `20MB`

Alters the max amount of memory allowed for running JavaScript. The default (if the setting is empty) is 20MB. Increasing the limit may help if error messages such as "`JavaScript exceeded scriptmem limit`" are encountered. Note that the Maximum Process Size limit setting may also need to be increased if this is increased.

4.5.71 JavaScript Timeout

Syntax: integer

Max time, in seconds, to allow for running JavaScript. The default (if the setting is empty) is 5 seconds. Large or complex JavaScript pages may require more time, eg. if "`JavaScript exceeded scripttimeout`" messages are received.

4.5.72 Protocols

Select which protocols to allow to be fetched. If a protocol is not enabled, but the Base URL uses it, it will be automatically enabled for the walk. The protocols currently supported are `http`, `https`, `ftp` and `gopher`.

4.5.73 HTTP Version

What HTTP version to use for requests. `HTTP/1.1` enables compression (gzip, chunked, compress, deflate Content-Encoding) and is the default for products using Texis version 6 and later. `HTTP/1.0` was the default for previous versions. `HTTP/0.9` is of limited/no use.

4.5.74 SSL Client Protocols

Which SSL protocols to allow for client HTTPS/SSL connections when crawling and searching, ie. for connections from Webinator to remote `https://` URLs. The default is to leave all protocols enabled for maximum compatibility; the most-secure protocol will then be negotiated. However, sometimes the connection fails at (or soon after) the negotiation, possibly with the error message "`Missing HTTP response line in reply from` ...". This may be due to settings on the remote server that disallow certain SSL protocols. In such cases, disabling various SSL protocols under **SSL Client Protocols** may enable the connection to succeed.

4.5.75 Authentication Schemes

Select which authentication schemes to allow for password-protected URLs. The settable schemes are `Basic`, `File` (for `file://` URLs), `NTLMv1` and `NTLMv2`. NTLMv2 requires Texis version 5.01.1213917000 20080619 or later. Note that the scheme(s) actually *accepted* for a given URL are determined by the server; if none of the server-offered schemes are enabled by this setting, then the protected URL cannot be walked. This setting can be used to disable less-secure or undesired schemes, such as Basic or NTLMv1 authentication.

4.5.76 Embedded Security

Select the security for embedded objects on a page (eg. frames, scripts). `Any` fetches any required object. `Non-decreasing` will fetch a required object if its security (`https://` vs. non-`https://` in the URL) is not less than the main page, ie. an `https://` object on an `http://` page will be fetched, but not vice-versa. `Non-increasing` is the opposite. `Same protocol` requires that the protocol of the object be the same as the main page.

4.5.77 Entropy Source

Selects standard (default) or alternate entropy source. Entropy is used to initialize the SSL/`https` plugin. The standard sources should be sufficient; the alternate source is only needed if the `prngd` daemon (some Unix platforms) is required but cannot be successfully run. *Note:* Setting the source to Alternate will decrease SSL/`https` security.

4.5.78 Multiple Fetches

Syntax: select Y or N

`Multiple Fetches` allows a page to be fetched multiple times, and can potentially slow down a crawl. It should only be used in specific situations in conjunction with `Off-Site Pages`.

For example, Consider the situation of crawling two sites, `a.com` and `b.com` with `Off-Site Pages` enabled. A link from a page on `a.com` to `b.com/page.htm` is considered off-site, so it will be crawled but its links won't. Then, when `b.com` starts its crawl, `b.com/page.htm` won't be processed because it's already been done, causing `b.com/page.htm`'s links to not be included.

`Multiple Fetches` allows the 2nd encounter of `b.com/page.htm` to be processed again, which will allow its links to be properly processed.

4.5.79 Follow Cross-Site Links

Syntax: select Y or N

When crawling multiple hosts, setting `Follow Cross-Site Links` to `Y` will allow links from one host to another to be respected, as opposed to only starting from each host's Base URLs.

If you have a lot of Base URLs that have lots of duplicate links to each other that would've been found on-site anyway, setting `Follow Cross-Site Links` to `N` can improve crawl performance.

4.5.80 Max Redirects

Syntax: a whole number from 0 up or -1

This indicates the maximum number of redirects that are followed when attempting to retrieve a page. If set to -1 then redirects will not be followed when attempting to retrieve the page, but will be treated as a link.

4.5.81 Empty Form Redirects

Syntax: select Y or N

Some crawled pages implement a redirect by having a HTML form that points to the target, and uses JavaScript to submit the form.

If `Empty Form Redirects` is set to `Y` and a page doesn't have any content, Webinator will treat any HTML `<form>` targets on the page as a redirect.

4.5.82 Index Name

Syntax: one or more filenames separated by space

Set the filename assumed for directory URLs. The default is "`index.html`" and "`index.htm`". This filename will be removed from stored URLs to prevent redundant fetches of the page. So the URLs "`http://www.mysite.com/fun/`" and "`http://www.mysite.com/fun/index.html`" will be considered the same and only be fetched once (as `http://www.mysite.com/fun/`).

4.5.83 DNS Mode

Syntax: choose from drop down list

This controls how Webinator looks up IP addresses for hostnames. "`Internal`" uses Texis's own internal parallelizing name lookup routines. "`System`" uses the standard system routines. You should use "`Internal`" unless it causes compatibility problems.

4.5.84 Net Mode

Syntax: choose from drop down list

This controls what API Webinator uses to access Web pages. "`Internal`" uses Texis's own internal parallelizing Web fetch routines. "`System`" uses the standard system routines. You should use "`Internal`" unless it causes compatibility problems.

Note: "`System`" only has effect for the Windows version of Webinator. It does not currently support parallel access and some other Web features of the "`Internal`" mode. However, it does provide an alternate way to access NTLM-controlled sites (using the user/password set in `Login Info`), in versions prior to October 2004. Later versions support NTLM authentication in the default "`Internal`" net mode.

4.5.85 User Agent

Syntax: full user-agent string

Set the User-Agent (browser type) to report to web servers. Normally Webinator reports itself as Mozilla version 4.0. Modify this setting to report as a different user agent. If you want to emulate a particular browser, you can access your site with that browser, then check the site's transfer log to see what user agent string was logged (typically the last double-quoted entry on the line).

4.5.86 Mime Types

Syntax: one or more acceptable MIME types, each on a separate line

These are the Multipurpose Internet Mail Extensions (MIME) types that Webinator informs the web server are acceptable. MIME types have the syntax `type/subtype`. Either `type` or `subtype` may be `*` to mean "any". By default all MIME types are allowed (`*/*`).

4.5.87 Respect Expires Header

Syntax: choose from drop down list

For `refresh`-type walks, this controls how the Expires header is used. Set to `No` the Expires header will be ignored. Set to `Limited` the Expires header will be used, but limited by the Minimum and Maximum Refresh Times. Set to `Yes` the Expires header will be treated as definitive.

Invalid and out of range headers will be ignored, with the exception of "0".

4.5.88 Default Refresh Time

Syntax: choose from drop down list

For `refresh`-type walks, this is the default time period to initially try refreshing a URL; typically set to 1 minute. Note that the actual refresh period is dynamically computed for each URL based on how often it changes.

4.5.89 Minimum Refresh Time

Syntax: choose from drop down list

For `refresh`-type walks, this is the minimum time period to try refreshing a URL. The actual refresh period is dynamically computed for each URL based on how often it changes, and will not be less than this value. This prevents too much time being spent refreshing a very dynamic page (ie. constantly refreshing it and loading the web server). Typically set to 1 minute.

4.5.90 Maximum Refresh Time

Syntax: choose from drop down list

For `refresh`-type walks, this is the maximum time period to try refreshing a URL. The actual refresh period is dynamically computed for each URL based on how often it changes, and will not be greater than this value. This ensures that all URLs – even relatively static ones – are eventually checked for changes.

4.5.91 Maximum Process Size

Syntax: choose from drop down list

Upper limit to memory size of walker processes. If a walker process exceeds this limit, it is re-started (at the same point it left off) by the dispatcher, at most once. If the same child repeatedly exceeds this limit, the walk may stop until it is re-started via schedule or manually.

4.5.92 Replication Settings

Syntax: List of hosts and profiles

A list of hosts and profiles to send walk data updates to. The hosts must have the sending server listed as a cluster member under the system-wide settings.

4.5.93 Debug Replication

Syntax: Choose Y or N

`Debug Replication` will log additional information about the replication process that can assist in troubleshooting. The log file written will be named `repl-{pid}.log`, where `{pid}` is the process ID of the replication process.

See also “Replication” 5.17.

4.6 Search Settings

This group of options applies to the standard search and provides a convenient way to make common changes to the search behavior and appearance. You are not limited to the features listed here. You may modify the search script to look however you want and to behave however you want.

See also “Customizing Webinator’s Appearance” 3.5.

4.6.1 Notes

This is the same setting as **Notes** under Walk Settings: a scratch pad area for the administrator of the profile. It in no way affects the walk or search.

4.6.2 Query Logging

Syntax: select Yes or No button

This indicates whether the search should log user queries. If Yes, users' queries are logged to the querylog table of the database. The contents of this table may be viewed from the `Query Log` menu of the Administrative Interface.

Note: The query log table gets erased during every new walk. You will only be able to view queries that have occurred since the latest new walk. Refresh walks do not cause the table to be erased.

4.6.3 Rotate Schedule

Syntax: The day of week (or daily) and the time of day to rotate

This selects when to rotate query logs on this profile. During a rotate action, the log table data is optionally e-mailed to someone, and then the data is erased from the log table.

See also `Attach Logs` (section 4.5.3).

4.6.4 Email

Syntax: A valid e-mail address

When the query log is rotated (according to the schedule set), an e-mail message with an attached file (containing the previous log data) is sent to this address. Multiple addresses may be specified, separated by commas.

4.6.5 Result Order

Syntax: select Relevance, Date, or URL button

This determines the default ordering of search results.

- `Rank` - search results are ordered by rank (or relevance) by default.
- `Date` - search results are ordered by date descending (newest first) by default.
- `URL` - search results ordered by their URLs alphabetically by default.

Search users may select the alternate ordering from this default in the Advanced search form.

4.6.6 Results Style

Syntax: choose from drop down list

This controls the style used for displaying individual answers to user queries. There are various styles from which to choose. The arrangement and amount of information varies in every style. In the administrative interface you may click the question mark (`?`) next to "`Results Style`" to see a sample of each of the available styles.

4.6.7 Allow RSS

Syntax: select Yes or No button

If `Allow RSS` is set to `Y`, then each search result page will include a reference to an RSS feed for that search, which users will be able to monitor.

Setting `Allow RSS` to `N` will both remove the reference from search result pages, and disallow the viewing of RSS feeds.

4.6.8 Format XSL Output

Syntax: select Yes or No button

If set to `Y`, then extra line breaks are added in to the output of the server-side XSL stylesheet processing. This has the following effects:

- It makes the HTML output more readable by humans, changing it from one extremely long line to a well formated document.
- It adds a small amount of size to the document (usually between 1-4%)
- Adding line breaks at certain locations can sometimes trigger odd rendering bugs in Internet Explorer (adding spaces where there shouldn't be spaces).

4.6.9 XSL File

Syntax: Browse local disk for a XSL file

This allows the use of a customized XSL file to format the output of a search. A default XSL style sheet is included with Webinator (`/webinator/xsl/default.xsl`). The **XSL File** option is used only if the **Results Style** is set to `XSL Stylesheet`. The links below this option display the current XSL stylesheets, which may be downloaded for editing and then re-uploaded with this option.

Note that the `/webinator/xsl` subdirectory of the web server's document root must be writable by the Texis user in order for this option to work. The install program normally does this at installation however.

4.6.10 Abstract Style

Syntax: choose from drop down list

This setting controls the short description or abstract that is generated for each search result. Choosing `Query` uses a snippet that matches the query. `Beginning` uses the start of the document's content. `Top` uses the top of the current page. `Description` uses the value of the `Description` meta tag.

4.6.11 Abstract Length

Syntax: enter number in text box

This determines the length in bytes of the document abstract.

4.6.12 Max Title Length

Syntax: enter number in text box

This determines the maximum length in bytes of the document title shown in the results. If the title is over this length, it will be truncated and ended with ellipses.

Title length may be expanded up to 10 characters over this setting in order to avoid cutting off in the middle of a word.

Set to `-1` to always use the full title.

4.6.13 Max URL Display Length

Syntax: enter number in text box

This determines the maximum length in bytes of the matching URL shown in the results. If the title is over this length, it will be truncated after the hostname with ellipses and ended with as much of the path and filename as it can.

Note that this does not affect the URL that is actually linked to - that URL is always the full, proper URL. This setting only affects the displayed URL.

Set to `-1` to always use the full URL.

4.6.14 Results per Page

Syntax: a whole number

This controls the number of results (answers) listed on each results page. When there are more than this many answers to a user's query the user will have to hit "next" to see more answers.

4.6.15 Max User Results per Page

Syntax: a whole number, or -1 to disable

Search users are able to customize how many hits per page they see by supplying the parameter `rpp`. This setting places an upper bound on how many results per page they can request. This prevents someone from requesting 1000000 results on a page and bogging down the search system.

If set to `-1`, then all `rpp` parameters are ignored.

4.6.16 Page Links Shown

Syntax: a whole number, defaults to `10`

This specifies the number of page links to include in the summary of the results.

For example, if we are on page 22 of 5,000 total results, by default direct links will be shown to pages 18 through 27 (for a total of 10 links). If `Page Links Shown` is set to `20`, it will show links 13 through 32, for a total of 20 page links.

4.6.17 Results Width

Syntax: a whole number or a percentage valid for an HTML <TABLE> WIDTH

This controls the width of the `<TABLE>`s used in the search results. This may be a number indicating a fixed width or a number from 1 to 100 followed by a percent sign(%). This tells the user's web browser how wide to make the table.

4.6.18 Box Color

Syntax: a color name or number valid for HTML color specification

This controls the color of the "gray" informational boxes at the top and bottom of search results pages.

4.6.19 Show Advanced Search

Syntax: select Yes or No button

This controls whether or not the Advanced Search button is displayed on the search form. If set to No then the button will be hidden, otherwise it will be displayed.

4.6.20 Results Highlighting

Syntax: select None, Classes, Inline or Bold

The user's query will be highlighted in various parts of the search results (Title, Abstract, etc.) with the selected method:

- `None` or `N` - No highlighting will be done in search results.
- `Classes` or `Y` - Terms will be highlighted with `<span>` tags that refer to classes that should be defined in a separate CSS file, e.g. the `/webinator/common/search.css` file, which can be edited to customize the highlighting style. Each term in the query is highlighted with a different class (in a different color, by default).
- `Inline` - Terms will be highlighted with `<span>` tags that directly specify a fixed CSS style. This is not customizable, but is self-contained and does not depend on a separate stylesheet or file. Same visual result as `Classes` with the default CSS.
- `Bold` - Terms will be highlighted with `<b>` tags.

The default is `Bold`.

4.6.21 Context Highlighting

Syntax: select None, Classes, Inline or Bold

The user's query will be highlighted in the context view (Match Info page) with the selected method. Same choices as for **Results Highlighting**. The default is `Classes`.

4.6.22 PDF Query Highlighting

Syntax: select Yes or No button

When making links to PDFs in search results, Webinator will add extra info to the link which will cause the user's query to be highlighted by the PDF viewer. Changing this setting to "N" will remove that extra information from the link, and no longer highlight the user's query in the PDF document.

4.6.23 Font

Syntax: a font name valid for HTML <FONT> specification

This specifies the font to use throughout the search interface.

4.6.24 Display Charset

Syntax: a standard IANA charset name

This sets the charset used to display search results in. The default if empty is the charset for Storage Charset under All Walk Settings. This charset should be a superset of `US-ASCII` (same 7-bit sequences), compatible with Top HTML, and translatable by Webinator from Storage Charset.

A `<META HTTP-EQUIV=Content-Type>` tag in Top HTML will be updated automatically to reflect this charset. This update can be disabled by putting 2 or more spaces between `META` and `HTTP-EQUIV` in Top HTML.

Note that if the Display Charset differs from the Storage Charset, search results must be converted on-the-fly, potentially degrading performance slightly. Thus, if Display Charset is ever changed, it is recommended that Storage Charset be changed as well, and after the next rewalk (when all the database data is now in the new Storage Charset), Display Charset be change back to default (empty, which will still display in the new Storage Charset).

4.6.25 Top HTML and Bottom HTML

Syntax: HTML

This is static HTML to place at the beginning and ending of every search page respectively. It is useful for setting styles and displaying navigation menus and otherwise making the search pages look like the rest of your site.

Top and Bottom HTML when placed together should be exactly what is required to create a complete and valid HTML page. You can use your favorite HTML editor to create a page with a placeholder for the search form and results. Then cut and paste the section of HTML before the placeholder into the Top HTML and the section of HTML after the placeholder into the Bottom HTML.

If `$query` occurs within these fields, it will be replaced by the user's query.

CSS Stylesheet

Top HTML should always include a `<link>` to the CSS stylesheet `/webinator/common/search.css`, which is in the web server document root tree. This contains styles for hit-highlighting and other search functionality. While it may be edited to change these styles, it is more portable to add a separate stylesheet with any custom styles, included after `search.css`. That way, any future Webinator upgrades will upgrade stock styles as needed, but not affect any custom styles in separate files.

4.6.26 Enable Sherlock

Syntax: select Yes or No button

This informs the search to include comment tags in the results page to allow Sherlock to process the list.

Sherlock is a metasearch tool for Macintosh computers.

4.6.27 Top Best Bet Title

Syntax: text

This is the title text of best bets displayed above the search results. Common choices are "`Best Bets`" and "`Suggested Links`". See `Using Best Bets` 5.15 for more details.

4.6.28 Right Best Bet Title

Syntax: text

The title text of best bets displayed to the right of search results. Common choices are "`Best Bets`" and "`Suggested Links`". See `Using Best Bets` 5.15 for more details.

4.6.29 Top Best Bet Group

Syntax: choose group from drop-down list

This controls which group of best bets will be shown above the results. The group must already be created. See `Using Best Bets` 5.15 for more details.

4.6.30 Right Best Bet Group

Syntax: choose group from drop-down list

This controls which group of best bets will be shown to the right of the results. The group must already be created. See `Using Best Bets` 5.15 for more details.

4.6.31 Top Best Bet Box Color

Syntax: valid HTML color

This controls the color to be used for the background of the top best bet box. See `Using Best Bets` 5.15 for more details.

4.6.32 Right Best Bet Box Color

Syntax: valid HTML color

This controls the color to be used for the background of the right-side best bet box. See `Using Best Bets` 5.15 for more details.

4.6.33 Top Best Bet Border Style

Syntax: select from drop-down list

This controls the style of the top best bet box border. You can choose to have no border, a border around all the best bets, or an individual border around each result. See `Using Best Bets` 5.15 for more details.

4.6.34 Right Best Bet Border Style

Syntax: select from drop-down list

This controls the style of the right-side best bet border. You can choose to have no border, a border around all the best bets, or an individual border around each result. See `Using Best Bets` 5.15 for more details.

4.6.35 Right Best Bet Box Width

Syntax: enter number in text box

This controls the width of the best bet boxes shown to the right of the regular search results. See `Using Best Bets` 5.15 for more details.

4.6.36 Authorization Method

The `Authorization Method` setting controls what Results Authorization method(s) are used by Webinator when verifying user access to search result URLs. See the Results Authorization section (p. 82) for details. The possible settings are:

- `None`: No access verification; return all search results to all users. This is the default. It is also the setting that should be used for a Meta Search profile, even if one or more of its back-end profiles does use Results Authorization: the request and response for credentials will automatically be passed back and forth from front-end Meta Search to back-end profiles, which will handle the authorization (not the front-end).

- `Forward login cookies`: Webinator will forward login cookies from the user to the result URL. This is for custom HTML-form-based single-sign-on systems.

- `Basic/NTLM/file - prompt via form`: Webinator will prompt the user for their credentials with a form, then send them to the result URL via HTTP Basic, NTLM or Windows/SMB file authentication.

4.6.37 Login Cookies

For the `Forward login cookies` Results Authorization method, one or more cookies must be named in the `Login Cookies` setting. No values are given, as they will be obtained automatically on a per-search basis from the user.

When a user conducts a search, if the named cookies are seen from the user's browser, the user is assumed to be logged in, and the cookies are forwarded to the results URLs for authorization. If the named cookies are not seen, the user is assumed not to have logged in yet, and is redirected to `Login URL` instead.

4.6.38 Login URL

For the `Forward login cookies` Results Authorization method, if none of the `Login Cookies` are seen at search time, the user is assumed not to have logged in yet, and will be redirected to this URL instead. The `Login URL` should be the URL to the site's form-based login page.

After logging in, the site's login page can be configured to re-redirect the user back to their original search if desired. The special token "`%REFERER%`", if used in the `Login URL`, will be replaced with the URL back to the user's search. Thus, it could be assigned to a query-string variable in the `Login URL` so that the login page can redirect back to the search. Eg. with this value for the `Login URL`:

```
http://login.acme.com/login.asp?searchurl=%REFERER%
```

Webinator would redirect the user to `http://login.acme.com/login.asp`, with the `searchurl` variable set to the Webinator search page (with query). The `login.asp` code could be modified to redirect the user back to the `searchurl` query variable after login.

4.6.39 Basic/NTLM/file Cookie Type

For the `Basic/NTLM/file - prompt via form` Results Authorization method, this setting controls what cookie type to use for Webinator's copy of the user's credentials.

With `Basic/NTLM/file - prompt via form` set, when a user conducts a search for the first time, a form is presented (from Webinator) asking for a user and password. The user/pass is sent back to the user as a cookie from Webinator for use in future searches without having to re-prompt. The user/pass is also simultaneously used to validate search results via HTTP Basic/NTLM or Windows/SMB file access.

The `Basic/NTLM/file Cookie Type` setting controls whether this cookie from Webinator should be `Persistent` (retained permanently so the user does not have to login again) or `Session` (discarded after browser closure for security).

Note that the `Basic/NTLM/file Cookie Type` cookie is distinct from the `Login Cookies`; they are used for different access methods. The former originates from Webinator and is only ever sent to/from the user and Webinator: non-cookie-based access methods are then used from Webinator to the result URLs for actual authentication. `Login Cookies`, however, originate from a third-party form-based login system, and pass from the login server to the user to Webinator to the result URLs.

4.6.40 Login Verification URL

For the `Basic/NTLM/file - prompt via form` Results Authorization method, the user is directly prompted for a login by Webinator. Since authentication is handled by another server, when search results are denied access, Webinator cannot know if the denial is URL-based (lack of access by the user), or login-based (mistyped/wrong password).

To differentiate the two and give users a chance to correct mistyped passwords, a `Login Verification URL` may be set. This should be a URL that *all* users have access to, but that is still

protected (ie. anonymous users are denied). It should be an actual file (not a directory), preferably small (a few KB), and permanent (not likely to move, be renamed or have perms changed).

If `Login Verification URL` is set, Webinator will verify a user's prompted-for login by accessing this page. Since all users have access to it, a denial is assumed to mean the login was incorrect, and the user will be re-prompted for their credentials. Without a `Login Verification URL` set, a mistyped password will result in no search results, but the user will not know if they do not have access to the results, or they merely mistyped their password.

4.6.41 Unauthorized Result Query

For all `Authorization Method` types of Results Authorization, it is assumed a protocol-level denial will be issued when Webinator accesses URL(s) that a user does not have access too. Eg. for HTTP URLs, a `401 Unauthorized` message should be issued.

However, some servers may only issue a human-readable denial message, but otherwise return an ok (eg. HTTP `200`) protocol message. For such results Webinator will assume the user has access, and will erroneously return the result.

To remedy this, `Unauthorized Result Query` may be set to a query that will match only denied pages (eg. "`Access Denied`"). The `Field/Type` box should be set to the query type (substring vs. REX) and field (raw HTML vs. formatted text) for the search. The `Query` field is set to the actual substring or REX query.

Note that this setting imposes an extra search load, as each search result must be verified with a full-page GET instead of a HEAD, as well as queried against. Thus, `Unauthorized Result Query` should only be set if absolutely necessary.

4.6.42 Username Fixup

Username Fixup allows you to make modifications to the Results Authorization username provided, such as adding or removing a domain. This allows multiple back-ends with slightly different authentication schemes to be searched simultaneously in a Meta Search.

- `Search` - the search expression to match on the incoming username. Unless you're stripping off a domain, this should be left blank to match everything.
- `Replace` - the replacement string used to modify what was matched in the search. Please see examples below, or the `Replacement Strings` section of the Vortex manual on our website for the exact syntax.

For example, suppose you have a wiki and a file server. They use the same authentication back-ends, but the wiki takes the format `username` and the file server takes the format `DOMAIN/username`. If you create a profile for each of them and set the `Username Fixup Replace` value for the file server to `DOMAIN/\1`, then you can meta-search both with `username` and each will get the format it needs.

Examples

- Changing `username` to `MYDOMAIN/username`
  - Search - *(Empty)*
  - Replace - `MYDOMAIN/\1`
- Changing `MYDOMAIN/username` to `username`
  - Search - `>>=!/+/=.+`
  - Replace - `\4`
- Changing `MYDOMAIN/username` to `OTHERDOMAIN/username`
  - Search - `>>=!/+/=.+`
  - Replace - `OTHERDOMAIN/\4`

4.6.43 Max Docs to Auth-Check

This setting is the maximum number of raw (pre-auth-check) search result URLs to examine for authorized results, during results authorization. Decreasing this limit can speed up searches and reduce origin server load, at the cost of possibly truncated displayed results. Eg. noisy queries that match many overall documents on the server, but few of which are authorized for the search user, may use a lot of server resources, so reducing this limit may reduce that load.

The maximum value is -1 or blank (the default), for no limit: ie. continue until all results are checked, or `Successful Auth Result Limit` or `Total Auth Timeout` is reached.

4.6.44 Successful Auth Result Limit

This setting is the maximum number of authorized (displayable, post-auth-check) results to try to establish, during results authorization. Increasing this limit makes it more likely to get an exact hit count for a search (instead of a single page), at the expense of more search time and more origin server load.

The minimum (and default if empty) is the same as the `Results per Page` setting (p. 63), which produces a page of results the fastest. The maximum is -1 for no limit, ie. continue until all results are checked, or `Max Docs to Auth-Check` or `Total Auth Timeout` is reached.

4.6.45 Total Auth Timeout

This setting is the maximum total time in seconds to spend searching and authorizing results, during Results Authorization. The maximum setting value is -1 for no limit, ie. let `Search Timeout` (p. 77) cancel the search if reached. Any other negative value is relative to `Search Timeout`. Thus the default (if empty) of -5 means stop searching 5 seconds before `Search Timeout`, so that there are a few seconds left to send the results to the user.

4.6.46 Allow Authorization URL

If enabled, the `Authorization URL` field of each document is used for Results Authorization instead of the document URL. (If the `Authorization URL` field of a document is empty, or this setting is disabled, the document URL is used.) Enabling this can speed up searches under certain circumstances.

Sometimes an entire group of documents share the same authorization. For example, on some systems the contents of a directory always have the same authorization as the directory itself. In other words, every user's permissions on the files in any directory is the same as their permissions on the directory itself. If this is the case, then Results Authorization can authorize all results in the directory just by authorizing the directory itself, once. This reduction in calls speeds up searches.

For this optimization to be effective, the `Authorization URL` field in the database must be populated (see **Data from Field**, p. 39). For example, on systems where the contents of a directory always have the same authorization as the directory itself, `Authorization URL` should be set to the parent dir of each URL. The more files there are (on average) in a given directory, the more effective this optimization will be. Additionally, the **Authorization Caching** setting should be set to `Session`, so that the one-time directory authorization can be reused for each result inside the directory. (Otherwise Results Authorization must repeat the directory authorization for every result in the directory, as normal.)

The `Authorization URL` field may also be used on systems that do not meet the group-authorization criteria (many docs sharing the same authorization) detailed above. An environment may exist where the crawled/result URL is simply not the same URL that should be used for Results Authorization. For example, the crawl/result URLs may be `file://` URLs, yet the authorization should take place with `http://` URLs of the same host and path. In such a case, the `Authorization URL` field could be populated with the `http://` variant to tell Results Authorization to use those URLs. In this instance, the field is being used to properly authorize URLs, and will not necessarily speed up searches (because the `Authorization URL`s are unique and not shared across groups).

4.6.47 Authorization Caching

Whether and how to cache Results Authorization traffic. The default of `None` does no caching. When set to `Session`, Results Authorization traffic is cached for the duration of the session, i.e. that search alone. Normally caching is of little benefit, because authorization URLs are typically the same as result URLs, and the latter are typically unique in a given search; thus caching will not help. However, if the `Authorization URL` field is populated, and **Allow Authorization URL** is enabled, enabling caching may speed up Results Authorization searches. See **Allow Authorization URL** (p. 72) for details.

4.6.48 Debug Results Authorization

Enabling this setting causes copious debugging information to be logged. It should only be enabled at the request of Tech Support for diagnosing Results Authorization problems.

4.6.49 Show Authorization Info

Enabling this causes details about the ResAuth process to be displayed on the search results page - which URL are being attempted, what the outcome is, how long it takes, etc. This can assist in troubleshooting why results aren't displaying when expected.

- `None` *(default)* - No information is displayed.
- `Admin Users Only` - information is displayed only if the browser is currently logged in to the admin interface. This allows admins to troubleshoot ResAuth without exposing information to all users.
- `All Users` - information is displayed for all search users.

WARNING - The information shown includes info about URLs that search users don't have access to (explaining how/why they failed). The Webinator acknowledging the existence of these URLs when they're unauthorized could be considered a security breach in some scenarios.

It is recommended to only set it to `Admin Users Only` when troubleshooting, and then set it back to `None` when no longer needed.

4.6.50 Enable Spell Check

Syntax: select Yes or No button

This turns on the spell check option. With this option on, any search which produces no results displays a list of alternate-spelling queries, which will produce more results. If a query produces one result, Webinator suggests other words similar in spelling to the words you entered. The suggestions are based on the actual walk database, so unusual spellings or terminology used on your site are picked up by the spell-checker. The number of suggestions varies, depending on the `Suggest Time Limit` and `Number of Suggestions` options. The default is on.

4.6.51 Suggest Time Limit

Syntax: choose from drop-down list

This controls the number of seconds Webinator allows for spelling suggestions to be made. See also `Enable Spell Check` 4.6.50 for more information.

4.6.52 Number of Suggestions

Syntax: choose from drop-down list

This controls the number of spelling suggestions offered. See also `Enable Spell Check` 4.6.50 for more information.

4.6.53 Synonyms

Syntax: choose from drop-down list

This allows you to select a level of equivalence matching. You can limit results to specific matches, or you can allow synonyms and phrases. The values are described as follows:

`Disabled`: no phrase recognition and no synonyms (equivalences). Only searches for the the actual terms in a query. This is regardless of ~ usage.

`Phrase recognition only`: recognize query word groups that are known phrases and search for them as phrases.

`Phrases & Allow synonyms`: phrase recognition plus allows the tilde ( ) operator to match synonyms on specific query terms

`Phrases & Use synonyms by default`: phrase recognition and matching synonyms on all query terms (tilde to turn off on specific terms).

See also `Using the Thesaurus` (section 5.3).

4.6.54 Main Thesaurus

Syntax: the symbolic name for the primary thesaurus

Here you can select a main thesaurus. A drop-down list allows you to select one of the thesauri that was defined in `Maintenance`, `Custom Thesaurus`.

See also `Using the Thesaurus` (section 5.3).

4.6.55 Secondary Thesaurus

Syntax: the symbolic name for the secondary thesaurus

Here you can select a secondary thesaurus. A drop-down list allows you to select one of the thesauri that was defined in m`Maintenance`, `Custom Thesaurus`.

See also `Using the Thesaurus` (section 5.3).

4.6.56 Translate Boolean

Syntax: select Yes or No button

Off by default. If on, Boolean keywords `and`, `or`, and `not` in the search query will be translated into set logic.

Webinator uses set logic internally, and this setting translates basic boolean statements into proper set logic automatically. This is a limited translation, and does not support nesting of statements.

For more information on Webinator's use of set logic, please see the `Using Set Logic to Weight Search Items` section of the Texis manual on our website.

4.6.57 Allow the @ Operator

Syntax: select Yes or No button

Off by default. If on, allow use of the `@` (intersections) operator in queries. Queries with few or no intersections (eg. `@0`) may be slower, as they can generate a copious number of hits.

4.6.58 Allow Linear

Syntax: select Yes or No button

Off by default. If on, an all-linear query –one without any indexable "anchor" words– is allowed. A query like "`/money #million`", where all the terms use unindexable pattern matchers (REX, NPM or XPM) is an example. Such a query requires a linear search of the entire table, and this can be very slow for a table of significant size.

If `allinear` is off, all queries must have at least one term that can be resolved with the Metamorph index, and a Metamorph index must exist on the field. Under such circumstances, other unindexable terms in the query can generally be resolved quickly, if the "anchor" term limits the linear search to a tiny fraction of the table. The error message "`Query would require linear search`" may be generated by linear queries if this is off.

4.6.59 Allow NOT Logic

Syntax: select Yes or No button

On by default. If on, allows "NOT" logic (eg. the `-` operator) in a query.

4.6.60 Allow Post-Processing

Syntax: select Yes or No button

Off by default. If on, post-processing of queries is allowed when needed after an index lookup, eg. to resolve unindexable terms like REX expressions, or only partially indexable terms. If off, some queries are faster, but they may not be as accurate if they aren't completely resolved. The error message "`Query would require post-processing`" may be generated by such queries if this is off.

4.6.61 Allow Wildcards

Syntax: select Yes or No button

On by default. If on, wildcards are allowed in queries. Wildcards can slow searches somewhat because potentially many words must be looked for.

4.6.62 Allow Leading Wildcards

Syntax: select Yes or No button

Off by default. If on, leading wildcards ("`*word`") are allowed in queries. **Allow Wildcards** must also be enabled. Note that leading-wildcard terms are significantly slower to search for than trailing-wildcard terms such as "`word*`".

4.6.63 Single-Word Wildcards

Syntax: select Yes or No button

On by default. If on, wildcard searches will span only one word in the text – instead of up to 80 characters across words – and will suffix-match. Eg. the query "`con*tion`" will match "`condition`" but not "`consider my position`" nor "`conditionally`".

4.6.64 Allow WITHIN Operators

Syntax: select Yes or No button

Off by default. If on, "within" operators (`w/`) are allowed. These generally require a post-process to resolve, and therefore they can slow searches. If off, the error message "`'delimiters' not allowed in query`" will be generated if the within operator is used in a query.

4.6.65 Require All Words

Syntax: select Yes or No button

By default, all words a user searches for must be in the result for it to match. If `Require All Words` is changed to `N`, a result will be shown if *any* of the query terms are in the result.

Results that match multiple words will be ranked higher than results that match fewer.

4.6.66 Resolve Phrase Noise Words

Syntax: select Yes or No button

Off by default. This indicates whether to exactly resolve the noise words in phrases. If on, a phrase such as "`state of the art`" will only match those exact words; however, this may require post-processing to resolve (potentially slower). If off, any word is permitted in place of the noise words, and no post-processing is needed; this is faster but potentially less accurate.

4.6.67 Keep Noise Words

Syntax: select Yes or No button

Off by default. This indicates whether to keep noise words (`Yes`) in the query during query processing and search for them, or remove them (`No`, the default) from the query and ignore them.

4.6.68 Noise List

Syntax: whitespace separated list of noise (stop) words

A list of words to be ignored in queries (if `Keep Noise Words` is No). If empty the default list will be used, which is:

a about after again ago all almost also always am am an and another any anybody anyhow anyone anything anyway are are as at away back be became because been before being between but by came can cannot come could did do does does doing done down each each else even ever every everyone everything everything for from front get getting go goes going gone got gotten had has has have have having he her here him his how i if in into is is isn't it just last least left less let like make many may maybe me mine mine more most much my my myself never no none not now of off on one onto or our ourselves out over per put putting same saw see seen shall shall she she should should so some somebody someone something stand such sure take than that the their their them them then there these they this this those through till to too two unless until up upon us us very was was we went were were what what's whatever when where whether which while who who whoever whom whose whose why will will with within without won't would wouldn't yet you your

4.6.69 Search Timeout

Syntax: integer number of seconds

This is the maximum overall time to spend searching and sending results. Exceeding this limit, whether due to server load, network slowness, etc. will result in a "Timeout" message to the user. This helps prevent heavy load from overwhelming the server. The default (if empty) is 30 seconds. The maximum is -1 for no limit, which is strongly discouraged.

4.6.70 Show Error Messages

Syntax: select box

Show Error Messages determines the disposition of error messages during searches. It may be set to one of the following values:

- `None`
  Suppress all errors
- `In HTML comments`
  Show errors in HTML comments (for HTML result styles) so that they are not normally visible to the user, but can be viewed via View Source in a browser. In XML result styles, errors will be suppressed.

- `In HTML comments & query errors visible`
  Show errors in HTML comments (for HTML result styles), but show query-related errors (eg. "`Your query was all noise words.`") visibly (in grey boxes).

The default is `In HTML comments & query errors visible`. Note that in admin (test search) mode, all errors are always shown visibly, for admin perusal.

4.6.71 Debug SQL Level

Syntax: integer number or empty/0 to disable

Setting Debug SQL Level to a non-empty/non-zero value (typically 3) enables extra debug messages for certain SQL statements. Generally only set at the request of tech support for diagnosing problems.

4.6.72 Fast Result Counts

Syntax: select Yes or No button

Off by default. Some complex queries involving categories or proximities closer than page can take more time to determine exact result hit counts. In some cases it may cause timeouts. Enabling this option will determine hit counts much faster, and using less CPU, in these cases at the expense of accuracy. The hit counts for complex queries will generally be overestimated (it will say there are more hits than there really are).

4.6.73 Proximity

Syntax: choose from drop-down list

Proximity gives the ability to locate answers with greater precision. The Webinator input form gives you several options to control the search proximity:

line All query terms must occur on the same line

sentence Query items must all reside within the same sentence

paragraph Within the same paragraph or text block

page All items must occur within same HTML document (the default)

A bar-graph display will be shown any time a ranking search was performed (eg. all searches except `Show Parents`).

4.6.74 Language Characters

Syntax: list or range of characters, as inside REX `[]`

The **Language Characters** setting controls what characters constitute a language query. Query terms composed entirely of these characters are considered language terms, and have **Word Forms** processing applied. The syntax is a list of characters (no separation), and/or a range of characters; the same as a REX character class (without the brackets). The default is `\alpha\'\x80-\xFF`, ie. alphabetic, hi-bit (for UTF-8) and apostrophe (for contractions). For best results, all characters that could match part of a **Word Definition** expression (p. 48) should usually also be listed in **Language Characters**.

4.6.75 Word Forms

Syntax: choose from drop-down list

The `Word forms` options give you control over how many variations of your query terms are sought in your search as follows:

Exact: Only exact matches are allowed. (the default)

Plural & possessives: Plural and possessive forms are found. (`s`, `es`, `'s`)

Any word forms: As many word forms as can be derived are located.

Custom: use the three custom settings below to determine word forms.

4.6.76 Custom Suffix List

Syntax: Space-separated list of suffixes

When using the Word Forms `Custom`, this is the space-separated list of suffixes to use. All of these will be repeatedly stripped off of words, as long as the word is longer than the `Custom Suffix Min Length`.

An example setting could be `s es ' a e i y`. For the word `smith's`, the `s`and `'` would be stripped, causing it to match `smith`, `smiths`, etc.

4.6.77 Custom Suffix Default Removal

Syntax: Y or N

When using the Word Forms `Custom`, this controls whether to remove a trailing vowel, or one of a trailing double consonant pair, after normal suffix processing is finished. This will not apply if it would take the word below the minimum word length.

For example, if `ing` is in the suffix list and `Default Removal`is `Y`, then the word `running` will have the `ing` stripped, and then the 2nd `n` will be removed via `Default Removal`, producing `run`.

Default Removal is set to `Y` when using `Plurals & Possessives` and `Any Word Forms`.

4.6.78 Custom Suffix Min Length

Syntax: Number

When using the Word Forms `Custom`, Webinator will not try to strip additional suffixes from any word shorter than this length. For example, if min length is `3`or more, the `es` on `yes` will not be treated as a suffix.

Min Length is set to `3` when using `Plurals & Possessives`, and `5` for `All Word Forms`.

4.6.79 Word Ordering

Syntax: choose from drop-down list

Controls how important word order is for results ranking: hits with terms in the same order as the query are considered better. For example, if searching for "`bear arms`", then the hit "`arm bears`", while matching both terms, is probably not as good as an in-order match. The default weight is Medium (500).

4.6.80 Word Proximity

Syntax: choose from drop down list

Controls how important proximity of terms is for results ranking. The closer the hit's terms are grouped together, the better the rank. The default weight is 500.

4.6.81 Database Frequency

Syntax: choose from drop down list

Controls how important frequency in the table is for results ranking. The more a term occurs in the table being searched, the *worse* its rank. Terms that occur in many documents are usually less relevant than rare terms. For example, in a web-walk database the word "`HTML`" is likely to occur in most documents: it thus has little use in finding a specific document. The default weight is 500.

4.6.82 Document Frequency

Syntax: choose from drop down list

Controls how important frequency in document is for results ranking. The more occurrences of a term in a document, the better its rank, up to a point. The default weight is 500.

4.6.83 Position in Text

Syntax: choose from drop down list

Controls how important closeness to document start is for results ranking. Hits closer to the top of the document are considered better. The default weight is 500.

4.6.84 Clicks from Home

Syntax: choose from drop down list

Controls how important being close to a Base URL is for results ranking. The more times the walk had to click on links to get to the page, the lower weight it will have. The default weight is off, ie. do not factor in clicks-from-home for results ranking.

4.6.85 Ranked Rows

Syntax: number

The maximum number of rows that can be scrolled to when returning ranked results. This can be set to 0 for all matching rows, or to any other number. The lower the number the better the performance, however users won't be able to scroll through as many results. The default is 200.

4.6.86 XML Export Variables

Syntax: names separated by newlines

`XML Export Variables` is a list of variables, one per line, that are to be displayed and propagated through XML search results. For example, if `cbtGroup` is specified in `XML Export Variables`, and the search query includes `...\&cbtGroup=user\&cbtGroup=backup...`, then the following block will appear in XML output, after all the `<Result>` tags:

```
<exportVar>
  <variable name=‘‘cbtGroup’’>user</variable>
  <variable name=‘‘cbtGroup’’>backup</variable>
</exportVar>
```

This setting only applies if `Results Style` is set to `XSL Stylesheet`.

4.6.87 Phishing Protection

Phishing Protection prevents Webinator from being used as a tool in a phishing attack.

Webinator has a redirect page as part of its Query Logging functionality, where it will provide a redirect to the URL specified. It would be possible for an attacker to specify a URL that, at first glance, looks like a link from Webinator, which the user may trust. After the redirect, it actually ends up somewhere else.

If Phishing Protection is enabled, the redirect page will make sure that any URL specified is actually in the profile's walk database before issuing the redirect to it.

4.6.88 Decode Displayed URLs

Decode Displayed URLs will cause the URL that is displayed in search results to be URL-decoded, which includes replacing sequences with their proper characters.

This can be useful when URLs have words separated with spaces, which are replaced with `%20` to be a valid URL. Decode Displayed URLs allows you to display the decoded version, making the files easier for search users to read.

`"this%20is%20a%0file.txt"` becomes `"this is a file.txt"`.

4.6.89 Visible

This controls whether this profile is visible to other Webinators (or even the same one) for use in a meta search. Any profile that is to be used as a part of a meta search must have the `Visible` flag set to `Y`.

If a profile has `Visible` set to `N` and is used as a back-end for meta search, it will return the error `Profile not Visible`.

4.7 Results Authorization

Results Authorization allows restriction of search results to authorized users only, on a per-URL basis. Only users with access to a given URL will ever see that URL in a result list, instead of all users seeing all matches (and potentially being denied access to results already shown).

Access to a URL, as well as the namespace of users, is determined by the URL's origin server, not Webinator, so no reconfiguration of users or access is needed – the pre-existing server access controls are just forwarded by Webinator. And since access is determined on a per-result, not per-search, basis, a single profile can serve a multitude of users with any combination of whole/partial access to the underlying data.

Results Authorization works at search time (late binding) by accessing each potential search result URL with the user's credentials. Only URLs authorized to that user are then shown in search results. The authentication method(s) used will depend on the existing system(s) already used by the indexed URLs. Various schemes are supported:

- `None`: No access verification; return all search results to all users. This is the default.
- `Cookie-based`: Custom HTML-form-based single-sign-on systems. Users first login on a web server (not a Windows workstation login), which then sends an access cookie to the user's browser. This cookie is automatically returned to the server when accessing future pages, and grants the user access.
- `Basic`: HTTP Basic authentication, for web servers.
- `NTLM`: Windows NTLM authentication, for web servers.
- `SMB/Windows`: SMB for Windows file servers (for Thunderstone products that support `file://` crawling).

For cookie-based systems, Webinator will merely forward the cookies the user has already received from the site login page. For all others (Basic/NTLM/SMB), Webinator must prompt for the user and password directly, as they are needed to verify result URLs. In the latter case, credentials will then be stored in a cookie by Webinator so that future searches do not need to re-prompt for a login. Note that NFS-mounted file servers are not currently supported by Results Authorization, due to limitations of NFS.

4.7.1 Results Authorization Crawl Settings

Webinator itself needs read access to the entire set of URLs in order to build a search index. Therefore, before walking a protected data set for Results Authorization, it may be necessary to fill out the `Login Info` setting (p. 52) under `All Walk Settings` with a full-access admin type account, so that Webinator can crawl the data.

Or it may be necessary to fill out a `Primer URL` (p. 51) containing login info to submit to a site's login form, so that Webinator can obtain the login cookies needed for access to the rest of the site.

4.7.2 Results Authorization Search Settings

After a successful crawl, Results Authorization is configured with the `Results Authorization Options` group on the `Search Settings` page. The primary setting is `Authorization Method` (p. 68), which is determined by the authentication system(s) in use by the indexed URLs. If cookie-based, this is set to `Forward login cookies`; for all other systems, it is set to `Basic/NTLM/file - Prompt via form`. Most of the remaining settings depend on which method was selected; see the `Authorization Method` setting (p. 68) for details.

There are also a few resource/tuning settings, such as `Max Docs to Auth-Check`, `Successful Auth Result Limit`, `Total Auth Timeout`, and `Debug Results Authorization`, which are not required, but merely fine-tune the results.

4.8 Meta Search - Search multiple profiles as one

Meta search allows you to search multiple profiles simultaneously and merge and display the results as if it was one big profile. The meta search can search and combine profiles from multiple Webinators.

4.8.1 Profile Creation

When creating a profile, change the `Standard` select box to `Meta Search` instead.

4.8.2 Meta Search Walk Settings

`Walk Settings` is somewhat of a misnomer for a meta search profile since it doesn't do any walking of it's own. On this page you list the host(s) and profile(s) to search and merge when this profile is accessed.

For each profile you want included in the search, list the full URL to that machine's `search` interface, e.g. e.g. `http://searchbox/texis/webinator/search/`. You can use `localhost` or `127.0.0.1` instead of the hostname when using profiles on the same machine as the meta search.

The `Display Name` column is used to provide a user friendly name for this profile that will be displayed if the user is allowed to choose which profiles to search.

The `Status` column shows the status of the remote profile once a host/profile has been entered and `Update` has been pressed. If the target is searchable, `OK` is displayed. Otherwise, text explaining the error is displayed. Refreshing the page re-queries the target profiles.

If `User Selection` is set to Y then the user will be presented with a list of `Display Names` and can choose which ones to search. Leaving them all unchecked will cause them all to be searched.

The `Meta Mode` setting controls whether profiles on the same host will be searched serially or in parallel. "Sameness" of host is determined by the `Target Search URL` setting, so using different names or a name and an IP address will allow you to mix serial and parallel.

The `Results Merge Method` setting controls how target profiles' results are merged and sorted by the meta profile. Two methods are available:

- `Requested order`
  The results will be sorted as requested, i.e. as specified by the `order` query-string variable (or if that is unset, the meta search `Result Order` search setting). This is the default. Thus, results from different target profiles may or may not be mixed together (depending on how they sort by `order`).
- `Target profile order`
  The results will be sorted by their `Profiles` setting order first, then by requested (`order` variable) order. This will result in *all* results of the first target profile being shown first, then all results of the next target profile, etc.

Note that in both cases, the *target* profiles still individually sort their results according to requested `order`. The `Results Merge Method` setting only affects how those top results are then merged and sorted by the *meta* profile.

All servers with profiles listed as targets of a meta search must have the IP address of that meta search server listed in the `Cluster Members` setting under `Maintenance->System Wide Settings`. Otherwise meta searches will return the error "Profile not Visible".

4.8.3 Search Settings

The appearance options control the appearance of the meta search results pages. Currently the Results Authorization and query options of the meta profile do not apply: use the target profiles' options instead.

When using best bets the meta search profile must have the same group names as the backend profiles. Any best bets from the backends that have group names that are not defined in the meta profile will not be shown.

Query logging of the meta search and the backends are independent of each other. The meta search will respect its own query logging setting as will each of the backend profiles. So it is possible to have multiple logs for the same query if both the meta search and the backend have query logging turned on.

4.9 Access Control

Access Control allows different administrative users to be given different levels of access to Webinator; normally, with access control off (the default) all users have access to all administrative functions. Access Control can only be enabled or disabled by the `webinator` user, on the Maintenance page.

Access Control is supported in the full Texis product, but not Webinator-only.

4.9.1 User Groups

User groups allow easier access control maintenance, as users with similar permissions can be administered together once rather than separately several times. The special group `Everyone` always exists and cannot be edited; it always contains all users as a convenience.

User groups may contain other groups as well as users, allowing complex hierarchies to be created if needed. Permissions for a user are affected by all groups a user is directly or indirectly a member of. For example, if user `Amy` is in group `Programmers`, and group `Programmers` is in group `IT`, then `Amy` is also indirectly a member of `IT`, and her permissions are affected by those granted to not only herself and `Programmers` but `IT` as well.

4.9.2 Object hierarchy

Each administrative action that can be access-controlled (eg. editing walk settings, creating accounts) can be thought of as an object. Some actions are broader than others and can be thought of as a superset, eg. editing *all* profiles is a superset of editing a *specific* profile. Thus, access control objects are arranged in a tree-like hierarchy, where each object has a parent object, and can inherit permissions from it. This makes setting privileges on a logical group of objects (eg. all profiles) easier, as only one object may need to be changed (the parent). Also, when new child members (eg. new profiles) are created, they will inherit the same privileges automatically. The access control object hierarchy in Webinator is as follows:

```
/                       Global root object
   Users/                  User accounts
      webinator               webinator user
      ...                     Other users
   Groups/                 User groups
   Profiles/               Profiles
      default                 default profile
      ...                     other profiles
   Settings/               Profile settings
   Maintenance/            Maintenance page
      Info/
      Updates/
```

Note that these “files” do not really exist: the objects are merely symbols representing actions that can be access-controlled.

4.9.3 Access Control Lists

An object may have an Access Control List (ACL) associated with it. ACLs determine what rights (Read/Write/Delete/Change perms) users have on objects. Each object's ACL contains one or more Access Control Entries (ACEs). An ACE identifies a trustee (a user or group), a set of rights, and whether those rights are allowed or denied the trustee on that object. In addition to the ACL explicitly set on an object, rights may be inherited from parent objects' ACLs, as mentioned above.

4.9.4 Determining Effective Rights

The effective rights a specific user has on an object – what the user can actually do with the object – are determined by examining ACEs in a specific order. The first ACE that matches both the user and the desired access right determines whether the user has that right on the object. An ACE matches the user if it specifies the user or any group the user is directly or indirectly a member of. An ACE matches the desired right if the right is listed in the ACE.

ACEs are examined in the following order[1]:

1. ACEs explicitly set on the object
2. ACEs explicitly set on the object's parent
3. ACEs explicitly set on the object's further ancestors, nearest ancestor first

At each object, ACEs are checked in ACL order (the order displayed for an object on the Access Control page). Order can be changed among multiple ACEs on the same object by using the `up arrow` and `down arrow` buttons next to the ACEs.

If no matching ACE is found after all levels are examined (back to the root or `Global` ACE), access is allowed by default (this is for back-compatibility with non-ACL mode).

4.9.5 Required Rights for Admin Actions

Certain ACL rights are required for certain administrative actions to be performed. In order to maximize rights-configuration flexibility, some actions require rights on multiple objects. For example, editing settings on a profile requires rights not only on the profile, but also on the setting itself. Note in the object hierarchy (p. 85) that profiles and settings are two "sibling" branches, rather than settings being replicated as descendants of every profile. Thus, profiles and settings can be thought of as a two-dimensional grid for permissions, and a user's rights can be tailored across that grid: access to one setting across all profiles, access to all settings one profile only, etc.

The rights needed for specific actions are listed below. If a user does not have the required rights for an action, either a red `Access denied` message will be displayed, or (if access still granted to other parts) the affected object may simply not appear (read access denied), or may appear grayed out (write access denied). For more information and some example permission schemes, see the Using Access Control section, p. 111.

[1] In versions 5.3.0 and earlier, deny ACEs were always required to be before allow ACEs for an object.

Walk and Search Settings

For settings under Basic, All Walk, and Search Settings, a user must have read access to the profile as well as read access to the specific setting in order to see the setting. Write access to the profile, and write and delete access to the setting, is needed in order to modify a setting. (Delete is needed to clear a setting, which may not be apparent from the form.) Note that some settings are grouped on a line, such as the `Enterprise` setting: permissions can be granted to the group as a whole (`Enterprise`), or only specific settings in the group (`Enterprise - Yes` or `Enterprise - Domain`). If a user has no read access to a setting, it will not be displayed on the page. If a user has no write access to a setting, it will be disabled (grayed out and not modifiable).

Starting and stopping a walk

Write access to the profile and write access to the `Walk now` setting is required to start a walk. Write access to the profile and write access to the `Stop walk` setting is required to stop a walk.

Best Bets

Write access to the profile and write access to the `Best Bet Groups` setting is needed to modify the Best Bet Groups for a profile, or to modify Best Bet words for a specific URL (under List/Edit URLs). Note that this is distinct from editing Best Bet *search* settings (eg. `Top Best Bet Title`), which only affect search, not the walk itself.

List/Edit URLs

Write access to the profile and write access to the `List/Edit URLs` setting is needed to modify URLs in the database, including using the `Update Soon` link. Read access to both is needed to view URLs.

List Duplicates

Read access to the profile and read access to the `List Duplicates` setting is needed read the error table and list the duplicates of a URL.

Walk Status

Read access to the profile and read access to the `Walk status` setting is needed to view Walk Status.

Query Log

Read access to the profile and read access to the `Query log` setting is needed to view the Query Log.

Profiles

Read access to the profile and read access to the desired setting(s) are needed to view the given setting. Write access to both is needed to modify a setting. Delete access to the profile is needed to delete the profile. Write access to `All Profiles` (the parent of profiles) is needed to create a new profile.

Accounts

Write access to `All Users` is needed to create a new user. Write access to the user is needed to change the password for a user. Delete access to the user is needed to delete a user.

User Groups

Write access to `All Groups` is needed to create a new group. Write access to the group, as well as write access to each member being added or removed, is needed to add or remove members to or from a group (except where the group is only indirectly being modified due to a member itself being deleted). Delete access to the group is needed to delete a group.

Access Control

Change-perms access to an object is needed in order to create, edit or delete an ACE on the object.

Maintenance

Read access to `Info` under `Maintenance` is needed to read the Information links. Write access to `Updates` under `Maintenance` is needed to install or upgrade software.

4.10 Running the Walker by Hand

4.10.1 Using dowalk

Normally a walk is initiated from the administrative interface. There may, however, be times when it is desirable to start a walk by hand from a shell (or command) prompt or as a part of some other automated task. When the administrative interface starts a walk it shows you the command line to use It is of the form:

```
texis profile=PROFILENAME dowalk/dispatch.txt
```

You may also specify the parameter `ttyverbose` to be 1, or higher, to tell `dowalk` to print various status messages to the screen when being run by hand. The form would be

```
texis profile=PROFILENAME ttyverbose=1 dowalk/dispatch.txt
```

Where `PROFILENAME` is the name of the profile you have configured using the administrative interface. You will need to supply the full path to texis if it is not in your PATH. You will also need to supply the path to the `dowalk` script if it is not in the current directory when you run the command.

```
INSTALLDIR/bin/texis profile=PROFILENAME□↝
   ↪INSTALLDIR/texis/scripts/webinator/dowalk/dispatch.txt
```

or

```
INSTALLDIR\texis profile=PROFILENAME□↝
   ↪INSTALLDIR\Texis\Scripts\Webinator\dowalk/dispatch.txt
```

The walker will behave the same as it does from the administrative interface. Walk info will be logged to the same files. See section 6.1.

There are several other "entry points" that can be used to get various different behaviors when starting the walker. They all take the same form as `dispatch` above except that `dispatch` is replaced by the name of the entry point. The entry points are:

- `dispatch`
  Start a complete new walk.
- `hold`
  To stop a walk that is in progress, create/update the search indices and make it the live search.
- `stop`
  To stop and abandon a walk that is in progress.
- `indexmakelive`
  To create/update the search indices on an abandoned walk and make it the live search.
- `refreshnow`
  To force soonest refresh of a particular URL. This requires an extra `u=THEURL` argument to tell it what URL to refresh. This will flag the page for refresh on the next refresh check. It will not refresh anything itself. So you need to have walk type set to refresh and a schedule set.
  `texis profile=PROFILENAME u=THEURL dowalk/refreshnow.txt`
- `ifmodified`
  Checks the `Watch URL`. If the watched page has changed a walk is started. If not no action is taken. This is generally used on a frequent schedule to automatically rewalk a site if it changes.
- `singles`
  Fetches and indexes any single pages specified in the profile that are not yet in the database. You would call this after adding adding to `Single Page`, `Page File`, or `Page URL`.
- `refresh`
  Start a "refresh" walk. This walk will check all pages already in the database and download only changed ones. Missing pages will be deleted. New pages discovered on modified pages will be added.
- `recat`
  Recategorize the database based on the current settings of `Categories`.

- `reindex`
  Drop and recreate the Metamorph index on the html table. This would be used after changing the `Word Definition` expressions.

- `updateindex`
  Update the Metamorph index on the html table. This would be used after performing manual sql operations against the html table.

- `remakeindex`
  Drop and recreate all (standard) indices on the database. This has little use except in the case where indices got corrupted by disk errors or such.

- `checkandbuild`
  Ensure that the proper search index exists for the search fields selected in the profile. Wouldn't generally be called except internally when the desired fields to search are changed.

- `tsverrors`
  Dumps the error table as tab separated values of Date, Url, Reason. Optional `start` and `end` date-times may be specified. Not specifying `start` means start at beginning. Not specifying `end` means continue to end. `texis profile=PROFILENAME start="2004-10-01"`□↪
  ↪`end="2004-11-01" dowalk/refreshnow.txt`

- `convert`
  The entry point `convert` has a different syntax than the others.

  `texis v2db=DB v2profile=PROFILE v4profile=PROFILE`□↪
  ↪`dowalk/convert.txt`

  It is used to convert Webinator 2 profiles to Webinator 4 profiles (as well as possible). Set `v2db` to the full path to the existing Webinator 2 database containing the profile to convert. Set `v2profile` to the name of the Webinator 2 profile in the specified database to convert. Set `v4profile` to the name of the new Webinator 4 profile to create in the global database.

  A walk is NOT started. After conversion you would select the new profile, make any adjustments or fixups, then start a new walk.

4.11 Running the Search Interface

See section 5.1, p. 97.

4.12 Maintenance

The Maintenance menu has the following structure. Each item is described in the pages that follow.

```
Information
   Thunderstone Information
Install/Upgrade
   Apply a License
System Settings
   System Wide Settings
   View/Edit Access Control Lists
   Enable or Disable Access Control Lists
   Custom Thesaurus
   Save Webinator settings
   Restore Webinator settings
   Test Network and Servers
   Advanced Support Tools
```

4.12.1 Information

The `Information` group provides links to a variety of information useful for monitoring the system and performing maintenance.

Thunderstone Information

This page provides Thunderstone software version numbers, Webinator serial number, Thunderstone contact information, and license information.

4.12.2 Install/Upgrade

The `Install/Upgrade` section provides links to pages for installing and upgrading software.

Apply a License

This allows a license update (obtained from Thunderstone) to be applied to Webinator, to upgrade features or increase limits. A form is provided to accept the license and verify credentials. Only the `webinator` user can apply a license update, and that account's password must be given again on the form for security. The license is provided either by selecting the file it was saved to via the "`License from file`" box, or cut-and-pasting it (eg. from an email) into the "`Or copy license text`" box.

Hit "`Apply License`" to apply the license. If the license was installed successfully, the message "`License applied successfully`" will be shown. If it could not be installed, an error message will be shown.

Note: A full install or upgrade (not just scripts) to Webinator version 6 or later is required for this feature. Also, it must have been enabled in `conf/texis.ini` via the `[License Update] User` setting (enabled by installer).

If the license could not be installed, the alternate method is to copy it to a file named `license.upd` in the install directory (typically `/usr/local/morph3` under Unix or `C:\Program Files\Thunderstone Software\Webinator` under Windows), then from a command prompt, cd to the install dir and run `texis -update` (Windows) or `bin/texis -update` (Unix) to apply the license.

Some typical errors that might occur when installing a license from the web form include the following:

- `License update unauthorized`
  The password may be incorrect, or the currently-logged-in user may not be permitted to update licenses (usually must be `webinator`).
- `Invalid license`
  The license supplied was invalid or unacceptable. Make sure it is a valid Webinator license, not a third-party license (eg. for the Language Analysis Module). Make sure it was cut-and-pasted cleanly from any email message it was embedded in.
- `Service Not Enabled`
  License updates via the web admin interface were not enabled at install/upgrade time. Use the command-line interface (above).
- `Secure Connection Required`
  A secure (SSL) connection from Webinator to the Texis Monitor process (which manages licenses) could not be established. Use the command-line interface (above).
- `Internal error`
  An internal Texis Monitor error occurred. Use the command-line interface (above).

4.12.3 System Settings

This area is for settings that affect Webinator as a whole and/or may be shared by multiple walk profiles.

System Wide Settings

Cluster Members

This field defines the machine(s) and/or network(s) that constitute a cluster of Webinators. If you have more than one, all of their IPs or a network prefix and wildcard (such as `10.10.10.*`) should be specified here. All machines matching these IPs will be allowed full access to Webinator internals without verification. This allows for replication and meta searching.

API Logging

Allows you to record the XML requests & responses of all dataload and SOAP admin API calls to `api.log` in the logs directory. This can be useful when troubleshooting why dataload requests aren't storing properly.

Dataload and replication are supported in the full Texis product, but not Webinator-only.

Disable All Walks

When this setting is on, no walks will launch for any profiles for any reason (manual, schedule, etc). Setting to `Y` will stop ALL profiles from walking, overriding any individual profile's `Disable Walks` setting.

This can be useful with machines that should be dataload-only, or for machines that want to guarantee their content won't change.

Log All Replication

Writes information for each replication queue processor to `replication.log`. This forces logging for all profiles, and also for non-profile, System data replication.

If both "Log All Replication" and a profile's "Log Replication" are set, logging for that profile will be the more verbose of the two.

Replication is supported in the full Texis product, but not Webinator-only.

Enable/Disable Access Control Lists, View/Edit Access Control Lists

This option turns on/off ACLs for accessing Webinator.

ACLs are supported in the full Texis product, but not Webinator-only.

The default permission scheme for managing Webinator is very basic and all accounts have full admin privileges. ACLs allow very fine grained control over which administrators can access which features and settings.

See p. 85 for details about access control lists.

Custom Thesaurus

This area allows you to upload one or more custom thesauri (synonym lists) for use by search profiles. An uploaded thesaurus is compiled and kept on Webinator. There is no way to download a thesaurus once uploaded so it's a good idea to keep a copy around in case you want to make modifications later on.

Each thesaurus may be used by zero or more profiles and should not be deleted if it is in use by a profile. Search options that affect the use of these thesauri are `Synonyms`(4.6.53), `Main Thesaurus`(4.6.54), and `Secondary Thesaurus`(4.6.55).

See section 5.3 for further details.

Save Webinator Settings

This allows you to save all of the current profile and most of the system settings from Webinator to an XML file on your local workstation. This file can be used to aid in cloning Webinators for a cluster and as a backup in the event the machine needs to be restored from scratch.

”System-Wide Settings” includes things not specific to a profile - admin logins, system-wide settings, etc. You can choose to download the settings for all profiles, or for some combination of profiles.

Click `Download` to save a copy of the current settings to your workstation.

Restore Webinator Settings

Use this option to restore settings that you’ve previously captured using `Save Webinator settings`.

Test Network and Servers

This area provides the ability to test the network connectivity of Webinator and find what web and file server documents look like to it. It is divided into two sections. The first section is for testing Webinator fetching and processing of urls. The second section is for testing Webinator’s general network connectivity.

Test URL fetch

First choose the profile whose settings you want to use for the fetch test or choose `-Defaults-` for all default settings. If you are currently working in a profile that profile will be automatically pre-selected.

Checkboxes are provided for most of the profile URL settings so you can test those without typing them in. Check as many of those URLs as you want to test.

The input `Other URL` is also provided so you can test any arbitrary URL. Multiple URLs may be entered separated by space. Be sure to properly encode any entered URL. In particular encode space as `%20`.

Several processing options are provided to control how much processing to do.

- Full Processing
  Perform full processing on the fetched file as if it is being prepared for the search database. Otherwise only perform the basic download of the page.
- Check Robots.txt
  Consult the site’s robots.txt file to see if the selected URL is acceptable to fetch by the crawler. The test will fetch the URL regardless of robots.txt settings since it’s a single page test not a full crawl.
- Keep Download
  Keep the raw undecoded download and decoded data for display. Using this can make the test results page particularly large for large source documents like PDFs etc.

Press `Test Page Fetch(es)` to begin the testing of the selected URLs.

Each selected URL will be fetched in sequence and results of the fetch(es) presented one after the other on the same page. A short summary will be shown for each fetched URL followed by various statistics and

other information about the page. Most of the information is collapsed (hidden) to reduce page clutter. Click the `+` next to an item to expand that item for viewing. Click the `-` to recollapse an item. Use the `Collapse all` and `Expand all` links to Collapse all items or expand all items respectively. Use `Show empty fields` to show all fields even if there was no data for them. That helps one determine that a value is actually missing as opposed to overlooked for display.

Large text fields will be shown in scrollable areas by default to avoid taking over the page. Click the `+` next to a scrolling area to let it fully expand onto the page. Click the `-` to reconfine and expanded field.

Test Network

There are several network tests available. As many as desired may be done together. Each will be executed in sequence one after the other and the results presented together on one page.

- Find IP
  Lookup an IP address for a given host. Options (correspond to walk DNS Mode settings):
  - Internal - Perform the lookup using internal parallelizing routines.
  - System - Perform the lookup using standard system routines.
- Ping
  Send ping packets to the given hostname or IP address to determine reachability and speed. Check `Gateway` to ping the configured gateway address. A handful of packets will be sent and statistics about each and a summary of response times and loss will be displayed. *Note that not all machines respond to ping and some firewalls block ping. Page fetching may still work even if ping doesn't.*
- Traceroute
  Trace the network route to the given hostname or IP address to determine reachability and spot possible problem areas. It will display one line for each hop along the network route to the target machine. Asterisks (*) indicate a problem finding the next hop. *Note that some firewalls and routers block traceroute. Page fetching may still work even if traceroute doesn't.*
- Email
  Send a small test email to the given email address. This will test Webinator's email configuration as well as the recipient's ability to receive emails from Webinator. If the recipient doesn't get the test email look in `Maintenance->Manage Logs->maillog` to see if the message was handed off successfully. If it was handed off check the recipient's spam folder.

Advanced Support Tools

This area contains tools that Thunderstone Tech Support may ask you to run. You don't need to do anything in here unless tech support tells you to.

- **Re-output XSL files** In the past, when a profile was restored from backup or made as a copy, it was possible for a profile's XSL files on disk to become out of sync with the profile's settings. This has been fixed, but customers that were running outdated scripts may have profiles in this situation.

  The `Re-output XSL files` section checks which profiles have settings and files, and will re-write the profile's XSL data to disk.

- **Re-schedule walks** In the past, when a profile was restored from backup, it was not actually scheduled with the walk schedule (despite its settings saying it was). This has been fixed, but customers that restored profiles with old scripts may still have improperly scheduled profiles.

  The `Re-schedule walks` section re-applies all profile's scheduled settings to the walk scheduler.

- **Version 6 Upgrade CSS Fix** When upgrading to Texis version 6 from an earlier version, **Top HTML** and **XSL File** contents should have a CSS `<link>` tag to the default CSS stylesheet. This option provides an interface to automatically update them.

Chapter 5

Procedures and Examples

5.1 Searching your Index

Search the pages you have indexed by entering the following URL into your Web browser:

- On Unix:
  `http://www.mysite.com/cgi-bin/texis/webinator/search/`
- On Windows using CGI:
  `http://www.mysite.com/scripts/texis.exe/webinator/search/`
- On Windows using ISAPI:
  `http://www.mysite.com/texis/webinator/search/`

The above is a virtual path comprised of 2 parts. "`.../cgi-bin/texis`" is the Texis Web Script interpreter and "`/webinator/search`" is the path to the search script relative to your installation's `ScriptRoot`, which is the `texis/scripts` subdir of your install dir.

You may have to use a slightly different URL if you specified a different CGI directory during installation.

The URL given above will search the live database specified in the default profile called "`default`". If that profile is not found it will try to search the default walk database, `INSTALLDIR/texis/db` on Unix or `INSTALLDIR\texis\db` on Windows.

You may specify an alternate profile by including its name in the URL.

```
.../webinator/search/?pr=MYPROFILE
```

Where `MYPROFILE` is the name of the profile you wish to use. The search will use the live database specified by that profile.

You may also specify a database to search instead of a profile.

```
.../webinator/search/?db=DATABASE
```

Where DATABASE is the name of the database you wish to use. This would generally be the live database for a given profile which may be found as the first item listed on the administrative interface's Walk Settings page. Databases used this way must exist under the texis subdirectory of the installation directory. What you specify for DATABASE is only the portion of the path and name under the texis directory. For example, to search the database /usr/local/morph3/texis/myprofile/db2 you would use:

```
.../webinator/search/?db=myprofile/db2
```

When using a database instead of a profile, the look and feel settings will be those that were live when the walk of that database was performed. The profile will not be consulted for more recent changes. A benefit of not consulting the profile, however, is some increased search speed, which may be useful on a very heavily searched system. A disadvantage of specifying the database is that it will no longer be correct if a new walk is performed.

To get help on constructing queries click on the Advanced button of the search form. On the advanced search form you will find hyperlinks into the search help, which is also included in this manual in section 7.

To place the search form onto your existing web page(s) call up the Live Search from the administrative interface main menu (or the URL you determined from the above). This will bring up the search form. Use your web browser's view page source option (MSIE: TopMenu->View->Source, Netscape: TopMenu->View->Page Source) to get the source of the page. Cut everything between and including the <FORM> and </FORM> tags. That form may then be pasted into the web page(s) of your choice. You may also rearrange the look of the form as long as the variables are still present. If you have categories there will be a cq select list in the form. You may leave this out if you always want to search everything. Or you may make it a hidden variable with a fixed value if you always want to search the same section.

5.2 Similarity Searching

The search script has a feature called "Find Similar" which allows a user to click on a search result record to find more pages within the database similar to that one. This feature may also be accessed from any web page by placing the appropriate URL on it. You may search for pages in your database that are similar to any other web page whether it's in the database or not. The URL for finding similar pages has the form shown below.

Note: On Windows the /cgi-bin/texis/ *portion of the following URLs will be something like* /scripts/texis.exe/ *but may vary depending upon your installation.*

```
http://www.mysite.com/cgi-bin/texis/webinator/search/↪
   ↪similar.html?pr=default&ref=http://somesite/somepage.html
```

If the page containing the similarity URL resides on the same server as the search the http://www.mysite.com portion may be omitted:

```
/cgi-bin/texis/webinator/search/similar.html?↪
   ↪

pr=default&ref=http://somesite/somepage.html
```

If the profile to be searched is "default" the `pr=default&` portion may be omitted:

```
/cgi-bin/texis/webinator/search/similar.html?↪
   ↪ref=http://somesite/somepage.html
```

If the profile to be searched is anything other than "default" that must be specified instead of `default`:

```
/cgi-bin/texis/webinator/search/similar.html?↪
   ↪pr=myprofile&ref=http://somesite/somepage.html
```

If the page to be located is the page the URL is on the `ref=URL` portion may be omitted:

```
/cgi-bin/texis/webinator/search/similar.html
or
/cgi-bin/texis/webinator/search/similar.html?pr=myprofile
```

The similar function will lookup the desired URL in the database or, if it's not in the database, fetch it from the webserver. It will then search the database looking for indexed pages similar to the specified page.

You could place a URL like this on all of your pages so users could, with one click, find all pages on your site similar in content to the one they were reading.

5.3 Using the Thesaurus Feature

You can create a thesaurus to either replace or add to the default thesaurus. The creation procedure is the same for either usage. Note that a thesaurus is not limited to synonyms. It can contain anything you wish to associate with a particular word: i.e., identities, generalities, or specifics of the word entry, plus associated phrases, acronyms, or spelling variations. Webinator maintains a collection of thesauri that you upload. For each profile you may select which, if any, thesaurus to use.

Here are the steps to use the thesaurus feature.

- Create a thesaurus file. Use the syntax described in the document "User Equivalence File Format" at the following URL: `http://www.thunderstone.com/site/texisman/↪`
  `↪user_equivalence_file_format.html`

  That document refers to the thesaurus as an "equivalence file".

- Upload your thesaurus to Webinator. At the main menu click `Maintenance` then under System Settings click `Custom Thesaurus`. The Custom Thesaurus page opens.

- In the `Name` field, enter a symbolic name that will be listed as an option in search settings. This name does not have to be related to the filename on disk in any way.

- In the Permutations field, choose a value. This value controls how many variations of your defined terms to create during indexing of your uploaded source file. Here is an example of the effect of the various values.

  Assume a thesaurus entry of: `car,ford,chevy,toyota`

`Permutation None`: Just the terms as you entered them. Query “car” would find “car”, “ford”, “chevy”, and “toyota”. Query “ford” would only find “ford”.

`Permutations Single`: The terms you entered and the reverse. Same as above plus a query for any of “ford”, “chevy”, or “toyota” would find “car”.

`Permutations Full`: Equate every term with every other in each entry. Same as above plus a query for “ford” would find “chevy” and “toyota”.

- In the `New File` field, enter (or browse to) the file on your disk to upload. Click `Save Changes` to upload and index the file. When indexing is completed, you will receive a report about the indexing. If `Show results of indexing` is checked, you will also get a summary of the indexed words.

- After your thesaurus is installed on Webinator you can go to `Search Settings` for a profile to activate the thesaurus. There are three related options: `Synonyms`, `Main thesaurus`, and `Secondary Thesaurus`.

- Set `Synonyms` using the following information. `Synonyms` indicates how you want to apply a thesaurus (either yours or the default) to queries.

  `Disabled`: no phrase recognition and no synonyms (equivalences)

  `Phrase recognition only`: recognize query word groups that are known phrases and search for them as phrases

  `Phrases & Allow synonyms`: phrase recognition plus allowing the tilde ( ) operator to match synonyms on specific query terms

  `Phrases & Use synonyms by default`: phrase recognition and matching synonyms on all query terms (tilde to turn off on specific terms).

- Set the `Main Thesaurus` and `Secondary Thesaurus` fields by using the following information. If you want to use only your thesaurus and not the default one, select yours for the `Main Thesaurus` option and leave verb‘Secondary Thesaurus‘ set to none. If you want the default in addition to your own, leave `Main Thesaurus` set to Built-In and set `Secondary Thesaurus` to yours. The names listed in these options are the symbolic names (`Name` field) you gave your thesauri when uploading them.

- Click `Update` to apply these settings. There is no need to check `Apply Appearance`, and these settings are applied to both `Test Search` and `Live Search`.

5.4 Page Exclusion, Robots.txt, and Meta-robots

On the first access to a site the file `/robots.txt` will be retrieved, if its exists. Settings there will be respected. Any encountered URL that is disallowed by `robots.txt` will be discarded. Meta robots is also respected for each page retrieved. See `http://www.robotstxt.org/wc/exclusion.html` for the robots.txt and meta robots standards.

If there are any HTML trees that you don’t want indexed you may want to setup a `robots.txt` file, meta robots within the HTML pages, or use the various exclusion options to Webinator. For example: if you had a

"text only" version of your web server that duplicated the content of your normal server you would not want to index it. (On the other hand if most of your meaningful text is contained in graphics, Java, or JavaScript you may want to walk the text tree instead of the normal one, since graphics and Java are not searchable.)

Suppose your "text only" pages were all under a directory called `/text`. The simplest way to prevent traversal of that tree would be to use the exclusion or exclusion prefix.

The exclusion would look something like this:

```
/text/
```

The exclusion prefix would look something like this:

```
http://www.mysite.com/text/
```

That will prevent retrieval of any pages under the `/text` tree. This does not prevent other Web robots from retrieving the `/text` tree. To setup a permanent global exclusion list you need to create a file called `robots.txt` in your document root directory. The format of that file is as follows:

```
User-agent: *
Disallow: /text
```

Where `*` is the name of the robot to block. `*` means any robot not specifically named (all robots in this case since no others are named). Or you could specify the name of the robot. For Webinator it would be `Webinator`. You may specify several "Disallow"s for any given robot (see below). The "Disallow"s are simple path prefixes. They may not contain wildcards.

You may also specify different "Disallow" sets for different robots. Simply insert a blank line and add another "User-agent" line followed by its "Disallow" lines.

Here's a larger example:

```
User-agent: *
Disallow: /text
Disallow: /junk

User-agent: Webinator
Disallow: /text
Disallow: /webinator

User-agent: Scooter
Disallow: /text
Disallow: /junk
Disallow: /big
```

The `Scooter` robot will be blocked from accessing any pages under the `/text`, `/junk`, and `/big` trees. Webinator will be blocked from accessing any pages under `/text` and `/webinator`. All other robots will be blocked from accessing pages under `/text` and `/junk`.

Use of `robots.txt` is not enforced in any way. Robots may or may not use it. Webinator will, by default, always look for it and use it if present. This may be disabled by turning off "Respect robots.txt". When using robots.txt you may still use "Exclusions" for manual exclusion.

Meta robots provides another method of controlling robots such as Webinator. Any HTML may contain a meta tag in the source of the form.

```
<meta name="robots" content="WHAT-TO-DO">
```

`WHAT-TO-DO` may contain any of the following keywords. Multiple keywords may be used by placing a comma(`,`) between them.

Table 5.1: Meta-Robots Flags

Keyword	Meaning
`INDEX`	Index the text of this page
`NOINDEX`	Don't index the text of this page
`FOLLOW`	Follow hyperlinks on this page
`NOFOLLOW`	Don't follow hyperlinks on this page
`ALL`	Synonym for `INDEX,FOLLOW`
`NONE`	Synonym for `NOINDEX,NOFOLLOW`

Like `robots.txt` this is not enforced in any way. Robots may or may not use it. Webinator always indexes and follows hyperlinks by default so it only looks for `NOINDEX` and/or `NOFOLLOW` and/or `NONE`.

5.5 Indexing Other Sites

You may index a site other than your own by specifying its URL just as you would for your own site.

```
http://www.anothersite.com
```

Please be kind when indexing other sites. Many are low bandwidth or heavily used already and won't appreciate being hit hard. If you want to index any significant number of sites, please contact Thunderstone, as we may have what you want already. Remember that we are one SQL statement away from turning off any individual free Webinator license.

5.6 Indexing Individual Pages

To add an individual HTML page to the database, but not go after any of its references, add it to the `Single Page` list box.

5.7 Reindexing on a Schedule

It is often desirable to reindex a given site on a regular basis because of continuously changing content. You may specify a `Rewalk Schedule` to handle this for you.

It is also useful to perform a single rewalk at a later time or date to avoid overloading a web server during heavy use periods.

5.8 Checking for Web Server Errors

When you start a walk you will be sent to the walk status page. You may also reach that page at any time by selecting `Walk Status` from the menu. This page will show you the summary status of the running walk. When the walk completes you will see a summary of the walk as well as a list of any errors encountered. Following the error list is a list of duplicate pages encountered.

You may also view document linkage and info and errors from the `List/Edit URLs` page (4.3.5) from the menu.

5.9 Removing Pages from the Database

Use the `List/Edit URLs` menu (4.3.5) to find and delete specific URLs from the the database. You may delete individual pages or many pages at once using wildcards.

5.10 Erasing the Entire Database

If you decide to wipe out your existing database and it's settings to start over go to "Profiles" and click "Delete" next to the profile you wish to delete. This will completely remove the selected walk database and all options related to it.

5.11 Using Multiple Databases

Once you have a live searchable database you may want to build a separate one to contain different kinds of pages or to experiment with, without destroying your live database. Use the `Profiles` menu to create a new profile and database. You create the new profile with default settings or with a copy of the settings from another profile.

5.12 Integrating Webinator with your Site

There are three main techniques to integrate Webinator with your site. The techniques are categorized as follows:

- Static Host
- Dynamic Host and HTML
- Dynamic Host and XML

The simplest technique, `Static Host`, uses the built-in capability of Webinator to present a search page directly to a site visitor. Although this technique can be used with a dynamic host, it is commonly used with a static host. On your site, you present either a search field or a simple link. If you present a search field, when a visitor submits a query, the query is sent to Webinator. If you present a link, when a visitor clicks the link, a search page generated by Webinator is presented to the visitor, and the visitor uses this search page to submit a query. In either case, after a query is sent to Webinator, it responds by sending the search results (HTML) to the visitor's browser. Note that you can customize the HTML of the search page, and this allows you to maintain a consistent appearance for your site and the search page generated by Webinator.

The `Dynamic Host and HTML` technique can be used in dynamically generated web sites. The host server sends a search query (HTTP request) to Webinator, which responds by sending search results as HTML to the host server. The host server is responsible for sending the search query to Webinator, handling the HTML search results from it, and for all interactions with the site visitor.

The `Dynamic Host and XML` technique can be used in dynamically generated web sites. The host server sends a search query (HTTP request) to Webinator, which responds with the search results as XML. The host server is responsible for sending the search query to Webinator, handling the XML search results from it, and for all interactions with the site visitor.

5.12.1 Static Host

Use the information in this section to perform the `Static Host` type of integration.

- Decide whether your existing pages will include a query field or just a link to the Webinator search page.
- If your page will use a query field (an HTML form), you can obtain the HTML code that you need from the Webinator's live search page as follows:

  On Webinator, in the Administrative Interface, at the main menu, click `Live Search`. This opens the search form.

  Use your web browser's view page source option (MSIE: `TopMenu->View->Source`, Netscape or Mozilla: `TopMenu->View->Page Source`) to open a window that contains the source code of the page.

  Cut everything between and including the `<form>` and `</form>`. Paste the form into your web page(s).
- If your page will just link to the Webinator search page, create the link using the URL of the Test Search. To obtain the URL of the Test Search, at the Administrator Interface, on the main menu, click `Test Search`. When the Webinator search page opens, cut the URL string from your browser, and paste it into your web page(s) at the appropriate link element.

- Optionally, if you want to change the appearance of the search page, you can do this by adding HTML to the Top and Bottom HTML settings. At the Administrator Interface main menu, click the `Search Settings` link, and scroll until Top HTML and Bottom HTML are in view. Add the desired HTML code. For information about using an HTML editor to make these code additions, refer to (section 4.6.25).

- The settings are ready for test runs. Use the `Update Test` button to apply the ”test” settings, and use the `Test Search` link on the left to try them out.

- After you are satisfied with the appearance and operation of searches using `Test Search`, you are ready to go live.

- On the `Search Settings` page, press the `Update Live and Test` button to make the settings live.

- If you are using a link to the Webinator search page, change the link to point to the Live Search, using the same steps you used to set up the link to point to the Test Search.

5.12.2 Dynamic Host and HTML

Issuing a Query Programmatically

Use the information in this section to issue a query programmatically.

You can use either POST or GET to issue the search query. The only required variables are `pr` (profile), `query` and `dropXSL`. Variables not specified in the query take default values.

Here is an example URL for a search.

```
http://HOSTNAME/texis/search/main.xml?dropXSL=0&pr=default↪
↪&prox=page&rorder=500&rprox=500&rdfreq=500&rwfreq=500↪
↪&rlead=500&sufs=2&order=r&query=query
```

Where *HOSTNAME* is the IP/hostname of your Webinator, and *query* is the user’s query.

The following table provides a description of the query variables.

Please see the `Additional Fields` section (p. 120) for information on specifying additional field

Table 5.2: Search Query Variables

Variable	Description
`pr`	Specifies the Webinator profile.
`prox`	Proximity: words should be in the same line, sentence, paragraph, or page.
`rorder`	Word order: terms same order as query are better (`0-1000`; `500` = medium).
`rprox`	Indicates how close the words need to be (`0-1000`; `500` = medium).
`rdfreq`	Importance of frequency in the table (`0-1000`; `500` = medium).
`rwfreq`	Importance of frequency in the document (`0-1000`; `500` = medium).
`rlead`	Importance of closeness to document start (`0-1000`; `500` = medium).
`sufs`	Word forms (suffix). `0` (exact), `1` (plurals), or `2` (any), or `3` (custom)
`order`	Controls the sort order. Values are `r` (relevance) or `d` (date).
`query`	Search query entered by site visitor.
`rpp`	Results Per Page.
`cq`	For categories. `cq=1` for 1st, `cq=2` for 2nd etc.
`tq`	Used for title-only queries.
`uq`	Used for URL Prefix queries.
`dq`	Used for depth queries.
`mtq`	Used for Mime Type queries.
`mdlt`	Modified Date less than query
`mdgt`	Modified Date greater than query
`dateSource`	What date to use, id or Modified

searches in the URL.

dateSource: id vs modified

The `dateSource` parameter allows you to determine which date associated with the URL gets used for display, sorting, etc.

- `Modified` *(default)* - The time the file or page was last modified is used.
- `id` - the time that Webinator last update its record of the file or page is used.

If a collection of files that were modified a year ago were moved and just now picked up by the Webinator crawl last night, then the `Modified` date would be a year ago, but the `id` date would be last night.

`id` is the default dateSource when requesting an RSS feed of a search.

Processing Search Results

Webinator returns search results as HTML with this method, so the results can be passed along to the site visitor without changes, or they can be modified or expanded before they are sent.

5.12.3 Dynamic Host and XML

This section provides information about issuing a query programmatically and receiving the XML search results from Webinator.

Issuing a Query Programmatically

You can use either POST or GET to issue the search query. The only required variables are `pr` (profile), `query` and `dropXSL`. Variables not specified in the query take default values.

Here is an example URL for a search.

```
http://HOSTNAME/texis/search/main.xml?dropXSL=1&pr=default↪
   ↪&prox=page&rorder=500&rprox=500&rdfreq=500&rwfreq=500↪
   ↪&rlead=500&sufs=2&order=r&query=query
```

Where *HOSTNAME* is the IP/host of your Webinator, and *query* is the user's query.

Refer to `Issuing a Query Programmatically` 5.12.2 for definitions of the query variables.

Processing Search Results

Settings in the administrative interface `Search Settings` control the format of the data Webinator returns in response to a query. Set the `Results Style` to `XSL Style` so dropXSL does not use the XSLT, and Webinator will just return the raw XML.

The XML elements are described in the XML Elements in Search Results section, p. 155.

Sample ASP Code

The following ASP code demonstrates sending an http GET command to Webinator and receiving XML search results from it.

```
<%
 on error resume next
 dim objSrvHTTP
 dim objXMLSend
 dim objXMLReceive

 set objSrvHTTP = Server.CreateObject("MSXML2.ServerXMLHTTP.4.0")
 set objXMLSend = Server.CreateObject("MSXML2.DOMDocument.4.0")
 set objXMLReceive = Server.CreateObject("MSXML2.DOMDocument.4.0")

  if err.number <> 0 then
      Response.Write err.description
      Response.Write "First error in code."
    end if
    err.clear

 objXMLSend.async = false
 objXMLSend.loadXML ("<msg><id>2</id></msg>")
 objSrvHTTP.open
"GET","http://HOSTNAME/texis/search/main.xml?dropXSL=1&
pr=test&prox=page&rorder=500&rprox=500&rdfreq=500&rwfreq=500&
rlead=500&sufs=0&query=test&submit=Submit",false

  if err.number <> 0 then
      Response.Write err.description
      Response.Write "Second error in code."
    end if
    err.clear

 objSrvHTTP.send objXMLSend
 set objXMLReceive = objSrvHTTP.responseXML
 Response.ContentType = "text/xml"
 Response.Write objXMLReceive.xml

  if err.number <> 0 then
      Response.Write err.description
      Response.Write "Third error in code."
    end if
    err.clear
%>
```

5.13 Search Result RSS Feeds

Search result RSS feeds can help you monitor a certain search query, and let you know when new results appear for the query.

All search result pages have an RSS link embedded in them. Recent versions of modern browsers, such as Internet Explorer and Firefox, have built-in features that notify you when an RSS feed you're subscribed to changes.

- IE 7 and 8 - `http://www.microsoft.com/windows/IE/ie7/tour/rss/`
- Firefox - `http://kb.mozillazine.org/Live_Bookmarks_-_Firefox`
- Opera - `http://www.opera.com/mail/rss/`

5.14 OpenSearch Support

The search interface also has an embedded Open Search description. This means that modern browsers can use the Quick Search box (to the right of the address bar) to perform searches on Webinator.

- Bring up the search interface for the profile of your choice
- Hit the "down" arrow next to the Quick Search box
- Choose "Add Search Provider..." to add Webinator to the list of available searches.

Internet Explorer users can find more detailed instructions at
`http://msdn.microsoft.com/en-us/library/cc848862.aspx`

5.15 Using Best Bets

Webinator allows you to create links that will appear either at the top or to the right of the search results when specific keywords are searched for. They can be used for suggested links, or to promote specific URLs so they stand out from the main results. The Best Bet links are arranged into groups, which allow you to enable or disable a group of results easily.

5.15.1 Quick Creation

The easiest way to create Best Bets is to directly add keywords to URLs. This skips the group and display settings, which can be customized later (and are detailed below).

From the "List/Edit URLs" page, enter the URL you want and click on the URL to get the details on that URL. There is a form on the page that allows you to add keywords to that URL. You can define a priority, title, description, and keywords for the URL (as listed in the table below).

The group will be listed as `(Create New)`. This will create a default group and automatically set it to display, instantly using the Best Bet you just created. The created group (`default`) can then be used to create any number of other keyword-URL associations.

You can go to the "Search Settings" page to customize how the Best Bets are displayed, as detailed below.

5.15.2 Fully Customized

The first step in create best bet links is to define a group. This is done from the "Group Settings" tab. You can name the group, and decide which information will be displayed about the group.

After creating a group you can add keywords to specific URLs. From the "List/Edit URLs" page enter the URL you want, and click on the URL to get the details on that URL. Currently you can only use URLs that have been walked and are in the database. The fields on the form are:

- `Priority` *(optional)* - The priority for this Best Bet. If multiple Best Bets match a single query, their `Priority` fields determine what order they're shown. If you're only setting one Best Bet per URL, or aren't particular about the order, you don't need to set a priority.

- `Title`- The title that will be displayed for the Best Bet on the search results page.

- `Keywords`- A comma separated list of what keywords will trigger this Best Bet. A Best Bet is displayed when the query is completely contained within one of the keywords for that Best Bet.

- `Group`- Which Best Bet Group this Best Bet will be created in. A Best Bet will only have a chance to match if its group is set to display as either `Top Best Bets` or `Right Best Bets`on the Search Settings page.

  If no groups currently exist, `(create new)` will be displayed, and a group will be created for you if you enter keywords and a title for this Best Bet.

- `Description` *(optional)* - The description to display for this Best Bet. The Best Bet Group for this Best Bet might be set to not display the description, so it's optional.

The title and description can contain HTML code. Be careful that it does not disrupt the rest of the page layout. You can create multiple entries for the same URL. Each time you save a new set of blank boxes will be shown.

Once the Best Bets are created you can go to the "Search Settings" page to set up how they are displayed. For the top and right placements you can define which group is shown there, what title if any to display above the links, and the color, size and style of the boxes around the Best Bets.

As with any of the Search Settings these will apply to the "Test Search" first, and then when you apply the settings be copied to the "Live Search", allowing you to test the settings and make sure they are appropriate before going live.

5.16 Using Access Control

The concepts and actions of access control in Webinator are discussed in detail in the Access Control section, p. 85. The following are some general tips on how to setup and maintain access control rights.

Access Control is supported in the full Texis product, but not Webinator-only.

5.16.1 Initial Lockdown

Since the default mode for Access Control when created is to allow all rights to all users for back-compatibility, it is recommended that perms be "locked down" first, and only granted as needed. The `webinator` user, having the irrevocable ability to reset ACLs, should remain a "superuser" with all access, and other accounts turned into lesser-permission users. Lockdown should happen in this order:

1. Allow superuser: The `webinator` user should have an `Allow` entry for all rights to the top-level `Global` object[1].
2. Deny everyone: The group `Everyone` should have a `Deny` entry for all rights to the top-level `Global` object.

With these perms, users other than `webinator` – including new users and profiles created in the future – will not be able to see or modify administrative settings. They can be granted perms as needed later, for example, the Read right could be removed from the `Global` deny ACE so that they can read but not modify any admin action/setting.

5.16.2 Example: User with Complete Control on One Profile

To configure a user that has complete access to just one specific profile (but no other profiles, nor the rest of administration such as creating accounts etc.), set up the lockdown settings above, then:

1. Create a Profile ACE on the specific profile, for that user, read and write access, and type `Allow`.
2. Create a Setting ACE for `All Settings`, for that user, read, write and delete access, type `Allow`.

The user will now be able to modify any setting on that profile, as well as start/stop walks on it, but will not be able to edit other profiles.

5.16.3 Example: User with Look and Feel Control on All Profiles

To configure a user that has the ability to change the Top and Bottom HTML on *any* profile, but cannot edit walk settings, nor start nor stop a walk, etc., set up the lockdown settings above, then:

[1] In version 5.3.0 and earlier, the webinator user should instead be explicitly granted all rights to each of the second-level objects (All Users, All Groups, All Profiles, All Settings, and Maintenance).

1. Create a Profile ACE on `All Profiles`, for that user, read and write access, and type `Allow`.
2. Create a Setting ACE for `Top HTML`, for that user, read, write and delete access, type `Allow`.
3. Create a Setting ACE for `Bottom HTML`, for that user, read, write and delete access, type `Allow`.

The user will now be able to change the top and bottom HTML for any profile.

5.17 Replication

5.17.1 Replication Overview

In replication, a server profile sends walk data to another server profile. The two profiles can be on different machines or they can be on the same one. If the profiles are on different machines, the sending and receiving profiles can have the same or different names. If the profiles are on the same machine, use different profile names.

Replication is supported in the full Texis product, but not Webinator-only.

Here is an example that illustrates the replication process. In this example, the `Sender` profile has been set up as the sender profile and `Receiver` is the receiver profile. After `Sender` performs a walk, it sends the walk data to `Receiver`. The `Receiver` profile accepts the data as-is, without regard to its own profile settings. Only the profile that performed the walk may send the walk data, so in this example `Receiver` cannot replicate (the data it received from `Sender`) to another profile.

To avoid undesired overwriting of replication walk data, you should not allow the receiver profile to perform walks.

Before the receiver will accept replication data, the sender(s) need to be granted permission to send the data. This permission is managed in a cluster member list.

A good use of replication is to set up multiple machines to replicate to a single receiving profile. For example, machines A, B, and C each have a different profile, and they each replicate their walk data to a profile on machine D, which is the receiver. Another use of replication is to send walk data from multiple profiles on a machine to a single receiver profile that is on the same machine. This provides a means of combining walk data into a single profile. Another use of replication is to replicate data from one sender to multiple receivers. This way multiple machines hold the same walk data.

5.17.2 Procedure

The procedure in this section is an example of setting up replication on a single machine. It can be adapted to multiple machine configurations by changing the Replication Settings.

Set up the Sender Profile

- Choose an existing, walkable profile to be the sender. Or go to the `Profiles` menu item and create one, filling in all fields for a normal walk. We'll assume this profile is called `Sender`.

- Go to the `All Walk Settings` menu item for the `Sender` profile.
- Scroll down to `Replication Settings`.
- Enter the information for the receiver. In this example, `Host IP or Name` is `localhost` because we'll be sending data to the same machine, and `Profile Name` is `Receiver`. The page now includes the location of the receiver profile.
- Click `Update and Go` button.
- After a moment, the `Walk Status` page opens. Notice that there are N items in the replication queue. The number N is similar to the number of pages that were walked. The items remain in the queue, because they cannot be sent until the receiver profile is created (below). Normally, when a receiver profile is present, the contents of the queue are automatically sent to the receiver.

Create the Receiver Profile

- Create a new profile called `Receiver` via the `Profiles` menu item. (This matches the receiving profile name we entered on the `Sender` profile.)
- At main menu click `Maintenance`, then under `Search Appliance Settings` heading, click `System Wide Settings`.
- At the `Cluster Members` field, enter the IP address for each server that will send walk data to this machine. Use a separate line for each entry. In this example, there is one sending IP address, and it is 127.0.0.1 (use IP numbers, not the word `localhost`). To enable an entire subnet to send data, use an IP prefix and wildcard, eg. `10.10.*`.
- Click `Update` button.
- At main menu, click `Profiles`.
- When `Profiles` page opens, click `Sender`. A `Walk Settings` page opens for the `Sender` profile.
- Click `Walk Status` button. The `Walk Status` page for the `Sender` profile opens.
- There are still N items in the replication queue.
- Click the `replication queue` link.
- The items in the replication queue are sent to the `Receiver` profile. On the `Walk Status` page, there are now 0 items in the replication queue, which indicates the items were sent.
- On main menu, click `Profiles`, click `Receiver`, click `Walk Status` and observe that there is a list of pages recently walked. These pages were not walked by `Receiver`, instead they were obtained from `Sender`, which performed the walk.

5.17.3 DataLoad API

The replication system can also be used to load data directly onto Webinator from an outside source, instead of "pulling" it from a URL or its links. This can be used for data that is not permanently stored at its URL (eg. generated data), and therefore cannot be fetched for indexing; it can instead be pushed to Webinator for indexing.

The DataLoad API is supported in the full Texis product, but not Webinator-only.

Before loading data onto Webinator, it must be configured to accept data from the IP address(es) that will be sending to it. This procedure is the same as for replication; see the Cluster Members setting, p. 113.

Submission Format

Data is submitted to Webinator with an HTTP POST request sent to a similar URL as the admin interface (eg. `http://.../dowalk`), but with `/recvdata.xml` appended. Eg.:

```
http://www.mysite.com/texis/webinator/dowalk/recvdata.xml
```

The following POST variables must be set in the request. Be sure to URL-encode the values:

- `profile`
  Set to the name of the receiving profile.

- `data`
  Set to an XML document containing the data, and what to do with it (insert/delete/etc.). See below for details.

Specifying all fields manually

Below is an example `data` document where all fields are specified. Be sure to HTML-encode values.

```
<?xml version="1.0" encoding="UTF-8"?>
<ThunderstoneReplication
      xmlns:dt="urn:schemas-microsoft-com:datatypes"
>
  <Item>
    <Type>I</Type>
    <Size>150369</Size>
    <Visited>2005-10-25 15:25:18</Visited>
    <Dlsecs>0</Dlsecs>
    <Depth>0</Depth>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Title>Sprocket Specifications</Title>
    <Body>...</Body>
    <Keywords>sprockets, gears, hubs</Keywords>
    <Description>Sprocket details</Description>
    <Meta></Meta>
    <Category>Mechanical</Category>
    <Modified>2005-10-25 11:21:07</Modified>
    <NextCheck>2005-10-25 16:25:18</NextCheck>
    <Views>0</Views>
    <Clicks>0</Clicks>
    <CTR>0.000000</CTR>
    <Pop>0</Pop>
    <MimeType>text/html</MimeType>
    <Charset>UTF-8</Charset>
    <Refs dt:dt="bin.base64">...</Refs>
    <Errors dt:dt="bin.base64">...</Errors>
    <RawData dt:dt="bin.base64"></RawData>
  </Item>
</ThunderstoneReplication>
```

Any element whose text data might not be XML-safe (eg. binary chars in the `<Body>`) should be base64-encoded, and the attribute `dt:dt="bin.base64"` set in the tag. Eg. the `<Refs>` and `<Errors>` elements' text data are always base64-encoded. Note that the XML namespace prefix `dt` should also then be set to `urn:schemas-microsoft-com:datatypes` in the root `<ThunderstoneReplication>` element.

The elements are:

- `<Type>` The action to take with this data. Text value may be one of:
  - `I` Insert the data (overwrite previous data for URL if any)
  - `D` Delete the URL
  - `DP` Delete the URL as a pattern (eg. `http://www.mysite.com/dir/*`)

  - `UI` Update search indexes (call after a batch of inserts/deletes)

- `<Size>` The integer size of the original document.
- `<Visited>` When the document was fetched, in `YYYY-MM-DD HH:MM:SS` format.
- `<Dlsecs>` Number of seconds taken to download the document.
- `<Depth>` Depth of URL from a Base URL, eg. 0 is a Base URL, 1 is one click away, etc.
- `<Url>` The URL of the document.
- `<Title>` The title of the document.
- `<Body>` The formatted body of the document.
- `<Keywords>` Any keywords for the document.
- `<Description>` The description of the document.
- `<Meta>` Any meta data for the document.
- `<Category>` The category the document is in, if any. Must be a category name from the profile's Categories.
- `<Modified>` The `Last-Modified` date of the document, in `YYYY-MM-DD HH:MM:SS` format.
- `<NextCheck>` When the document should be refreshed, in `YYYY-MM-DD HH:MM:SS` format.
- `<Views>` Number of views of the document: how many times it's been shown in search results.
- `<Clicks>` Number of clicks of the document: how many times it's been clicked on in search results.
- `<CTR>` Click-through-ratio: floating-point number ratio of clicks to views.
- `<Pop>` Document popularity: number of references (links) to it.
- `<MimeType>` The MIME type of the content served at the URL, or provided in RawData.
- `<Charset>` Character set of `<Body>` data. Should correspond with `Storage Charset` profile setting (p. 44). If a charset other than the `Storage Charset` is used, it should be a standard IANA charset that Webinator can convert to the `Storage Charset`.
- `<Refs>` Optional element with references (child links) of the document.
- `<Errors>` Optional element with errors of the document.

Uploading a binary file

If you have a binary file, such as a PDF or an Office document, you can send it with the dataload API and let the Webinator extract the text from it.

```
<?xml version="1.0" encoding="UTF-8"?>
<ThunderstoneReplication
      xmlns:dt="urn:schemas-microsoft-com:datatypes"
>
    <Item>
        <Type>I</Type>
        <Url>http://www.example.com/dataload.pdf</Url>
        <RawData dt:dt="bin.base64">0M8R4KGxGu....</RawData>
    </Item>
</ThunderstoneReplication>
```

The elements are:

- `<Type>` The action to take with this data. Text value may be one of:
  - `I` Insert the data (overwrite previous data for URL if any)
- `<Url>` The URL of the document.
- `<RawData>` element with the base64 encoding of raw document. It must include the `dt:dt="bin.base64"` attribute.

Combining the two: binary files with custom fields

It is possible to specify both a `<RawData>` document, *and* fields such as `<Title>`, `<Description>`, etc. The binary document will be processed, and any other fields provided will override the values that came from the document.

This can be useful in situations where you have a Content Management System (CMS) that contains metadata about a document that doesn't actually *occur* anywhere in the document. You can do a custom dataload that pushes in the document, and the custom Title/Description/etc.

Additional Fields

Each profile-specific Additional Field is optionally sent in a single element named after the field, with the XML namespace prefix `u`. The value of the field is the content of the XML element. Note that the `u` XML namespace prefix should be declared in the root `<ThunderstoneReplication>` node, as shown earlier.

For example, an `Integer` field `Quantity` and a `Text` field `State` may be given as:

```
<u:Quantity>57</u:Quantity>
<u:State>NY</u:State>
```

Other Details

The optional `<Refs>` element lists the links (references) from the given document, for parent-child linking. Its text value is a base64-encoded XML document with the following format when decoded:

```
<results xmlns:dt="urn:schemas-microsoft-com:datatypes">
  <result>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Ref>http://www.mysite.com/dir/otherpage.html/</Ref>
  </result>
  ...
</results>
```

Each `<Url>` should be the same as the `<Url>` in the above `<Item>` block. The `<Ref>` is a single link from the page. Only one `<Ref>` may be listed per `<result>`; additional links should be sent with additional `<result>` elements.

The optional `<Errors>` element contains any errors to be logged for the document. Note that this may be empty or not present if no errors are to be logged. Its text value is a base64-encoded XML document with the following format when decoded:

```
<results xmlns:dt="urn:schemas-microsoft-com:datatypes">
  <result>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Reason>Document not found: 404 (Not Found)</Reason>
  </result>
  ...
</results>
```

As with the `<Refs>` element, the `<Url>` must correspond with the original `<Item>` `<Url>`, and multiple errors must be listed in separate `<result>` elements.

Reply Format

The response to a DataLoad request is an XML document:

```
<ThunderstoneReplicationResult>
  <ItemResult>
    <rid>000000000</rid>
    <Type>D</Type>
    <DP>1</DP>
    <Status>OK</Status>
    <Info>Not found</Info>
  </ItemResult>
  <Rows>1</Rows>
  <Version>Version 5.01.1234567890 20051010 (...)
           2005-10-10 12:34:56</Version>
</ThunderstoneReplicationResult>
```

The elements are:

- `<rid>` The replication id. Ignored.
- `<Type>` The action type specified in the request.
- `<DP>` The number of URLs deleted by a `<Type>DP</Type>` action. Element is not present for other `<Type>`.
- `<Status>` Result code:
  - `OK` Success
  - `FAIL_UNKNOWNTYPE` The `<Type>` was not recognized
  - `NODATA` No parsable data in request
  - `Not Allowed` Sender is not a `Cluster Member`
  - `No Profile` No `profile` set in request POST
  - `FAIL` Failed, unknown reason
- `<Info>` Optional additional message; eg. `Not found` if a non-existent URL is deleted
- `<Rows>` How many request `<Item>`s were processed.
- `<Version>` Version and release date of the software.

Once data has been successfully loaded onto Webinator, if the profile has any receiver profiles defined under `Replication Settings`, the data will also be queued for replication to those receivers.

Dataload SOAP API

There is a SOAP API available for dataload, allowing you to use a SOAP library to communicate with the Webinator. For an overview of SOAP, Please see the `SOAP API` section (p. 121).

The WSDLs for the dataload API can be found on the `Profile Tools` page. Providing these WSDLs to whatever tool your language uses, such as Visual Studio's `wsdl.exe` program, should generate the necessary wrapper class.

The parameters are defined within the WSDL itself, and are generally the same as mentioned above in the `Submission Format` and `Reply Format` sections, with a few exceptions:

- The entire transactions are wrapped by SOAP envelops and the top-level elements are called `dataload` and `dataloadResponse` instead of `ThunderstoneReplicationResult`, respectively.
- The `dataload` element contains a `profile` element in addition to all the `Item`s.

C# Example Project

A C# example project is available that demonstrates using both the search and dataload SOAP interfaces. In the `Maintenance` section of the administration interface, choose `Extra Downloads`, and then `Thunderstone Soap Example`. Instructions are listed on that page and within the zip itself for how to use the project.

5.18 Additional Fields

5.18.1 Overview

The additional fields feature in the Webinator allows you to define structured data that can be searched on, sorted by, and included in the results when using an XSL stylesheet. Typical uses might include having prices, dates or ratings associated with the documents.

Additional Fields are supported in the full Texis product, but not Webinator-only.

5.18.2 Populating

To populate the additional fields they should first be defined in the additional fields section of the settings. You can specify a name, which is used as the name of the XML element when displaying the results, as well as when using the DataLoad API.

Once the field has been defined it can be populated either via the DataLoad API or through the Data From Field section. The fields are positionally numbered, and you can load Extra Field 1, 2 and/or 3 from the page that is read. If you are loading from a META field you will typically want a search of `.+` and the Meta Field you are loading from.

5.18.3 Sorting

To sort the results you can use the `order` form variable. To specify the first field you can set the value to `af1`, for the second `af2` and for the third `af3`. If you want to reverse the sort order you add a `d` to the value, i.e. `af1d`, `af2d`, `af3d`.

5.18.4 Searching

To add a search restriction to the query you can specify form variables with a name constructed as `af#OP`, where # is the number of the additional field, 1, 2, or 3, and OP is one of the following:

- `eq` - the field is equal to the form variable (e.g. af1eq)
- `gt` - the field is greater than the form variable (e.g. af2gt)
- `gte` - the field is greater or equal to than the form variable
- `lt` - the field is less than the form variable
- `lte` - the field is less or equal to than the form variable

5.19 SOAP API

5.19.1 SOAP Overview

The Simple Object Access Protocol (SOAP) is a W3C (World Wide Web Consortium) recommendation that essentially allows for Remote Procedure Call (RPC) functionality over HTTP, via XML. (This is a simplification of the 120-page SOAP spec, but it suits our purposes). SOAP web services provides a systematic, defined way of communicating function requests and responses over a network transport.

SOAP interfaces are described by another W3C recommendation, WSDL documents - Web Services Definition Language. WSDL documents are the prototypes for SOAP functions. They define what parameters are expected to the functions, what formats are/aren't allowed, what will be returned, etc. Given the WSDL of a SOAP web service, it's possible to generate the client code that interacts with the services (as is demonstrated in the C# example project later).

Specifically for the Webinator, the SOAP interface provides, when using a language that has a SOAP API, a way to invoke a search and on the Webinator and insert data as if it were a local function call.

5.19.2 SOAP API vs. XML Output

The SOAP interface provides functionality very similar to the "XML output" search interface. So why use one as opposed to the other?

Use SOAP if the language you are writing has a SOAP interface available for you. Many languages and environments (including Visual Studio) provide SOAP tools, where you provide the WSDL to the

webservice, and it will generate "wrapper" classes for you, allowing you to interact with Webinator as if it were simply a local function.

If whatever development environment you're using doesn't have a real SOAP interface, then use the XML API instead of the SOAP API. All the added information/rules of SOAP that make it easy for programs to exchange data will instead make it more cumbersome to use manually.

5.19.3 Getting the WSDL

The WSDLs can be found on the `Profile Tools` page for the profile. It links to the Dataload WSDL selector and a search WSDL selector, which lets you choose either a WSDL for this profile, or for all profiles (as explained below).

5.19.4 Global vs. per-profile WSDLs

When viewing search WSDLs, you have the option of requesting a WSDL specific to a single profile, or a global `All Profiles` WSDL, which can be used for any profile.

If you don't make use of Additional Fields, then there will be no difference between per-profile and global WSDLs.

Both per-profile and global WSDLs refer to the same search interface. The same SOAP response is generated for both WSDLs. The only difference is in how specific the WSDLs are - per-profile WSDLs specify which Additional Fields occur in the results, but the global WSDL must use `<xsd:any>` as a catch-all, as the Additional Fields may change from one profile to another.

Which you use is a trade-off that you must decide on.

- **per-profile WSDLs**
  - **Advantage** Additional Fields for the profile are "hard-coded" in the WSDL itself, so a SOAP client consuming the WSDL can make better use of the Additional Fields.
    For example, if your profile has Additional Fields called `price` and `location`, then a per-profile WSDL will specify that each result contains `<price>` and `<location>` elements. WSDL tools can do things like declare `response.price` and `response.location` variables.
  - **Disadvantage** Because the per-profile WSDL is specific to that profile's Additional Fields, a different WSDL must be used for every profile you want to interact with. If you're interacting with many different profiles (or it often changes), an global WSDL may be better suited.
- **Global WSDLs**
  - **Advantage** The `All Profiles` wsdl can be used for any profile. This is better if your application needs to query multiple profiles, or if you don't work with Additional Fields.
  - **Disadvantage** Additional Fields are represented in the `All Profiles` WSDL with `<xsd:any>`, which allows it to not declare which Additional Fields will occur in the XML (as it may change from one profile to another).

This means that programs consuming the WSDL cannot know which parametric fields will be returned, and will instead do things like offer an array of Additional Field XML elements that you must manually loop over to find the ones you want.

5.19.5 Configuring the SOAP Interface

The WSDL for Webinator is accessible in the following URL path from Webinator:

`/texis/ThunderstoneSearchService/describe.wsdl`

This link is also available from the `SOAP Tools` section, within `Profile Tools` in the admin interface.

Dataload SOAP API

The Dataload SOAP API takes the same parameters as the normal dataload API, please see the `Submission Format` (p. 114) and `Reply Format` (p. 119) sections, with a few exceptions:

- The entire transactions are wrapped by SOAP envelopes and the top-level elements are called `dataload` and `dataloadResponse` instead of `ThunderstoneReplicationResult`, respectively.
- The `dataload` element contains a `profile` element in addition to all the `Item`s.

5.19.6 C# example project

A C# example project is available that demonstrates using both the search and dataload SOAP interfaces. In the `Maintenance` section of the administration interface, choose `Extra Downloads`, and then `Thunderstone Soap Example`. Instructions are listed on that page and within the zip itself for how to use the project.

5.19.7 SOAP Links for Languages

This section contains links for recommendations of SOAP implementations in other languages. Thunderstone makes no guarantees to the completeness or quality of these projects, we simply provide links for convenience.

- ASP.NET - the same C# API code can be compiled into a .NET assembly and used from an ASP.NET page.
  - `http://www.codeproject.com/webservices/aspwebsvr.asp`
- Perl - SOAP::Lite for Perl is a collection of Perl modules which provides a simple and lightweight SOAP interface.
  - `http://www.soaplite.com/`

- Python The Web Services for Python Project provides libraries implementing the various protocols used when writing web services including SOAP, WSDL, and other related protocols.
  - `http://pywebsvcs.sourceforge.net/`
- Java - The Java API for XML Messaging (JAXM) provides a framework for sending and receiving SOAP messages.
  - `http://java.sun.com/webservices/jaxm/`
- c++ - EasySoap++ is a C++ library for SOAP.
  - `http://easysoap.sourceforge.net/`

5.19.8 SOAP API search Reference

Webinator exposes three functions for searching: `search`, which performs a normal search, `moreLikeThis`, which finds similar pages, and `showParents`, which shows which pages link to a page.

There are some parameters that are standard to all the functions:

- `jump` - number of user-visible results to skip. 0 would return the first page of results, 10 the second page, etc. (assuming 10 results per page).
- `order` - specifies how the results should be ordered. Possible values are: `r` - sort by relevance (default) `dd` - newest pages first `da` - oldest pages first

RankKnobs structure

There's an optional "rankKnobs" parameter for many of the functions that can specify how things should be ranked (each function notes whether it accepts rankKnobs). All of these can be set from 0-1000, where the higher the value, the more heavily that aspect is weighed; 500 is the default. These parameters correspond directly to the "Ranking Factors" settings on the Advanced Search page.

RankKnobs has the following parameters:

- `order` - importance of the words being in the proper order
- `proximity` - importance that the words are close together
- `dbFreq` - importance of the frequency of a word in the database
- `docFreq` - importance of the frequency of a word within the document
- `leadBias` - importance of closeness to the start of the document

search

This performs a normal search based on the query/queries provided.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `query` - Required. The Metamorph query to search for.

The following parameters can be provided to further refine the search:

- `urlQuery` - Used for URL Prefix queries. Corresponds to the default search interface's `uq` query-string variable (p. 105).
- `depthQuery` - the maximum depth that would be allowed. Supplying a value of `3` would only return pages that are no more than 3 clicks away from a Base URL. Corresponds to the default interface's `dq` variable.
- `categoryQuery` - numeric index for a category to use. 1 is the first category, etc. Corresponds to the default interface's `cq` variable.
- `proximity` - specifies a required proximity for the words in the query. Corresponds to the default interface's `prox` variable. Possible values are:
  - `page` - words must occur on the same page (default)
  - `paragraph` - words must occur in the same paragraph
  - `sentence` - words must occur in the same sentence.
  - `line` - words must occur on the same line.

The `search` function may also use the rankKnobs structure (section 5.19.8, p. 124).

The `search` function may also take a number of Additional Field parameters, as described in the `Searching Additional Fields` section (p. 39).

The output of the function is described in the XML Elements in Search Results section (6.8, p. 155).

moreLikeThis

`moreLikeThis` returns results that are similar to a result already found.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `id` - Required. the id of a single URL, returned from a previous function.

The `moreLikeThis` function may also use the `<rankKnobs>` structure, as described in section 5.19.8, p. 124.

The output of the function is described in the XML Elements in the Search Results section (6.8, p. 155).

showParents function

`showParents` lists all the pages that link to a previous retrieved search results.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `id` - Required. the id of a single URL, returned from a previous function.

The output of the function is described in the XML Elements in the Search Results section (6.8, p. 155).

5.19.9 SOAP API admin Reference

login

Parameters:

- `username` - the user being logged in
- `password` - the password for the user

Returns:

- `authToken` - an authentication token for use in further requests

`login` logs you in to the appliance, supplying you with an authentication token that will be included in all further requests to show that you're logged in. All other Admin SOAP API calls require an authToken for use. It's sent in further requests via a SOAP Header.

listProfiles

Parameters:

- *none*

Returns:

- `ProfileName` - an array of profile names requests

Returns a list of all profiles that currently exist on the appliance. If no profiles exist, a successful response with no `ProfileNames` is returned.

getProfileStatus

Parameters:

- `Profile` - name of the profile

Returns:

- `IsRunning` - whether or not the crawl is currently running, set to `true` or `false`.

Returns information about the profile, currently just whether or not the profile is running.

addProfile

Parameters:

- `Profile` - name of the new profile
- `Type` *(optional)* - the type of profile, `standard` or `metasearch`. Defaults to `standard`.
- `CopyOf` *(optional)* - name of the profile to copy
- `ParametricField` - *unused in Webinator*
- `PrimaryKey` - *unused in Webinator*
- `Dataspace` - the directory to use for the profile's data

Returns:

- `Success` - will be set to `ok`, indicating the profile was created successfully.

Adds a new profile to Webinator. If there's any problem (already exists, invalid profile name, etc), a SOAP Fault will be thrown.

deleteProfile

Parameters:

- `Profile` - name of the profile to be deleted

Returns:

- `Success` - will be set to `ok`, indicating the profile was deleted successfully.

Deletes a profile from Webinator. If the profile didn't exist, the call will still succeed.

getSettings

Parameters:

- `Profile` - name of the profile
- `Name` *(optional)* - array of setting names to get. If no `Name` is provided, all settings are returned.
- `TestOrLive` *(optional)* - whether to return the test settings, or live settings. Returns live settings by default.

Returns:

- `Setting` - an array of name/value pairs for the requested settings. A `TestOrLive` attribute on the settings indicate whether this applies to test, live, or both.

Gets a list of settings for the requested profile. You can request one or more specific settings by passing in `Name` parameters, or get all settings by not supplying a `Name`.

Some settings have test and live versions. You can request which version you'd like (defaults to live), and the returned settings indicate whether they apply to test or live. "Both" indicates that setting doesn't have different test and live versions.

setSettings

Parameters:

- `Profile` - name of the profile
- `TestOrLive` *(optional)* - whether this should apply to test settings, live settings, or both. Defaults to both.
- `Setting` - multiple name/value pairs of settings that you'd like to set

Returns:

- `Success` - set to `ok`, indicating the settings were set properly.

Applies an array of settings for the given profile.

If there is any problem (such as an invalid setting name) in any one of the settings, a SOAP Fault is returned, and NONE of the settings are applied. This allows you to tweak the problem settings, and re-submit the entire batch again, without having them "partially applied" in between.

getThesauruses

Parameters:

- *none*

Returns:

- `Thesaurus` - an array of Thesaurus information
  - `Name` - name of the thesaurus
  - `Permutations` - what permutations apply to this thesaurus
  - `NumProfileUsing` - the number of profiles that are currently using this thesaurus

Returns information about all thesauruses that exist in Webinator.

setThesaurus

Parameters:

- `Name` - name of the thesaurus. If this thesaurus already exists, it will be replaced.
- `Permutations` - what permutations apply to this thesaurus. Possible values are `Full`, `Single`, or `None`. Defaults to `Single`.
- `Verbose` *(optional)* - If set to `Y`, verbose output of processing the thesaurus content will be included in the response.
- `Content` - the text content that should be used for the thesaurus. See the Thesaurus section for details on the format (5.3, p. 99).

Returns:

- `Output` - output of the thesaurus processing operation. Any errors will be listed in the text.

Creates or updates a thesaurus in Webinator. Once created, a thesaurus can be used in a profile by setting its `SSc_eqprefix` or `SSc_ueqprefix` to this thesaurus' `Name`.

deleteThesaurus

Parameters:

- `Name` - name of the thesaurus to delete

Returns:

- `Output` - output of the thesaurus deletion operation. Any errors will be listed in the text.

Deletes the thesaurus `Name`. Any profiles using this thesaurus will have their setting properly cleared.

5.20 Thunderstone ISAPI Proxy Module

5.20.1 Overview

The Thunderstone ISAPI Proxy Module is an IIS add-on that allows you to proxy requests through another machine.

Users' search requests are not made directly to Webinator, but to the Proxy Module, which then passes the request along to Webinator.

The Proxy Module also has an optional AuthProxy, which allows the use of automatic Active Directory credentials when authenticating Webinator search results. Internet Explorer will be configured to automatically pass along the authentication information, allowing for Single Sign On. Webinator communicates with the Proxy Module to authorize the results.

5.20.2 Requirements

For the Proxy Module:

The Thunderstone Proxy Module must be installed on a machine with IIS.

For the Proxy Module with the authProxy:

The Proxy Module with authProxy requires IIS 6, which is only available on Windows 2003 or later. The machine will need to be added to Internet Explorer's `Local Internet` zone, so if you have a server already listed there, you may want to use it.

This may be the same machine as Webinator is installed on (if it is Windows 2003 or later with IIS 6 or later), and the machine the Proxy Module is installed on must be a member of the Active Directory domain.

Ensure that the machine that the authProxy is being installed on is a secure machine. Because the Proxy Module is dealing with authorization functionality, a user with Administrative privileges could potentially tamper with operations.

5.20.3 Installing the Proxy Module

Before installing the proxy module the only thing you need to know is:

- The full hostname of the Webinator machine (eg. `thunderstone.mysite.com`) that the Proxy Module will be communicating with.

You can download the installer that contains the Proxy Module and authProxy from the Webinator machine by going to the `Maintenance` section, selecting `Extra Downloads`, then

`Thunderstone Proxy Module`, and finally click the `Download proxyModuleInstaller.exe` link for the installer. Once downloaded, the installer must be run on the Windows machine that you wish to make the proxy.

When installing you will be asked for a few items:

- `Destination Location` - This is where the actual DLL for the proxy module and its supporting files are placed. The directory `windows\system32\inetsrv` is recommended by default.
- `Configuration Directory` - This is the path that will be used for the IIS virtual directory that the proxy module is assigned to. Its actual location does not matter, as the proxy module will intercept all requests, but IIS still requires that all virtual directories point to a real path. The directory `Program Files\Thunderstone Software\Thunderstone ISAPI Proxy Module` is suggested by default.
- `Target` - The full hostname of the Webinator machine that this Proxy Module should connect to.
- `Active Directory Authentication` - If the Target you entered supports using the authProxy, you will be asked if you'd like to use Active Directory Authentication.

  If you are using the Proxy Module simply to allow access to Webinator through the proxy, choose "No".

  If you're using the Proxy Module to enable Single Sign On behavior in an Active Directory environment, choose "Yes".

5.20.4 Post-Install Setup

If you're using the authProxy, there are some configuration steps that must be manually performed, as they occur on machines other than the Proxy Module's machine. Please perform these before attempting to use the Proxy Module with authProxy.

Grant "Trust for Delegation" to the proxy machine

The machine that runs the Proxy Module & authProxy must be marked as `trusted for delegation` by the Active Directory domain controller. This is necessary for the proxy to automatically "pass along" the users' authentication to the searched web sites.

- **On the domain controller**, go to `Start - Programs - Administrative Tools - Active Directory Users and Computers`.
- Choose `Computers` on the left.
- Locate the computer that is running the Proxy Module, right-click on it, and choose `Properties`.
- Check `Trust computer for delegation`. A message box warning you that "this is a security-sensitive operation and it should not be done indiscriminately" will pop up. Click `OK`.

- Click `OK` to close the machine's properties, and close the `Active Directory Users and Computers` window.

Configuring Internet Explorer for Passing Credentials

The Proxy Module & authProxy machine must be listed in Internet Explorer's `Local Internet` security zone for all computers using it in order to function properly. If it is not in the `Local Internet`, then credentials will not be automatically provided. Even if the credentials are entered manually, the Proxy Module cannot authenticate with results when not listed in Internet Explorer's `Local Internet`.

If the Proxy Module machine is already in the `Local Internet` settings, you may skip this step.

The following steps adds the Proxy Module machine to Internet Explorer's `Local Internet`:

- Start Internet Explorer.
- Choose `Tools` from the menu, and select `Internet Options`.
- Choose the `Security` tab.
- Choose `Local Internet` from the list of zones.
- Click the `Sites` button to edit the local internet.
- Click `Advanced` to manually add a site.
- *Un*check `Require server verification (https:) for all sites in this zone`
- Enter the full hostname of your proxying machine, for example `proxyMachine.example.com`.
- Click `Add` to add the site to the Trusted Sites.
- Click `Close` to close the `Advanced` window.
- Click `OK` to close the `Local Intranet` window.
- Click `OK` to close the `Internet Options` window.

Internet Explorer is now configured to pass credentials to the proxy machine. This is a per-user configuration, and will need to be configured for any user that is authenticating via the Proxy Module.

Configuring Webinator

There are three things that must be done in Webinator to configure it to accept authentication information from the authProxy, one of them global and two on a per-profile basis.

Add the Proxy Machine to Cluster Members

The IP address of the machine that the authProxy is installed on must be added to the list of `Cluster Members` to tell Webinator to trust the proxy machine.

- Choose `Maintenance` on the left.
- Choose `System Wide Settings`, under the `Search Appliance Settings` section.
- Enter the proxy machine's IP address in the `Cluster Members` field on a new line.

Make the Target Profiles Visible

The profiles that you want to search with the authProxy must be set `Visible`, which enables the profile for things like meta searching and the proxy module.

- Select the profile in the `Profiles` page.
- Choose `Search Settings` on the left.
- Set the `Visible` setting to `Y`.
- Click `Update` at the bottom.

Enable Results Authorization for the Target Profile

Also, Results Authorization must be enabled for the target profile, if it's not already enabled.

- Select the profile in the `Profiles` page.
- Choose `Search Settings` on the left.
- Set the radio button for `Authorization Method` to `Basic/NTLM/file` (occurs beneath `Login Cookies` and `Login URL`).
- Click `Update` at the bottom.

5.20.5 Manually Configuring the Proxy Module

This section describes how to manually configure IIS for use of the Thunderstone Proxy Module. The will be described in more detail in the next section. This is **not** necessary for normal operations - these actions are normally performed automatically by InstallShield upon installation. These steps are only necessary if IIS's configuration gets wiped out and needs to be redone.

The Thunderstone Proxy Module is an ISAPI Extension, two if using the authProxy. They are assigned as Global Application Maps to Virtual Directories in IIS. All requests to the directories are not be served from the file system that the virtual directory points to, but instead go through the Proxy Module `dll`s.

One virtual directory is required per extension: `texis`, which gets assigned `proxyModule.dll`, and `authProxy`, which gets assigned `authProxy.dll`.

If using the authProxy, `texis` must have anonymous access disabled and Integrated Authentication enabled, while `authProxy` must have anonymous access allowed (which is allowed by default).

These are the steps that must be done if you are manually setting up IIS for using the Proxy Module. **Note that these are done automatically by the InstallShield wizard** and do *not* need to be manually done under normal circumstances.

- Open the IIS Configuration
  - Right click on `My Computer` on the desktop.
  - Select `Manage...`
  - Open `Services and Applications` in the tree.
  - Open `Internet Information Services`.
  - Open `Web Sites`.
  - Select the web site you want to add the Proxy Module to (most likely `Default Web Site`).
- Add the `texis` virtual directory
  - Right click on the web site and select `New -> Virtual Directory...`
  - The `Virtual Directory Creation Wizard` opens. Click `Next>`.
  - In the `Alias` box, enter `texis` and click `Next>`.
  - In the `Path` box, enter the real physical path you want the virtual directory to map to, and click `Next`. the Proxy Module uses the directory `<INSTALLDIR>/etc/ISAPI-virtualdir` by default.

    Note that it doesn't matter what directory is selected. This directory will never be used because all requests will be intercepted by the Proxy Module. The only reason a directory must be selected is because IIS insists that *all* virtual directories map to a real physical location.
  - At the `Virtual Directory Access Permissions` screen, just click `Next` to complete the wizard, as we won't be using any of the permissions.
  - Click `Finish` to complete the wizard and return to the `Computer Management` window.
- Apply `proxyModule.dll` as a Wildcard Application Map
  - Right-click on the newly created virtual directory and select `Properties`.
  - The lower half of the properties window is labeled `Application Settings`. Click `Create` to make a custom set of application settings for this virtual directory.
  - After clicking `Create`, the `Configuration` should no longer be disabled. Click `Configuration`.
  - The lower half of the new `Application Configuration` window details Wildcard Application Maps, which is currently empty. Click `Insert`.

  - Next to the `Executable` field, click the `Browse` button and locate `ProxyModule.dll`, which is in the directory you installed the Proxy Module to.
    * (The default location for this file in `C:\windows\system32\inetsrv` on 32bit windows, `C:\windows\SysWOW64\inetsrv` on 64bit windows).
  - **Uncheck** the box next to `Verify that file exists`, and click `OK`.
  - `ProxyModule.dll` will now be in the list of Wildcard Application Maps. Click `OK` to close the `Application Configuration` window.

- Configure `texis` for authentication

  Only necessary if using the authProxy.

  - While still in the `texis Properties` window for the new virtual directory, Select the `Directory Security` tab.
  - In the top section, labeled `Authentication and Access Control`, click the `Edit...` button.
  - **Un**check `Enable Anonymous Access` and ensure that `Integrated Windows Authentication` is **checked**.
  - Click `OK` to close the `Authentication Methods` window.
  - Click `OK` to close the `texis Properties` window.

  Now we need to create the `authProxy` directory in a similar manner, although it doesn't need anonymous access disabled.

  If you're not using the authProxy, please skip to the "Add the Proxy Module files to IIS' list of allowed extensions" section below.

- Add the `authProxy` virtual directory

  - Right click on the website and select `New -> Virtual Directory...`
  - The `Virtual Directory Creation Wizard` opens. Click `Next>`.
  - In the `Alias` box, enter `authProxy` and click `Next>`.
  - In the `Path` box, enter the real physical path you want the virtual directory to map to, and click `Next`. the Proxy Module uses the directory `<INSTALLDIR>/etc/ISAPI-virtualdir` by default.

    Note that it doesn't matter what directory is selected. This directory will never be used because all requests will be intercepted by the Proxy Module. The only reason a directory must be selected is because IIS insists that *all* virtual directories map to a real physical location.
  - At the `Virtual Directory Access Permissions` screen, just click `Next` to complete the wizard, as we won't be using any of the permissions.
  - Click `Finish` to complete the wizard and return to the `Computer Management` window.

- Apply `authProxy.dll` as a Wildcard Application Map

  - Right-click on the newly created `authProxy` virtual directory and select `Properties`.

  - The lower half of the properties window is labeled `Application Settings`. Click `Create` to make a custom set of application settings for this virtual directory.
  - After clicking `Create`, the `Configuration` should no longer be disabled. Click `Configuration`.
  - The lower half of the new `Application Configuration` window details `Wildcard Application Maps`, which is currently empty. Click `Insert`.
  - Next to the `Executable` field, click the `Browse` button and locate `authProxy.dll`, which is in the directory you installed the Proxy Module to.
    * (The default location for this file in `C:\windows\system32\inetsrv` on 32bit windows, `C:\windows\SysWOW64\inetsrv` on 64bit windows).
  - **Uncheck** the box next to `Verify that file exists`, and click `OK`.
  - `authProxy.dll` will now be in the list of Wildcard Application Maps. Click `OK` to close the `Application Configuration` window.

- Add the Proxy Module files to IIS' list of allowed extensions

  By default IIS blocks all ISAPI extensions as a security measure. The Proxy Module must be explicitly allowed in IIS' configuration.

  - Back in the `Computer Management` window, open `Web Service Extensions`, underneath `Internet Information Services`.
  - The right side of the window should now have a list of rules. Right-click beneath the existing rules and select `Add a new web service extension...`
  - In the `Extension Name` field, enter `Thunderstone Proxy Module`.
  - Next to the `Required files` text area, click the `Add...` button.
  - Next to `Path to file:`, click `Browse...` and locate `ProxyModule.dll`, (just as in the previous set of instructions), and click `OK` to close the `Add File` dialog.
  - If using the authProxy, click `Add` again, and this time choose the `authProxy.dll` file.
  - Check the box next to `Set extension status to Allowed`, and click `OK` to close the window.

IIS is now set up properly to use the Proxy Module. Note that if using the authProxy, changes still need to be made to the network and Webinator, as detailed in the **Post-Install Setup** and **Configuring Webinator** sections, on pages 131 and 132, respectively.

5.20.6 Troubleshooting the Proxy Module Authentication

This section details some troubleshooting steps you can go through if Proxy Module Authentication isn't working.

Review Installation Steps

There are a number of steps that must be manually performed after the Proxy Module install (due to them being done on different machines or as different accounts). Please ensure the following steps have been performed:

- Grant "Trust for Delegation" to the proxy machine (p. 131)
- Configuring Internet Explorer for Passing Credentials (p. 132)
- Configuring Webinator (p. 132)

Machine names and SPNs

A Service Principle Name (SPN) is the name by which a client uniquely identifies an instance of a service. By default your IIS machine has SPNs for its hostname, such as myServer, and its Fully Qualified Domain Name (FQDN), such as myServer.branch.example.com.

If the proxy machine is accessed by a name other than either of these, such as myServer.example.com, otherName.company.com, or its IP address, then Active Directory authentication will not work. Your choices are:

- Access the machine using either its host name or FQDN.
- Register an additional SPN for the proxy machine on the domain controller.

SPNs can be viewed and changed with the `setspn.exe` tool, which Microsoft provides as part of its `Windows 2003 Support Tools`. This package is available on the Windows 2003 CD, or, at the time of writing, at `http://go.microsoft.com/fwlink/?LinkId=100114`.

DelegConfig Diagnostic Tool

For general Active Directory troubleshooting, Thunderstone has found the `DelegConfig` tool to be handy. It's an ASP.NET application used to help troubleshoot and configure IIS and Active Directory to allow Kerberos and delegating Active Directory credentials. At the time of this writing, it is available at:

`http://www.iis.net/downloads/default.aspx?tabid=34&g=6&i=1434`

Thunderstone did not create DelegConfig, and does not make any guarantees to its accuracy or availability.

Launch IE as a different user

When testing multiple user accounts, you can Right-Click on an Internet Explorer shortcut, select `Run As...`, and enter another user's credentials to launch Internet Explorer as that user. This way you can try out IE as them without needing to go through the full login process.

- Note that the user you run IE as must have logged in at least once in order to properly configure IE, otherwise the `Add to Local Internet` dialog will not function properly.
- If there is no `Run As...` option, create a new shortcut to Internet Explorer and use that - the IE icon on that's on the desktop by default is not a normal shortcut.
- You **cannot** test another user by entering their credentials within the browser when accessing the proxy machine. Credentials MUST be passed to the proxy machine automatically by IE by having the proxy machine in IE's `Local Internet` settings.

  If credentials are entered manually in IE when accessing the proxy machine (even the credentials of the current user), then IE will use NTLM to authenticate with the proxy machine instead of Active Directory, which prevents proper delegation from occurring.

Chapter 6

Reference

6.1 Database and File Usage

Webinator maintains a database that contains text from HTML pages, links to other pages, and a list of categories.

When the Webinator walker runs it creates a new database, under your specified data directory, to hold the new walk. It then dispatches a separate process for each web site it needs to visit and another to handle all of the "Single Pages". Each of these retrieves all of the pages in it's base list and stores the text of the HTML page to the html table and the hyperlinks to the refs table. All of the desirable URLs from the page that have not been seen before are placed into an internal "todo" list. After all of the base URLs are processed the process repeats with the internal todo list. When there's nothing left in the todo list processing is complete.

Once all of the walking is complete the indices needed for searching are created on the data. Then the new database is flagged as the "live" one and the old database is deleted. Therefore your disk must have sufficient space for 2 complete databases plus temporary space used during the indexing step.

The databases are stored under your specified data directory. The databases are called `db1` and `db2`. Webinator alternates between using these two names.

Note that the above applies to a walk type of `New`. During a walk type of `Refresh` only one database, the "live" one, is used.

Webinator also maintains a file containing the detailed report for each walk. This file has the same name as the database with `.long` appended to the end. Also, a single file called `summary` is maintained with short summary information about the state of the database.

Given a data directory named `.../default` there may also be the following:

`.../default/db1` an actual walk database

`.../default/db2` an actual walk database

`.../default/db1.long` detailed walk report. Displayed when viewing `Walk Status`

`.../default/db2.long` detailed walk report. Displayed when viewing `Walk Status`

`.../default/summary` summary walk report. Displayed as `Walk summary` when viewing `Walk Settings`

Webinator, being based on Texis, also has the notion of a global "default" database. This database resides in the installation directory. On Unix it is called `INSTALLDIR/texis/testdb`. On Windows it is called `INSTALLDIR\texis\testdb`. This database is used to hold all of the profile and account settings. It does not contain any walked data. It is recommended that you *not* use this as your data directory.

Each setting has a record in the `options` table of the default database. See section 6.3 (p. 142) for the list of fields in the table. At each complete rewalk the current options settings are copied into an options table in the walk database. These options are not changed as settings are modified and are not otherwise used unless a search is performed setting the database with `db` instead of setting the profile with `pr`.

6.2 Walk Database Tables and Fields

Table 6.1: Fields in `html` table

Field	Description
`id`	Unique record id
`Hash`	Document hash for duplicate content detection
`Size`	Size of retrieved raw document (ie. HTML)
`Visited`	The date the page was modified (or fetched if modified not set)
`Dlsecs`	The number of seconds needed to fetch the page
`Depth`	The number of URLs traversed to reach the page
`Url`	The URL of the real HTML page
`Title`	The title of the page
`Body`	The formatted textual content of the page, in Storage Charset (UTF-8)
`Keywords`	The `keywords` meta data from the page
`Description`	The `description` meta data from the page
`Meta`	Other meta data from the page, separated by newlines
`Catno`	List of categories to which the URL belongs
`CatnoLowest`	Lowest `Catno` value
`Modified`	The date the page was modified
`NextCheck`	The date the page should next be refreshed
`Views`	The number of times this URL has been viewed (shown in results)
`Clicks`	The number of times this URL has been clicked (in results)
`CTR`	Click-through ratio
`Pop`	Popularity (number of pages linking to this page)
`MimeType`	MIME type of original page
`Charset`	Character set of page as stored (usually Storage Charset)

Table 6.2: Fields in `refs` table

Field	Description
`Url`	The URL of the HTML page
`Ref`	The URL of a reference (link) on the HTML page

Table 6.3: Fields in `categories` table

Field	Description
`Catno`	The number for the category
`OverlapsLower`	Y if some member(s) also in a lower category
`Url`	The URL pattern for the category
`Category`	The name of the category

Table 6.4: Fields in `error` table

Field	Description
`Url`	The URL of an HTML page that could not be retrieved
`Reason`	The reason it could not be retrieved
`id`	Unique record id (includes timestamp info).

Table 6.5: Fields in `querylog` table (if query logging enabled)

Field	Description
`id`	Contains the date and time of the query (unique record id)
`Client`	The hostname of the web client that performed the query
`Query`	The user's query as entered

6.3 Options Table Fields

These are the options table fields (maintained in the default database):

Table 6.6: Fields in `options` table

Field	Description
`id`	Unique id for the record
`Profile`	The name of the profile that the record belongs to
`Name`	The name of the setting
`Type`	The data type of the setting (always `String`)
`String`	The value of the setting
`Int`	Unused
`Float`	Unused
`Strlist`	Unused

You can look at the `SYSCOLUMNS` and `SYSINDEX` tables of the database for details about the field types, sizes, and indices.

6.4 Customizing the Search

You may make common changes to Webinator's search appearance by using `Search Settings` from the administrative interface main menu. But you are not limited to those features. You may change any and all aspects of the search program's appearance and behavior by modifying the supplied `search` script or writing an altogether new one.

For details on programming with Texis Web Script (Vortex), see the manual at the Thunderstone web site, `http://www.thunderstone.com/site/vortexman/`.

The following section describes some important points about the internals of the default search script that comes with Webinator. The search script is fairly heavily commented to aid in finding your way around within it.

The `init` function is called from every entry point. It is a good place to put settings that should always (or often) apply. This function understands the old (version 2) style specification of database by the `db` variable as well as the current method of extracting the database name from the profile named by the `pr` variable.

The `top` function displays the common HTML for the beginning of every page generated by the search script. This does not include the search form. This function is where you would place styles and navigation menus.

The `bottom` function is the complement to the `top` function. It displays the common HTML footer for the end of every page.

The `showform` function displays the search form with all current settings indicated.

The `qpar` and `fpar` functions process the user's form submission and apply appropriate search settings.

The `credit` function displays the Thunderstone credit on the search results. This is required for free users but may be changed or emptied for paid users.

The `result` function is called for each matching record to display. It then calls the configured `result*` function to generate the desired output style.

The `mlt` function is called to setup the search when the end user selects "`Find Similar`" (aka More Like This).

The `similar` function may be called directly to find pages within the database that are similar to the content of the URL specified. It has the same concept of "`Find Similar`" but will work on any specified URL, not just those displayed as the result of a search. It would be invoked something like this on any HTML page.

```
<a href="/cgi-bin/texis/webinator/search/similar.html?↪
   ↪pr=default&ref=http://somesite/somepage.html">↪
   ↪Find pages similar to somepage.html</a>
```

or

```
<a href="/scripts/texis.exe/webinator/search/similar.html?↪
   ↪pr=default&ref=http://somesite/somepage.html">↪
   ↪Find pages similar to somepage.html</a>
```

Set "`default`" above to the search profile you're using.

It will lookup that URL in the database or, if it's not in the database, fetch it from the webserver. It will then search the database looking for indexed pages similar to the specified page.

The `main` function is the standard Vortex default entry point. This is the function that is first called when users click "Submit" on the search form.

The `search` function does the core work of finding matching documents within the database. It calls `showform` and `qpar` then starts searching. For every match the `result` function is called. The `summary` function is called before the first match is displayed to display the search results summary. It is called again at the end of the results list.

The `putmsg` function handles errors that may occur and displays them in a somewhat more user friendly fashion. See the Vortex manual for details about how `putmsg` is used to capture errors.

6.5 Customizing the Walker

You may make many changes to Webinator's walk behavior by using `Walk Settings` from the administrative interface main menu.

But you are not limited to these features. You may change any and all aspects of the walker's behavior by modifying the supplied `dowalk` script. (The `webinatoradmin` script supplied with version 4 and earlier releases has been combined into `dowalk` for atomicity.)

For details on programming with Texis Web Script (Vortex), see the manual at the Thunderstone web site, `http://www.thunderstone.com/`.

The following describes some important points about the internals of the `dowalk` script that comes with Webinator. The `dowalk` script is fairly heavily commented to aid in finding your way around within it.

The `dowalk` script actually consists of 2 Vortex script files concatenated. The first part contains the walker/indexer and settings reading code. The second part of the file provides the management interface that is used from a web browser.

The `dispatch` function is the primary external entry point for performing a new walk. It load settings, sets up logging and databases, then invokes other processes in parallel (according to maximum servers setting). When all of the walking is complete it removes commonality from pages (if that option is set), creates the indices needed for searching the database, then makes the new database live and deletes the old database.

The `stop` function is an external entry point that is used to signal (using `<loguser>`) a walk that is in progress that it should stop. The walkers check for this signal (using `<userstats>`) at various points and will quit when it is detected.

The `reindex` function is an external entry point that is used to drop and recreate the Metamorph index on the html table. This is needed after changing the word definition expressions.

The `remakeindex` function is an external entry point that is used to drop and recreate all indices on the database. It it only for use if one or more non-Metamorph indices get corrupted by disk errors or such.

The `recat` function is an external entry point that is used to recategorize the html table based on the

current (presumably changed) categories.

The `ifmodified` function is an external entry point that is used to tell the dispatcher to run only if chkneedwalk indicates a walk is needed.

The `usage` function is called when you invoke `dowalk` incorrectly and prints a terse summary or correct usage options.

The `doplugin` function handles files that are not HTML or text, such as PDF and MSWord. It determines the correct options for `anytotx` based on the fetched page's MIME type or extension. It then calls the `dofilt` function which actually runs `anytotx` to perform the conversion to text and the extraction of meta information such as Title. It will make up a title for the document if none is returned by `anytotx`.

The `settings` function calls the defaults, readsettings, and applysettings functions, in order. This function is called by most entry points to get default and current settings for a given profile before proceeding with any work.

The `updatemmindex` function is called (sometime after having called settings) to create or update the Metamorph index on the html table.

The `maketables` function is called (sometime after having called settings) to create all of the Webinator tables. This function does nothing for Webinator-only licenses. For Webinator-only licenses the tables are created automatically by Texis when the database is created. The schema may not be changed.

The `walk` function is the core which walks all desired URLs on a single site. It always processes breadth first (ie it gets all URLs at a given depth before proceeding to the next level down). Any desired URLs that reside on a different site are placed into the database's todo table for processing by the dispatcher.

The `fetchset` function is used in various places to fetch one or more URLs (using the maximum threads setting) simultaneously.

The `manglepage` function is called before extracting text and hyperlinks from an HTML page. It allows the page to be modified before processing. This is where the ignore/keep tags are handled.

The `getrobotstxt` function fetches the robots.txt file from a given site and checks for any exclusions for Webinator. These exclusions are later added to the list of URL rejection patterns.

The `chkneedwalk` function is called to check if a rewalk is required. It fetches the page to see if the modification date has changed. Or, if the web server does not provide a modification date it compares the content to what it was previously. It sets an internal flag if a rewalk is needed.

The `putmsg` function intercepts error messages to provide special handling for some, and recording of most.

The `go` function is an external entry point used by the dispatcher when it starts up child processes to walk a specific site or set of URLs.

The `singles` function is an external entry point that is used to fetch all of the single page URL. It is called by the dispatcher as the first parallel process. Therefore single pages will generally be fetched earliest in a new walk.

The `rmlocks` function is used to remove any stale locks and monitor processes on a database and dismantle the locking structure. This is done before physically removing a database from the system.

The `geturl` function is a utility function that may be used to find out what the walker will think about a given URL using the current walk settings. It is invoked as follows:

```
texis profile=PROFILE top=THEURL dowalk/geturl.txt
```

This can generate a lot of output for a page of any size so you may want redirect it to a file that you can examine with your favorite viewer/editor.

```
texis profile=PROFILE top=THEURL dowalk/geturl.txt >FILE.txt
```

The `getrobots` function is a utility function that may be used to find out what the walker will think about a given robots.txt using the current walk settings. It is invoked as follows:

```
texis profile=PROFILE top=THEURL dowalk/getrobots.txt
```

This can generate a lot of output for a page of any size so you may want redirect it to a file that you can examine with your favorite viewer/editor.

```
texis profile=PROFILE top=THEURL dowalk/getrobots.txt >FILE.txt
```

6.6 Texis ISAPI

6.6.1 Overview

Texis ISAPI uses IIS's Internet Server API (ISAPI) to allow Webinator on IIS to use a Unix-style web address (no .exe in the path). Many URL-scanning programs see having `.exe` in the address path as an indication of an attempted exploit, so removing this can be desirable.

The other advantage is that Texis ISAPI can be installed on an IIS machine different from the machine that Webinator is installed on. Webinator can be installed on a dedicated machine inside your intranet, and your IIS web server can use ISAPI to display its content.

Installation and usage of Texis ISAPI does not in any way prevent usage of the conventional CGI method. Both can be used simultaneously, if desired.

6.6.2 How it Works

The Texis ISAPI software acts as a pass-through. IIS is configured to give requests to Texis ISAPI, which in turn passes it along to Webinator. Texis ISAPI receives Webinator's response, and passes that response back to the web browser.

There are two types of ISAPI programs: filters and extensions. Texis ISAPI contains both a filter and an extension in `ProxyModule.dll`, although you will only use one or the other (which one depends on

which version of IIS you're running). Both the Texis ISAPI filter and Texis ISAPI extension offer the same features and functionality, they only differ in how they are implemented in and communicate with IIS (the installer is able to set up either for you automatically).

- **IIS 5 or earlier**
  In IIS 5 or earlier, an ISAPI Filter is used. This is installed as a global filter, as is required of all filters that use `SF_NOTIFY_READ_RAW_DATA`. It is invoked for every request, but takes no action unless the request begins with `/texis`. If the request does, Texis ISAPI takes control of the request and processes it appropriately.

- **IIS 6 or later**
  On IIS 6 or later, `SF_NOTIFY_READ_RAW_DATA` for filters is explicitly denied, so Texis ISAPI uses an ISAPI Extension mapped as a Wildcard Application Mapping. The installer accesses the default site and creates a new virtual directory, `texis`. It creates custom Application Settings for that virtual directory, and adds Texis ISAPI's DLL file as a Wildcard Application Map. This means that any request that comes to that virtual directory (i.e. `/texis/...`) will not map to a real file location, but will instead be handed off to Texis ISAPI. The installer also adds Texis ISAPI as an "allowed" extension to the IIS Web Service Extensions Restriction List.

  Wildcard Application Maps are not available prior to IIS 6, so the filter is still used for earlier versions.

6.6.3 Settings for Texis ISAPI

Texis ISAPI must be configured for how to contact Webinator. It needs a port number and a host (if it's not the localhost). Regardless of whether loading settings was successful or not, an entry will be made in the `Application` Event Log detailing either what settings were loaded, or why settings couldn't be loaded.

Texis ISAPI will first attempt to read a port number from a locally installed Webinator's conf/texis.ini.

Reading values from conf/texis.ini

Texis ISAPI does this by first looking in the registry for an `InstallDir` value in the `HKEY_LOCAL_MACHINE\Software\Thunderstone Software` key to locate the installation path.

It tries to read the following values from the `[Httpd]` section of `conf/texis.ini`:

`Port`: This setting serves double-duty: it tells monitor web server what port it should listen on, and it tells Texis ISAPI what port it should use to connect to the monitor web server. The default `10700` should be fine.

If it is unsuccessful in reading values from `conf/texis.ini`, it will then attempt to read values from the registry.

Reading values from the Registry

Texis ISAPI will attempt to read the following values from the registry key `HKEY_LOCAL_MACHINE\Thunderstone Software\ISAPI`:

`port` (DWORD): the port that Texis ISAPI should use to connect to the remote monitor web server.

`host` (String): the hostname of the webinator machine that Texis ISAPI should use. If no hostname is found, `localhost` is assumed.

If Texis ISAPI is unable to read a `port` value from the registry, then it will disable itself and makes an Event Log entry detailing why it couldn't read from conf/texis.ini and why it couldn't read from the registry.

6.6.4 IIS Manual Configuration

This section describes how to manually configure IIS for use of Texis ISAPI. This is **not** necessary for normal operations - these actions are performed automatically by InstallShield upon installation. These steps are only necessary if IIS's configuration gets wiped out and needs to be redone.

IIS 5.X or earlier

In IIS 5 and earlier, Texis ISAPI is applied as an "ISAPI Filter". It is applied as a global filter that chooses to only intercepts certain requests. What path it should watch for and what webinator it should contact are configured via conf/texis.ini or the registry, as described above.

To manually add the Texis ISAPI Filter to IIS 5.x:

- Open the IIS Configuration
  - Right click on `My Computer` on the desktop.
  - Select `Manage...`.
  - Open `Services and Applications` in the tree.
  - Open `Internet Information Services`.
- Bring up ISAPI properties
  - Right-click on the `Web Sites` Folder (not an individual site).
  - Select `Properties`.
  - In the new `Web Sites Properties` window, select the `ISAPI Filters` tab.
- Add the Texis ISAPI Filter
  - Click the `Add...` button.
  - In the new `Filter Properties` window, enter 'Texis ISAPI' in the `Filter Name` field.
  - Click the `Browse...` button next to the `Executable` field, and browse to your `ProxyModule.dll` file.
    - (By default Webinator places this file in `C:\windows\system32\inetsrv` on 32bit windows, `C:\windows\SysWOW64\inetsrv` on 64bit windows).
  - Click `OK` to close the `Filter Properties` window.

  - Click `OK` to close the `Web Site Properties` window.

- Restart IIS
  - Right-click on `Internet Information Services`.
  - Open the sub-menu `All Tasks`
  - Select `Restart IIS...`
  - in the 'What do you want IIS to do?' box, select `Restart Internet Services on [YourComputer]`
  - Click `OK` to restart IIS.

IIS is now fully configured to use the Texis ISAPI Filter. You should have an entry in the Event Log detailing the results of Texis ISAPI's attempt to read settings.

IIS 6 or later

In IIS 6 and later, Texis ISAPI is used as an ISAPI Extension, not an ISAPI Filter. The extension is applied as a "wildcard application map" on a virtual directory. This means that all requests that come to the specified virtual directory will **not** map to the real location of the virtual directory, but instead be processed by Texis ISAPI.

For IIS 6 to use Texis ISAPI, there are two separate things that need to be done. A virtual directory needs to be set up to use `ProxyModule.dll`, and Texis ISAPI needs to be added it to IIS 6's `Allowed Extensions` list.

To create a virtual directory that invokes `ProxyModule.dll` on IIS 6:

- Open the IIS Configuration
  - Right click on `My Computer` on the desktop.
  - Select `Manage...`
  - Open `Services and Applications` in the tree.
  - Open `Internet Information Services`.
  - Open `Web Sites`.
  - Open the website you want to add Texis ISAPI to (most likely `Default Web Site`).

- Add a new virtual directory
  - Right click on the website you want to add Texis ISAPI to, and select `New -> Virtual Directory...`
  - The `Virtual Directory Creation Wizard` opens. Click `Next>`.
  - In the `Alias` box, enter `texis`, and click `Next`.

  - In the `Path` box, enter the real physical path you want the virtual directory to map to, and click `Next`. Webinator uses the directory `<INSTALLDIR>/etc/ISAPI-virtualdir` by default.

    Note that it doesn't matter what directory is selected. This directory will never be used because all requests will be intercepted by Texis ISAPI. The only reason a directory must be selected is because IIS insists that *all* virtual directories map to a real physical location.
  - At the `Virtual Directory Access Permissions` screen, just click `Next` to complete the wizard, as we won't be using any of the permissions.
  - Click `Finish` to complete the wizard and return to the `Computer Management` window.

- Apply `ProxyModule.dll` as a Wildcard Application Map

  - Back in the Right-click on the newly created virtual directory and select `Properties`.
  - The lower half of the properties window is labeled `Application Settings`. Click `Create` to make a custom set of application settings for this virtual directory.
  - After clicking `Create`, the `Configuration` should no longer be disabled. Click `Configuration`.
  - The lower half of the new `Application Configuration` window details Wildcard Application Maps, which is currently empty. Click `Insert`.
  - Next to the `Executable` field, click the `Browse` button and locate `ProxyModule.dll`.
    * (By default Webinator places this file in `C:\windows\system32\inetsrv` on 32bit windows, `C:\windows\SysWOW64\inetsrv` on 64bit windows).
  - **Uncheck** the box next to `Verify that file exists`, and click `OK`.
  - `ProxyModule.dll` will now be in the list of Wildcard Application Maps. Click `OK` to close the `Application Configuration` window, and `OK` to close the virtual directory's properties window.

To add Texis ISAPI to IIS' list of allowed extensions on IIS 6:

By default IIS blocks all ISAPI extensions as a security measure. Texis ISAPI must be explicitly allowed in IIS' configuration.

- Back in the `Computer Management` window, open `Web Service Extensions`, underneath `Internet Information Services`.
- The right side of the window should now have a list of rules. Right-click beneath the existing rules and select `Add a new web service extension...`
- In the `Extension Name` field, enter `Texis ISAPI`.
- Next to the `Required files` text area, click the `Add...` button.
- Next to `Path to file:`, click `Browse...` and locate `ProxyModule.dll`, (just as in the previous set of instructions), and click `OK` to close the `Add File` dialog.

- Check the box next to `Set extension status to Allowed`, and click `OK` to close the window.

IIS 6 should now be properly set up to use Texis ISAPI. Note that the extension doesn't get loaded until a request is made, so no entry will be made in the Event Log about startup until at least one request that uses the extension has been made.

6.7 CGI Mapping by Vortex File Extension

The following sections detail how to manually configure some web servers to run Webinator's Vortex scripts by URL file extension (eg. `.vs` or `.vtx`), instead of by URL directory (eg. `/cgi-bin/texis` or `/scripts/texis.exe`). This will allow[1] URLs such as `/webinator/dowalk.vs` to be run, instead of `/scripts/texis.exe/webinator/dowalk.vs`.

Notes:

- Windows: if the Texis ISAPI filter/extension is being used (p. 146), this procedure may not be needed, nor does it affect those URLs.
- This procedure may already have been performed by the Webinator installation wizard (especially under Windows).
- This procedure requires Vortex version 6 or later, with the default configuration of `.vs` in `Vortex Source Extensions` in `conf/texis.ini`. (Run `texis -version` from a command prompt to determine your version.)

6.7.1 Microsoft IIS

To manually configure Microsoft IIS to map Vortex scripts by file extension (`.vs`), Vortex version 6, and IIS version 6.0 or later are required. Earlier versions may not map extended-Vortex-syntax URLs (eg. with `/func.html` appended) properly. Note that this procedure may already have been performed by the Webinator installation wizard. This procedure should be performed as an administrator. It is intended for Windows Server 2003, but should be similar for other Windows operating systems as well:

1. **Open the IIS configuration:**
   (a) Right-click on `My Computer` on the desktop.
   (b) Select `Manage...` to open the `Computer Management` application.
   (c) Expand the `Services and Applications` tree (click on its plus-sign).
   (d) Expand the `Internet Information Services` tree.
2. **Add `texis.exe` to Web Service Extensions:**

[1] Depending on the license obtained from Thunderstone; this generally requires at least a purchased license.

(a) Double-click on `Web Service Extensions`.

(b) If an item named `Texis` already exists under the `Web Service Extension` list, double-click it, and click the `Required Files` tab.
Otherwise, if an item named `Texis` does *not* already exist under the `Web Service Extension` list, click the `Add a new Web service extension...` link and enter `Texis` into the `Extension name:` box.

(c) Click `Add...` on the `Properties` or `New Web Service Extension` dialog.

(d) Enter the path to the `texis.exe` executable in the `Path to file:` box. This is your Webinator install directory plus `\texis.exe`, eg. `C:\morph3\texis.exe` or `"C:\Program Files\Thunderstone Software\Webinator\texis.exe"`. Double-quote the path if it contains spaces.

(e) Click `OK` to close the `Add file` dialog box. The `texis.exe` path should now be listed under `Files:` as `Allowed`, or under `Required files:`.

(f) Click `OK` to close the `Web Service Extension Properties - Texis` or `New Web Service Extension` dialog box.

3. **Add a file-extension application mapping to the web site:**

(a) Expand the `Web Sites` tree (on the left).

(b) Right-click the appropriate web site.

(c) Select `Properties...` from the popup menu.

(d) Click the `Home Directory` tab (at top of dialog).

(e) Click the `Configuration...` button (lower right).

(f) Scroll through the `Application extensions` list. If an `Extension` for `.vs` (Vortex) already exists, double-click it. If not, click `Add...` to create one.

(g) In the `Executable:` box, enter the *identical* path to the `texis.exe` executable that you entered above.

(h) In the `Extension:` box, enter `.vs` to map Vortex files with that extension in URLs to the `texis.exe` executable.

(i) Check `All verbs` and `Script engine`.

(j) *Un*check `Verify that file exists`[2]

(k) Click `OK` to close the `Add/Edit Application Extension Mapping` dialog.

(l) Click `OK` to close the `Application Configuration` dialog.

(m) Click `OK` to close the `... Web Site Properties` dialog.

4. **Close the `Computer Management` application**

6.7.2 Apache

To manually configure the Apache web server to map Vortex scripts by file extension (`.vs`), either a redirect handler can be added to run Vortex scripts (preferred), or the scripts themselves can be directly executed (alternate). The redirect-handler method is preferable, as the latter method has some drawbacks.

[2] If `Verify that file exists` were checked, Vortex URLs would result in 404 errors, because IIS would not find the Vortex files in the web site's document root as it expects. They are in Vortex's `ScriptRoot` dir instead.

Preferred Method: Redirect Handler

To configure Apache to run Vortex scripts by file extension (`.vs`) using the preferred redirect-handler method, Vortex version 6, and Apache version 2.1 or later are required[3]. This procedure is intended for Unix systems and should be performed by `root`. It configures `texis` as a redirect handler to run Vortex scripts. Consult your web server manual (or online at `http://www.apache.org/`) for details and consequences on the directives used in this procedure:

1. Find the existing URL (not filesystem dir) on your server that runs `texis`. Typically this is `/cgi-bin/texis`. If there is no existing URL to run `texis`, consult your web server manual and configure the web server to do so. Typically this is done with a directive similar to:
   `ScriptAlias /cgi-bin/ /var/www/cgi-bin/`
   Note that you may also have to copy or symlink the `texis` executable to `/var/www/cgi-bin` (though this is typically already performed by the Webinator installation wizard).
2. Open the Apache configuration file (typically `/etc/httpd/conf/httpd.conf`; see your web server manual) with a text editor.
3. Add the following two lines, preferably near (but outside) the existing CGI directives:

   ```
   Action texis-vortex /cgi-bin/texis virtual
   AddHandler texis-vortex .vs
   ```

   If your existing configuration uses a URL other than `/cgi-bin/texis` to run `texis`, then change the `Action` directive appropriately. Note that the `Action` directive requires that it *must* be a URL, not filesystem, path. Note that Apache version 2.1 or later is needed for the `virtual` keyword; with 2.0, omit it.
4. Save the configuration file and exit the editor.
5. Re-start the web server (typically with `/etc/init.d/httpd restart`; check your web server manual).

Note that despite the `virtual` keyword in the `Action` directive, Apache may still require that the parent dir(s) of Vortex scripts exist in the document root (even though Vortex scripts are typically in Vortex's `ScriptRoot` dir, ie. `texis/scripts` in the Webinator install dir). This is believed to be a bug in Apache; it was noted in version 2.2.4 at least. Therefore, you may have to create a `webinator` (or other) directory in your web server's document root (usually `/var/www/html`). This dir may have already been created by the Webinator installation wizard. If the parent dir(s) are missing in the document root, it may one potential cause of `404 Not Found` errors when attempting to run Vortex scripts by extension.

Alternate Method: Direct Execution

An alternate method for configuring Apache to run Vortex scripts by extension (`.vs`) is to have Apache execute the scripts directly, instead of using a redirect handler. This method is less desirable than the redirect-handler method (above) for several reasons:

[3] Apache version 2.0 may be used, but it does not support the `virtual` keyword for `Action`. Consequences are noted in the procedure.

- The Vortex scripts must exist in the web server document root, not Vortex's `ScriptRoot`. This means the document root must be writable by the `texis` or CGI user (for `.vtx` compilation), and the script sources are accessible to users (though it may be possible to configure the web server to prevent the latter).
- Every script must be edited to contain a `#!/usr/local/morph3/bin/texis` prefix at the start.
- Every script must be have its execute bits set via `chmod`.

However, this alternate procedure may be used if the preferred redirect-handler method is not possible or practical for some reason. Consult your web server manual (or online at `http://www.apache.org/`) for details and consequences on the directives used in this procedure:

1. Open the Apache configuration file (typically `/etc/httpd/conf/httpd.conf`; see your web server manual) with a text editor.
2. Find the `<Directory ...>` directive that applies to the same directory as the `DocumentRoot` directive. This is typically (but not always) `<Directory /var/www/html>`.
3. Inside this `<Directory ...>` directive, add `ExecCGI` as an option to the `Options` directive.
4. Outside the directive (ie. the line after `</Directive>`), add this line:

   ```
   AddHandler cgi-script .vs
   ```

5. Save the configuration file and exit the editor.
6. Copy all Vortex scripts you wish to run from the Web by this method, from `ScriptRoot` to the same dir under document root. (This step is not needed if for some reason you have edited your Vortex `conf/texis.ini` file and set `ScriptRoot` to `%DOCUMENT_ROOT%`.) For example, the `dowalk.vs` and `search.vs` scripts in `/usr/local/morph3/texis/scripts/webinator` should be copied to the `/var/www/html/webinator` directory (create it if needed), assuming your Apache `DocumentRoot` is `/var/www/html`.
7. Edit every script you copied in the previous step with a text editor, and add this line as the *very* first line in the file:

   ```
   #!/usr/local/morph3/bin/texis
   ```

   (If your Webinator install directory is not `/usr/local/morph3`, change it appropriately.) This step and the next are needed even if `ScriptRoot` is `%DOCUMENT_ROOT%`.
8. Run `chmod a+rx` on every script you copied and edited.
9. Re-start the web server (typically with `/etc/init.d/httpd restart`; check your web server manual).

Note that if you are running Apache for Windows, you may also need to set or edit the `ScriptInterpreterSource` directive; see your web server manual.

6.8 XML Elements in Search Results

Search results can be sent as XML from Webinator to the host server. This section describes the XML elements. The elements are listed below in the approximate order that they are sent.

- `<?xml version="1.0"?>` The version of this XML.
- `<ThunderstoneResults>` Root element that encloses all results.
- `<Query>` Main text search string that was submitted by user.
- `<TitleQuery>` User's title query
- `<UrlQuery>` User's Url query
- `<DepthQuery>` User's maximum depth specified
- `<CategoryQuery>` User's category query
- `<ModifiedDateLessThan>` The `mdlt` query used, if any.
- `<ModifiedDateGreaterThan>` The `mdgt` query used, if any.
- `<UrlRoot>` The URL root of the search script.
- `<Profile>` The profile used.
- `<dropXSL>` If yes removes XSL from results.
- `<AdvancedSearch>` 1 if an advanced search form should be printed.
- `<Proximity>` What the proximity for the search was (line, sentence, paragraph, page)
- `<Suffixes>` What suffix processing occurred in the search
  - 0 - no suffix processing
  - 1 - plurals and possessives
  - 2 - all word forms
- `<Thesaurus>` 1 indicates the thesaurus was used.
- `<Order>` How the search was ordered.
  - r - by rank
  - dd - by date, descending
  - da- by date, ascending
- `<RankOrder>` The ranking weight of word order, from 0-1000
- `<RankProximity>` The ranking weight of query word proximity, from 0-1000
- `<rankDatabaseFrequency>` The ranking weight of rarity of words in the database, from 0-1000

- `<RankDocumentFrequency>` The ranking weight of frequency of words in the document, from 0-1000
- `<RankPosition>` The ranking weight of position in the document, from 0-1000
- `<RankDepth>` The ranking weight of depth of the document, from 0-1000
- `<mode>` set to `admin` if this is an admin search
- `<opts>` Internal use only.
- `<metasearchTarget>` Indicates what backend metasearch targets are available, one element for each target. Currently selected targets will have a selected="selected" attribute.
- `<AdminUrl>` The URL to the admin version of the search interface
- `<MakeLiveUrl>` The URL to make this look and feel live
- `<authUser>` The user that was authenticated via the Proxy Module.
- `<Category>` Information about what categories are available. Occurs multiple times.
  - `<CatVisible>` Set to `Y` if the category should be selectable in the list of categories.
  - `<CatSel>` Set to `Y` if this category is currently selected.
  - `<CatVal>` The numeric ID associated with this category, used for the select box.
  - `<CatName>` The displayed name for this category.
- `<TopBestBets>` Contains information on the BestBets that display above the results.
  - `<BBTitle>` The title for this section of BestBets
  - `<BestBet>` The list of BestBets in this group
    - `<BBResultNum>` The ordered numbering for this Best Bet, starting at 1.
    - `<BBPriority>` The priority for this BestBet, as assigned in the admin interface. The Best bets will already be in the proper priority order.
    - `<BBLink>` The URL that this BestBet links to.
    - `<BBLinkDisplay>` The URL that displays for this BestBet. Long URLs are intelligently truncated for display.
    - `<BBResult>` The link text for this individual BestBet, as assigned in the admin interface.
    - `<BBDescription>` The description for this individual BestBet, as assigned in the admin interface.
    - `<BBGroupname>` The name of the BestBet group this BestBet belongs to.
    - `<BBGroupid>` The id of the BestBet group this BestBet belongs to.
    - `<BBKeywords>` The keywords that trigger this BestBet record to display. This is all keywords for this individual record, not just the one that triggered this activation.
- `<ProfileInfo>` Encloses some profile summary info. Child elements include:
  - `<Profile>` The profile to which the `<ProfileInfo>` element refers to.
  - `<ExitIsEarly>` `Y` if search abort (`<UserResultsNum>` may be short), `N` if not.

  - `<ExitReason>` `ok` if search finished, otherwise token indicating reason (see table of reasons below).
  - `<RedirectUrl>` If present, external URL to redirect the user to. Eg. the external Login URL for Results Authorization, to get third-party login cookies.
  - `<LoginUrl>` If present, local URL to login with `rauser`/`rapass` for Results Authorization.

- `<Summary>` Encloses search results summary; only sent if a query was actually performed. Child elements include:
  - `<Profile>` The profile that the `<Summary>` element applies to.
  - `<Start>` First result item to list.
  - `<End>` Last result item to list.
  - `<TotalNum>` Total number of result items found, *before* ResAuth etc.
  - `<TotalIsEstimate>` `Y` if `<TotalNum>` is an estimate, `N` if not.
  - `<TotalIsShort>` `Y` if `<TotalNum>` is short (eg. early exit), `N` if not
  - `<UserResultsNum>` Total number of result items found, *after* ResAuth etc.
  - `<UserResultsIsEstimate>` `Y` if `<UserResultsNum>` is an estimate, `N` if not.
  - `<UserResultsIsShort>` `Y` if `<UserResultsNum>` is short (eg. early exit), `N` if not.
  - `<ResultsAuthorization>` `Y` if Results Authorization used for query, `N` if not.
  - `<Total>` Readable text for total number of results, *after* ResAuth etc.
  - `<CurOrder>` Text that describes the order by which results are listed.
  - `<OrderLink>` Link that provides an alternative sorting order results list.
  - `<OrderType>` Text that describes `<OrderLink>`.
  - `<NewSkip>` (Metasearch only) The skip value to use for any further request. Only needed with the SOAP API.
  - `<PreviousLink>` Link to the previous page of results (current page minus 1).
  - `<FirstPage>` `1` if this is the first page of results, `0` if later page.
  - `<Pages>` Tag that groups tags for a specific page of results. Child elements include:
    * `<PageLink>` Link to a certain page of the results.
    * `<PageNumber>` Page number a page of results.
  - `<NextLink>` Link to the next page of results (current page plus 1).
  - `<LastPage>` `1` if this is the last page of results, `0` if earlier page.
  - `<Credit>` Text to introduce credit image.
  - `<CreditImage>` The URL of credit image.

- `<Result>` Tag that contains all elements for a given result. Child elements include:
  - `<Profile>` Name of the profile for this `<Result>`. Note that results from meta-search back-ends are re-labeled to the front-end profile.
  - `<Num>` Number of this result item.
  - `<Skip>` Internal use: raw skip(s) for result. Valid for Meta Search back-ends.

- `<Id>` Identifier for this result item.
- `<ResultTitle>` Title of the page of this result item.
- `<Url>` The URL of the page for this result item.
- `<ClickUrl>` The URL for this result item, as should be clicked by the user. The default (if not present) is `<Url>`. Only sent if Query Logging is enabled, in which case it contains redirect for logging the click-through.
- `<UrlPDFHi>` The URL to highlight this PDF item in user's Acrobat Viewer.
- `<UrlDisplay>` Displayed URL for this result item.
- `<RawRank>` The raw relevance rank value for this result item (0-1000).
- `<ScaledRank>` Raw rank scaled up for a more-like-this search (0-1000).
- `<PercentRank>` ScaledRank as a percentage (0-100).
- `<DocSize>` Size (bytes) of the page of this result item.
- `<Depth>` Number of links walked from Base URL to this URL.
- `<UrlSimilar>` The URL to search for pages similar to this result item.
- `<UrlInfo>` The URL for context of answers within a matching document.
- `<UrlParents>` The URL of pages that link to the page of this search result item.
- `<Modified>` Date and time that the page of this result item was last modified.
- `<Visited>` Date and time that the page of this result item was crawled.
- `<Abstract>` Brief text surrounding the matched word or phrase.
- `<Charset>` Character set of the formatted text of the page (typically Storage Charset unless conversion failure).

The following table lists the possible value tokens for the `<ExitReason>` element:

Table 6.7: XML `<ExitReason>` Tokens

Token	Description
ok	Normal exit
ResAuth-ExternalLoginRequired	Need Login Cookies: redirect to `<RedirectUrl>`
ResAuth-CredentialsRequired	Need user/pass: send `rauser`/`rapass` to `<LoginUrl>`
ResAuth-LoginIncorrect	User/pass incorrect; re-send to `<LoginUrl>`
ResAuth-SuccessLimit	Successful Auth Result Limit reached
ResAuth-Timeout	Results Authorization timeout
ResAuth-MaxDocsCheck	Max Docs to Auth-Check exceeded
ResAuth-SmbError	SMB error
ResAuth-NoSmb	SMB unavailable/could not be run

6.9 Third-Party Software

See the Vortex manual for a list of third-party software used by Webinator.

6.10 Version Differences

See the Vortex manual for a list of features and differences between major versions of Webinator.

Chapter 7

Search Interface Help

7.1 Forming a Query

Webinator's search can be as simple or as complex as you need it to be. Usually you will just need to enter a few words that best describe that which you are trying to locate. To perform more complicated searches you might use any combination of logic operators, special pattern matchers, concept expansion, or proximity operations.

Example: `nature conservation organization`

7.1.1 Query Rules of Thumb

- If you get too many junk or nonsense answers, try:
  - Add some more words to your query.
  - Decrease the range of the `Proximity` control.
  - Change the `Word Forms` control to `Exact`.
  - Look at the Match Info and see why they are showing up.
  - Use the Exclusion Operator `(-)` to remove unwanted terms.
  - If you are searching for a phrase, hyphenate the words together.
- If you don't get any answers, or just too few:
  - Remove some more words to your query.
  - Examine your spelling.
  - Increase the scope of the `Proximity` control.
  - It just might not be there?

7.1.2 Overview of Query Abilities

Webinator is based on Texis and as such it shares its text query abilities with all of Thunderstone's products. Throughout our documentation you will see references to Metamorph or Texis. This is because all of our products share a common text query language. This document provides only a brief overview of this language.

If you'd like to know more see the online manual at `http://www.thunderstone.com/site/texisman/link_mmq.html`.

7.1.3 Controlling Proximity

Mastering the usage of proximity gives the ability to locate answers with greater precision. The Webinator input form gives you several options to control the search proximity:

line All query terms must occur on the same line

sentence Query items should all reside within the same sentence

paragraph Within the same paragraph or text block

page All items must occur within same HTML document (the default)

A bar-graph display will be shown any time a ranking search was performed (eg. all searches except `Show Parents`).

7.1.4 Ranking Factors

The ranking algorithm takes into consideration relative word ordering, word proximity, database frequency, document frequency, and position in text. The relative importance of these factors in computing the quality of a hit can be altered under `RANKING FACTORS` on the `Options` page.

7.1.5 Keywords Phrases and Wild-cards

To locate words, just type them in as you would in a word processor. Letter cases will be ignored.

The wild-card character `*` (asterisk) may be used to match just the prefix of a word or to ignore the middle of something.

If the item you wish to locate is more complicated than the simple `*` wild-card can accomplish, try using the regular expression matcher (http://www.thunderstone.com/texis/site/pages/regexp.html).

To locate a number of adjacent words in a specific order, surround them with `"` (double quotation) characters. Putting a `-` (hyphen) between words will also force order and one word proximity.

* see `Word Forms` (7.2, p. 165)

Table 7.1: Query examples

Query	Locates
`john`	john, John
`"john public"`	John Public
`web-browser`	Web browser, web-browser
`John*Public`	John Q. Public, John Public
`456*a*def`	1-456-789-ABCDEF
`activate`	activate, activation, activated, ... *

7.1.6 Applying Search Logic

Texis and Metamorph use set logic for text queries. Set logic is easier to use and provides more abilities than boolean. The examples below make reference to single keywords, but keep in mind that each keyword can represent an entire list of things or any of the special pattern matchers.

Sets (or lists) of things are specified by placing the elements within parenthesis, separated by commas. Example: *(bob,joe,sam,sue)* . In the examples below, you could replace any of the keywords with a list like this.

The default behavior of the search is to locate an intersection (or 'AND') of every element within a query. This means that the query: *"microsoft bob interface"* is the equivalent to the boolean query: *"microsoft AND bob AND interface"* .

- (without) The `-` (minus) is the most commonly used logic symbol. It means the answer should EXCLUDE references to that item.

+ (mandatory) The `+` (plus) symbol in front of a search item means that the answer MUST INCLUDE that item. This is generally used in conjunction with the permutation operation.

@N (permute) The `@` followed by a number indicates how many intersections to locate of the terms in your query. This may be confusing at first, but it is very powerful.

Table 7.2: Search Logic Examples

Query	Finds
`bob sam joe`	Bob with Sam and Joe
`bob sam -joe`	Bob with Sam without Joe
`bob sam joe @1`	Bob with Sam, or Bob with Joe, or Joe with Sam
`A B C D @1`	AB or AC or AD or BC or BD or CD
`+A B C D @1`	ABC or ABD or ACD
`A B C -D @1`	(AB or AC or BC) without D

The plus(`+`) and minus(`-`) operators must be attached to the term to which they apply. There must be a space between the operator and any preceding term.

Correct	Incorrect
`bob +sam -joe`	`bob + sam - joe` `bob+sam-joe`

7.1.7 Natural Language Query

You may enter a query in the form of a sentence or question. The software will automatically identify the important words and phrases within your query and remove the "noise words".

Example: What is the state of the art in text retrieval?

The software will search for: `state of the art` AND `text` AND `retrieval`

7.1.8 Using the Special Pattern Matchers

These pattern matchers are used to locate hard-to-find items within text:

- Regular expression matching for complex patterns
  `http://www.thunderstone.com/texis/site/pages/regexp.html`
- Approximate pattern matching for fuzzy searches
  `http://www.thunderstone.com/texis/site/pages/xpm.html`
- Numeric pattern matching for finding quantities
  `http://www.thunderstone.com/texis/site/pages/npm.html`

If improperly used these pattern matchers can slow queries. Therefore they require other keyword(s) in the query and are disabled entirely under Page proximity. For more details see the Vortex manual on Query Protection (`http://www.thunderstone.com/site/vortexman/link_qprot.html`).

Table 7.3: Pattern Matcher Examples

Query	Matcher	Finds
`ronald %regan`	Approx	Ronald Raygun, Ronald Re-an, Ronald 8eagan
`%75MYPARTNO9045d/6a`	Approx	Anything within 75% of looking like MYPARTNO9045d/6a
`/19[789][0-9]`	RegEXpr	1970-1999
`/[1-9]{3}\-=[0-9]{4}`	RegEXpr	Phone numbers: 555-1212, 820-2200
`#87`	Numeric	four score and seven, 87
`#>0<1`	Numeric	Fractions like 9/16, 55%, 0.123, 15 nanoseconds

Table 7.4: Word Form Examples

Word	president
EXACT	president
PLURAL	(above) + presidents president's
ANY	(above) + presidential presidency preside presides presiding presided
Word	tight
EXACT	tight
PLURAL	(above) + tights
ANY	(above) + tightly tightening tightened tighter tightest
Word	program
EXACT	program
PLURAL	(above) + programs program's
ANY	(above) + programming programmatic programmed programmer programmable

7.1.9 Invoking Thesaurus Expansion

Webinator has a vocabulary of over 250,000 word and phrase associations. Each entry is generally classifiable by either its meaning or part of speech.

Depending on the administrator's Synonyms setting for this profile, synonyms may already be included for each term in your query. If not, synonyms may be included for individual terms within your query by preceding them with a ~ (tilde) character.

7.2 Using Word Forms

The `Word forms` options give you control over how many variations of your query terms will be sought in your search.

Exact: Only exact matches will be allowed. (the default)

Plural & possessives: Plural and possessive forms will be found. (s, es, 's)

Any word forms: As many word forms as can be derived will be located.

We call this morpheme processing, and it is generally smarter than a traditional "stemming" algorithm. It doesn't just rip the end off a word, it actually checks to see if it could be a valid form of the search term. More information is available at
`http://www.thunderstone.com/site/texisman/link_ling.html`.

Notes: Thesaurus terms are also treated in the same manner. Words smaller than 4-5 characters will not be morpheme processed.

7.3 Controlling Proximity

These options give you control over the region in which a match must be found.

line: match terms must be located within the same line.

sentence: all terms within the same sentence.

paragraph: match terms must be located within the same paragraph.

page: (default) all terms within the same document.

In all cases the best possible matches for your query are located and ordered by decreasing quality. A bar graph is produced to indicate the quality of each answer.

7.4 Interpreting Search Results

Note: *The look and feel described here is for the standard search interface. The interface may have been customized by the web site administrator.*

When a query is submitted it will come back with another query form and up to 10 matching documents. If there are more than 10 answers, a link at the top and bottom of the list will allow you to view the next 10 in sequence.

The input form at the top allows you to further tailor your query to home-in on the desired answers, or to submit a completely new query without having to navigate back to the original input form.

Each answer in the result set will have a format similar to the following:

```
1: THE DOCUMENT TITLE (hyperlink to original)    84%*******___
   This is the document abstract. It consists    Size: 11K
   of the text around the first hit within the   Depth: 3
   matching document...                          Find Similar
   http://www.thesite.com/thepage.html           Match Info
                                                 Show Parents
```

The components of each result are:

- Result number
- Document title *(Clicking on this will take you to the original document)*
- Abstract *(The first few hundred characters of the document)*
- Match quality graph. `84%*******___` *(Only shown if relevance ranking was used)*
- Size *(How big is the original document)*
- Depth *(How many clicks from the top of the site)*

- Find Similar *(Find other documents similar to this one)*
- Match Info *(View the matches and other information about the document)*
- Show Parents *(List pages that link to this one)*

7.4.1 Viewing Match Info

The `Match Info` link will show you the context of your answers within the matching document. Matching words will be shown as hyperlinks. Clicking on any match term will take you to the next matching term. A summary at the top of the in-context view shows information about the document, including the time it was last modified.

7.4.2 Finding Similar Documents

The `Find Similar` link will find documents that are similar to the corresponding result. It does this by reading the original document to ascertain its main subject matter, and then conducting a relevance ranked search for those subjects.

Result documents are ordered from best to worst match. The bar graph display will indicate the overall quality of the match.

Note: The document you click on may not be ranked as the best match. This is because other documents may contain more information about the overall subject matter than the original.

7.4.3 Showing Document Parents

Often it is difficult to navigate using a search engine because there is no *back-link* present on the matching document. The `Show Parents` link solves this.

This link will show other documents that contain hyperlinks to the one you click on. In other words, it is an automated back button.