

**Thunderstone Webinator  
WWW Site Indexer Version 5.1.87**

Thunderstone Software

March 11, 2010



# Contents

<b>1</b>	<b>Document Conventions</b>	<b>11</b>
<b>2</b>	<b>Overview</b>	<b>13</b>
2.1	Features . . . . .	13
2.2	Obtaining Webinator . . . . .	14
2.3	Technical Support . . . . .	14
<b>3</b>	<b>Installation</b>	<b>15</b>
3.1	Unix Download and Installation . . . . .	15
3.2	Windows Download and Installation . . . . .	17
3.3	Filesystem Layout . . . . .	19
3.4	File Permissions and OS Specific Notes . . . . .	21
3.5	Customizing Webinator's Appearance . . . . .	21
<b>4</b>	<b>Operation</b>	<b>23</b>
4.1	Running the Administrative Interface . . . . .	23
4.2	First Time Run: Quick Start . . . . .	24
4.3	Administrative Interface Overview . . . . .	26
4.3.1	Entry . . . . .	26
4.3.2	Basic Walk Settings . . . . .	27
4.3.3	All Walk Settings . . . . .	27
4.3.4	Search Settings . . . . .	28
4.3.5	Best Bet Groups . . . . .	28
4.3.6	Profile Tools . . . . .	28

4.3.7	Walk Status . . . . .	29
4.3.8	Query Log . . . . .	30
4.3.9	Test Search . . . . .	31
4.3.10	Live Search . . . . .	31
4.3.11	Profiles . . . . .	31
4.3.12	Accounts . . . . .	32
4.3.13	Documentation . . . . .	34
4.3.14	Webinator Home . . . . .	34
4.3.15	Logout . . . . .	34
4.4	Basic Walk Settings . . . . .	34
4.4.1	Database . . . . .	34
4.4.2	Walk Summary . . . . .	34
4.4.3	Notes . . . . .	34
4.4.4	Base URL . . . . .	35
4.4.5	Enterprise . . . . .	35
4.4.6	Robots . . . . .	36
4.4.7	Extensions . . . . .	36
4.4.8	Exclusions . . . . .	36
4.4.9	Crawl Delay . . . . .	37
4.4.10	Parallelism . . . . .	37
4.4.11	Verbosity . . . . .	37
4.4.12	Rewalk Type . . . . .	38
4.4.13	Rewalk Schedule . . . . .	39
4.4.14	Action Buttons . . . . .	39
4.5	Advanced Walk Settings . . . . .	39
4.5.1	Watch URL . . . . .	40
4.5.2	Notify . . . . .	40
4.5.3	Attach Logs . . . . .	40
4.5.4	Categories . . . . .	40
4.5.5	URL File . . . . .	41

- 4.5.6 URL URL . . . . . 41
- 4.5.7 Single Page . . . . . 41
- 4.5.8 Page File . . . . . 42
- 4.5.9 Page URL . . . . . 42
- 4.5.10 Strip Queries . . . . . 42
- 4.5.11 Ignore Case . . . . . 42
- 4.5.12 Extra Domains . . . . . 43
- 4.5.13 Extra Networks . . . . . 43
- 4.5.14 Extra URLs REX . . . . . 43
- 4.5.15 Exclusion REX . . . . . 44
- 4.5.16 Exclusion Prefix . . . . . 44
- 4.5.17 Exclude by Field . . . . . 44
- 4.5.18 Data from Field . . . . . 45
- 4.5.19 Required REX . . . . . 47
- 4.5.20 Required Prefix . . . . . 47
- 4.5.21 Max Page Size . . . . . 47
- 4.5.22 Max Pages . . . . . 48
- 4.5.23 Max Bytes . . . . . 48
- 4.5.24 Max Depth . . . . . 48
- 4.5.25 Max URL Size . . . . . 48
- 4.5.26 Max Requests . . . . . 48
- 4.5.27 Max Connection Lifetime . . . . . 48
- 4.5.28 Page Timeout . . . . . 49
- 4.5.29 Meta Tags . . . . . 49
- 4.5.30 Standard Meta . . . . . 49
- 4.5.31 All Meta . . . . . 49
- 4.5.32 Storage Charset . . . . . 49
- 4.5.33 Source Default Charset . . . . . 50
- 4.5.34 XML UTF-8 . . . . . 50
- 4.5.35 Keep HTML . . . . . 50

4.5.36	Keep Links . . . . .	50
4.5.37	Remove Common . . . . .	51
4.5.38	Ignore Tags . . . . .	51
4.5.39	Keep Tags . . . . .	51
4.5.40	Ignore Characters . . . . .	51
4.5.41	Plugin Split . . . . .	51
4.5.42	Word Definition . . . . .	52
4.5.43	Index Fields . . . . .	52
4.5.44	Compound Index Fields . . . . .	53
4.5.45	Extra Indexes . . . . .	53
4.5.46	Spell-check Dictionaries . . . . .	53
4.5.47	Primer Type . . . . .	54
4.5.48	Primer URLs . . . . .	54
4.5.49	Login Info . . . . .	56
4.5.50	Proxy . . . . .	56
4.5.51	Proxy Login Info . . . . .	56
4.5.52	Cookie Source Path . . . . .	56
4.5.53	Off-Site Pages . . . . .	57
4.5.54	Stay Under . . . . .	57
4.5.55	Prevent Duplicates . . . . .	57
4.5.56	Duplicate Check Fields . . . . .	57
4.5.57	All Extensions . . . . .	58
4.5.58	Store Refs . . . . .	58
4.5.59	Inline Iframes . . . . .	58
4.5.60	Max Frames . . . . .	58
4.5.61	Execute JavaScript . . . . .	58
4.5.62	Fetch JavaScript . . . . .	58
4.5.63	JavaScript String Links . . . . .	59
4.5.64	Debug JavaScript . . . . .	59
4.5.65	JavaScript Memory . . . . .	59

4.5.66	JavaScript Timeout . . . . .	59
4.5.67	Protocols . . . . .	59
4.5.68	SSL Client Protocols . . . . .	60
4.5.69	Authentication Schemes . . . . .	60
4.5.70	Embedded Security . . . . .	60
4.5.71	Entropy Source . . . . .	60
4.5.72	Max Redirects . . . . .	60
4.5.73	Index Name . . . . .	61
4.5.74	DNS Mode . . . . .	61
4.5.75	Net Mode . . . . .	61
4.5.76	User Agent . . . . .	61
4.5.77	Mime Types . . . . .	61
4.5.78	Respect Expires Header . . . . .	62
4.5.79	Default Refresh Time . . . . .	62
4.5.80	Minimum Refresh Time . . . . .	62
4.5.81	Maximum Refresh Time . . . . .	62
4.5.82	Maximum Process Size . . . . .	62
4.6	Search Settings . . . . .	63
4.6.1	Notes . . . . .	63
4.6.2	Query Logging . . . . .	63
4.6.3	Rotate Schedule . . . . .	63
4.6.4	Email . . . . .	63
4.6.5	Result Order . . . . .	63
4.6.6	Results Style . . . . .	64
4.6.7	Abstract Style . . . . .	64
4.6.8	Abstract Length . . . . .	64
4.6.9	Max Title Length . . . . .	64
4.6.10	Max URL Display Length . . . . .	64
4.6.11	Results per Page . . . . .	65
4.6.12	Max User Results per Page . . . . .	65

4.6.13	Results Width . . . . .	65
4.6.14	Box Color . . . . .	65
4.6.15	Show Advanced Search . . . . .	65
4.6.16	Query Highlighting . . . . .	66
4.6.17	PDF Query Highlighting . . . . .	66
4.6.18	Font . . . . .	66
4.6.19	Display Charset . . . . .	66
4.6.20	Top HTML and Bottom HTML . . . . .	66
4.6.21	Enable Sherlock . . . . .	67
4.6.22	Apply Appearance and Revert Appearance . . . . .	67
4.6.23	Top Best Bet Title . . . . .	67
4.6.24	Right Best Bet Title . . . . .	67
4.6.25	Top Best Bet Group . . . . .	68
4.6.26	Right Best Bet Group . . . . .	68
4.6.27	Top Best Bet Box Color . . . . .	68
4.6.28	Right Best Bet Box Color . . . . .	68
4.6.29	Top Best Bet Border Style . . . . .	68
4.6.30	Right Best Bet Border Style . . . . .	68
4.6.31	Right Best Bet Box Width . . . . .	69
4.6.32	Enable Spell Check . . . . .	69
4.6.33	Suggest Time Limit . . . . .	69
4.6.34	Number of Suggestions . . . . .	69
4.6.35	Synonyms . . . . .	69
4.6.36	Translate Boolean . . . . .	70
4.6.37	Allow the @ Operator . . . . .	70
4.6.38	Allow Linear . . . . .	70
4.6.39	Allow NOT Logic . . . . .	70
4.6.40	Allow Post-Processing . . . . .	71
4.6.41	Allow Wildcards . . . . .	71
4.6.42	Allow Leading Wildcards . . . . .	71

4.6.43	Single-Word Wildcards . . . . .	71
4.6.44	Allow WITHIN Operators . . . . .	71
4.6.45	Resolve Phrase Noise Words . . . . .	71
4.6.46	Keep Noise Words . . . . .	72
4.6.47	Noise List . . . . .	72
4.6.48	Search Timeout . . . . .	72
4.6.49	Show Error Messages . . . . .	72
4.6.50	Debug SQL Level . . . . .	73
4.6.51	Fast Result Counts . . . . .	73
4.6.52	Proximity . . . . .	73
4.6.53	Language Characters . . . . .	74
4.6.54	Word Forms . . . . .	74
4.6.55	Word Ordering . . . . .	74
4.6.56	Word Proximity . . . . .	74
4.6.57	Database Frequency . . . . .	74
4.6.58	Document Frequency . . . . .	75
4.6.59	Position in Text . . . . .	75
4.6.60	Clicks from Home . . . . .	75
4.6.61	Ranked Rows . . . . .	75
4.6.62	Phishing Protection . . . . .	75
4.6.63	Decode Displayed URLs . . . . .	76
4.7	Running the Walker by Hand . . . . .	76
4.7.1	Using dowalk . . . . .	76
4.7.2	Using gw . . . . .	78
4.8	Running the Search Interface . . . . .	78
<b>5</b>	<b>Procedures and Examples</b>	<b>79</b>
5.1	Searching your Index . . . . .	79
5.2	Similarity Searching . . . . .	80
5.3	Page Exclusion, Robots.txt, and Meta-robots . . . . .	81
5.4	Indexing Other Sites . . . . .	83

5.5	Indexing Individual Pages . . . . .	83
5.6	Reindexing on a Schedule . . . . .	83
5.7	Checking for Web Server Errors . . . . .	83
5.8	Removing Pages from the Database . . . . .	84
5.9	Erasing the Entire Database . . . . .	84
5.10	Using Multiple Databases . . . . .	84
5.11	Using Best Bets . . . . .	84
5.11.1	Quick Creation . . . . .	84
5.11.2	Fully Customized . . . . .	85
<b>6</b>	<b>Reference</b>	<b>87</b>
6.1	Database and File Usage . . . . .	87
6.2	Walk Database Tables and Fields . . . . .	88
6.3	Options Table Fields . . . . .	90
6.4	Customizing the Search . . . . .	91
6.5	Customizing the Walker . . . . .	92
6.6	Taxis ISAPI . . . . .	94
6.6.1	Overview . . . . .	94
6.6.2	How it Works . . . . .	94
6.6.3	Settings for Taxis ISAPI . . . . .	95
6.6.4	IIS Manual Configuration . . . . .	96
6.7	Third-Party Software . . . . .	100
6.7.1	Antiword . . . . .	100
6.7.2	Aspell . . . . .	100
6.7.3	Catdoc xls2csv . . . . .	100
6.7.4	Cole library . . . . .	100
6.7.5	iconv . . . . .	101
6.7.6	ppt2html, msg2html . . . . .	101
6.7.7	SSL/HTTPS plugin . . . . .	101
6.7.8	unrar . . . . .	104
6.7.9	unzip . . . . .	105

6.7.10	zlib	106
6.7.11	SpiderMonkey (JavaScript-C) Engine	106
6.7.12	PDF/anytotx plugin	106
6.7.13	thttpd - throttling HTTP server	106
6.7.14	prngd	107
6.7.15	GNU General Public License	107
6.7.16	GNU Lesser General Public License	114
6.7.17	GNU Library General Public License	124
6.7.18	Netscape Public License	133
<b>7</b>	<b>Search Interface Help</b>	<b>143</b>
7.1	Forming a Query	143
7.1.1	Query Rules of Thumb	143
7.1.2	Overview of Query Abilities	144
7.1.3	Controlling Proximity	144
7.1.4	Ranking Factors	144
7.1.5	Keywords Phrases and Wild-cards	144
7.1.6	Applying Search Logic	145
7.1.7	Natural Language Query	146
7.1.8	Using the Special Pattern Matchers	146
7.1.9	Invoking Thesaurus Expansion	147
7.2	Using Word Forms	147
7.3	Controlling Proximity	148
7.4	Interpreting Search Results	148
7.4.1	Viewing Match Info	149
7.4.2	Finding Similar Documents	149
7.4.3	Showing Document Parents	149



# Chapter 1

## Document Conventions

Webinator runs on Windows NT, Windows 2000, 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 Taxis and is written in Taxis's Web Script language named Vortex.

It consists primarily of the Taxis binary program and two Vortex scripts that are run by the Taxis 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/webinator/>. 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 "unsubscribe" 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-5.1.tar.gz` file, from the URL given to you during the registration procedure, to a temporary directory on your machine. (The number `5.1` 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-5.1.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 *inaccessible* 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 `taxis` 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-5.1.tar.gz webinator.tar□~>
```

```
↪install.sum webinator.sum
```

**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, and Windows XP. Download and run the installation program `webinator-5.1.exe` from the URL you were given during the registration procedure. (The number 5.1 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 `/taxis.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:

- **Taxis 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 `Taxis ISAPI` (section 6.6, p. 94) for more information.
- **Install directory**  
This is the directory where Webinator will place its files and subdirectories. The directory you choose should be *inaccessible* 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 known 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 Taxis ISAPI, then you will see the following:

- **IIS Version**

Taxis ISAPI needs to know what version of IIS it is working with. Taxis 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 Taxis ISAPI and Webinator will use to talk to each other. Taxis ISAPI will be configured to attempt to connect to Webinator with this port, and Webinator will have its Taxis 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
  texis.cnf.sample
  .htaccess
  webinator/
    newindex.html
    dowalk
    search
    webinator5man.pdf
    swdrmlog.gif
  bin/
    texis
    monitor
    anytotx
    gw
  texis/
    monitor.log
    scripts/
      errorscript
      webinator/
        dowalk
        search
    vortex.log
    testdb/
    default/
      db1/
      db2/
HTML Directory
  webinator/
    index.html
    webinator5man.pdf
    swdrmlog.gif
CGI Directory
  texis
```

**Note 1:** The files under `webinator` are also installed into the `texis/scripts/webinator` directory, which is the live version. In previous (version 4) releases these were copied to your document root.

`newindex.html` will be named `index.html` if an old copy was not already present.

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 choose to install under your web server's document tree. If you installed 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
  license.key
  taxis.cnf.sample
  taxis.exe
  monitor.exe
  anytotx.exe
  gw.exe
  taxis\
    monitor.log
    scripts/
      errorscript
      webinator/
        dowalk
        search
    vortex.log
    testdb\
    default\
      db1\
      db2\
HTML Directory
  webinator\
    default.htm
    webinator5man.pdf
    swdrmlog.gif
CGI Directory
  taxis.exe

```

The `bin` or `Install` directory contains the `taxis` program and other related utility programs. The `gw` program from version 2.5 Webinator is included in this release. It may go away in future releases. It will work with existing Webinator 2.5 databases. **Note:** `gw` will **not** work with Webinator 4 or later databases.

The `taxis` directory contains the databases and Taxis log files.

## 3.4 File Permissions and OS Specific Notes

- **Windows**

IIS will typically run `taxis.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: `taxis/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.

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

```
monitor -U
```

- **Unix**

It is important that `taxis` 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 `taxis` 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 taxis  
chmod u+s taxis
```

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/taxis/site/pages/webinator5\\_scripts.html](http://www.thunderstone.com/taxis/site/pages/webinator5_scripts.html) for some examples of custom scripts.

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

See also *Customizing the Search* (section 6.4, p. 91) for some insight into the inner workings of the default search script.

# Chapter 4

## Operation

### 4.1 Running the Administrative Interface

*Note to Webinator 2 users: The Webinator 2 gw command is included in the Webinator 4 and later packages, but it does not contain most of the newer features of later versions, and it might be discontinued in future releases. See “Using gw” (4.7.2, p. 78) for details.*

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 Taxis 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.

If your site has pages that you want indexed, and these pages have extensions other than `.html`, `.htm`, or `.txt`, add the extensions to the `Extensions` list. Also note that extensions are case sensitive, unless you use `Ignore case` under `All Walk Settings`.

### 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
  Best Bet Groups
  Profile Tools
    Live Search Database % swd ?
    New Walking Database % swd ?
  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
  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 4.3.11, p. 31).

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** 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 4.4.12). 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 two 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.  
Changes made to the appearance options will be immediately visible in the test search. If Apply Appearance is checked, the changes will also be immediately visible in the live search.
- **Reset**  
This button reverts all settings on the page to what they were when the page was first loaded.

### 4.3.5 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 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:

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

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.

### **Live Search and New Walking Database**

These options are presented on the `List/Edit URLs` page (see 4.3.6) 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.

### **4.3.7 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

fetches, 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.

## 4.3.8 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.

### 4.3.9 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.

### 4.3.10 Live Search

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

### 4.3.11 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.

You can also import options from an existing Webinator 2 database. To do this, fill in the `New Profile Name` and `Data Directory` normally, then also fill in the `Webinator 2 database` field with the full path to the old database from which you would like to import options. If the options were stored in a profile other than the default of `lastrun` using the `gw -save`, fill in the name of the desired profile in the `Webinator 2 profile` box. Then click `Import Settings`. This creates a walk and load settings from the specified old database.

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

- **-[no]unique**: Databases are unique by default (“Prevent Duplicates” 4.5.55).
- **-j**: It is automatically implied (“Stay Under” 4.5.54).
- **-k**: The default expressions are broader (“Word Definition” 4.5.42). 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 “Extensions”4.4.7 list and/or MIME types to the “Mime Types”4.5.77 list 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.

### 4.3.12 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 Taxis 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 `taxis -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.

### 4.3.13 Documentation

This provides a hyperlink to the online version of this document.

<http://www.thunderstone.com/site/webinator5man/>

### 4.3.14 Webinator Home

This provides a hyperlink to the online home of Webinator.

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

### 4.3.15 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 4.5.5`, `URL URL 4.5.6`, `Single page 4.5.7`, `Page file 4.5.8`, and `Page URL 4.5.9` 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 4.5.12` 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` 5.3.

#### **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` 5.3.

### 4.4.7 Extensions

Syntax: one or more file extensions separated by space

A list of the URL extensions that the crawler will accept. The defaults are

```
.html  
.htm  
.txt  
.pdf
```

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 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` 4.5.15 and `Exclusion prefix` 4.5.16 for more ways to exclude URLs.

#### 4.4.9 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`(4.4.10) to 1. A delay defeats the advantage of multiple threads and large delays could cause unexpected page fetch timeouts.

#### 4.4.10 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.

#### 4.4.11 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.

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

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

### 4.4.12 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.

#### 4.4.13 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.

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

#### 4.4.14 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 accommodate 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

Table 4.2: Example Categories

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

provides an example.

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 (\*).

#### 4.5.5 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.

#### 4.5.6 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.

#### 4.5.7 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.

#### 4.5.8 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` 4.5.7.

#### 4.5.9 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` 4.5.7.

#### 4.5.10 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.

#### 4.5.11 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.

### 4.5.12 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 “restrictor” 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.

### 4.5.13 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.

### 4.5.14 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. 43.

### 4.5.15 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
<code>/scratch[0-9]/</code>	a subdirectory named <code>scratch</code> followed by a single digit
<code>[^\a\l]test[^\a\l]</code>	the word <code>test</code> (but not <code>retest</code> or <code>tester</code> etc.)

See also `Exclusions` 4.4.8, `Exclusion prefix` 4.5.16 and `Exclude by Field` 4.5.17.

### 4.5.16 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` 4.4.8, `Exclusion REX` 4.5.15 and `Exclude by Field` 4.5.17.

### 4.5.17 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 4.4.8` and `Exclusion REX 4.5.15`.

### 4.5.18 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 the 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. `Pre%20%2D%20Expense%20Report.doc` is decoded into `Pre - Expense Report.doc`.

**From Meta Field** - If you're using `Meta Field->` as your source, this field is used to specify which meta field's contents you want to use as your source. If you're not using `Meta Field->`, leave this field blank.

Entering text in this field will force the use of `Meta Field->`, regardless of `From Field`'s settings.

**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.

#### 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>.

#### 4.5.19 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.

#### 4.5.20 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.

#### 4.5.21 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.

#### 4.5.22 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.

#### 4.5.23 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.

#### 4.5.24 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.

#### 4.5.25 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.

#### 4.5.26 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.

#### 4.5.27 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).

### 4.5.28 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.

### 4.5.29 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.

### 4.5.30 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.

### 4.5.31 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.

### 4.5.32 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.

### 4.5.33 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.

### 4.5.34 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.

### 4.5.35 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.

### 4.5.36 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.

#### 4.5.37 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.

#### 4.5.38 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.

#### 4.5.39 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.

#### 4.5.40 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.

#### 4.5.41 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.

#### 4.5.42 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. 74) 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.

#### 4.5.43 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.

#### 4.5.44 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.

#### 4.5.45 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.

#### 4.5.46 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.

### 4.5.47 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.

### 4.5.48 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.

#### 4.5.49 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://www.mysite.com/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.

#### 4.5.50 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.

#### 4.5.51 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.

#### 4.5.52 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.

### 4.5.53 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.

### 4.5.54 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.

### 4.5.55 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.

### 4.5.56 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.

### 4.5.57 All Extensions

Syntax: select Yes or No button

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

### 4.5.58 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.

### 4.5.59 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.

### 4.5.60 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.

### 4.5.61 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.

### 4.5.62 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).

### 4.5.63 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.

### 4.5.64 Debug JavaScript

Syntax: select Yes or No button

Print additional debugging messages for JavaScript errors.

### 4.5.65 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.

### 4.5.66 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.

### 4.5.67 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`.

### 4.5.68 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.

### 4.5.69 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 Taxis 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.

### 4.5.70 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.

### 4.5.71 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.

### 4.5.72 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.

### 4.5.73 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/).

### 4.5.74 DNS Mode

Syntax: choose from drop down list

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

### 4.5.75 Net Mode

Syntax: choose from drop down list

This controls what API Webinator uses to access Web pages. “Internal” uses Taxis’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.

### 4.5.76 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).

### 4.5.77 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 (\*/\*).

#### 4.5.78 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".

#### 4.5.79 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.

#### 4.5.80 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.

#### 4.5.81 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.

#### 4.5.82 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.

## 4.6 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 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 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.

#### 4.6.8 Abstract Length

Syntax: enter number in text box

This determines the length in bytes of the document abstract.

#### 4.6.9 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.

#### 4.6.10 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.

#### **4.6.11 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.

#### **4.6.12 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.

#### **4.6.13 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.

#### **4.6.14 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.

#### **4.6.15 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.

### 4.6.16 Query Highlighting

Syntax: select Yes or No button

The user's query will be highlighted (bold tags) in various parts of the results (Title, Abstract, etc.) by default. Setting this to "N" will stop the query from being highlighted in results.

This is also configurable at search time with the `h1` search parameter, which overrides this setting.

### 4.6.17 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.

### 4.6.18 Font

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

This specifies the font to use throughout the search interface.

### 4.6.19 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).

### 4.6.20 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.

#### **4.6.21 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.

#### **4.6.22 Apply Appearance and Revert Appearance**

Syntax: select checkbox

Changes made to the search settings are not normally immediately visible to end users. They may be tested using the "Test Search" menu item. This allows you to see the effects of your changes before committing to them.

Selecting `Apply Appearance` causes the settings currently shown on the form to be made live so that end users will see them. Once this is done, it is permanent, and you must edit the settings to get back the earlier appearance. There is no undo.

Selecting `Revert Appearance` causes the unapplied search settings to be discarded. The settings on the form are reset to those being used on the live search.

#### **4.6.23 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 5.11` for more details.

#### **4.6.24 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 5.11` for more details.

#### 4.6.25 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 5.11` for more details.

#### 4.6.26 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 5.11` for more details.

#### 4.6.27 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 5.11` for more details.

#### 4.6.28 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 5.11` for more details.

#### 4.6.29 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 5.11` for more details.

#### 4.6.30 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 5.11` for more details.

### 4.6.31 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 5.11* for more details.

### 4.6.32 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.

### 4.6.33 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 4.6.32* for more information.

### 4.6.34 Number of Suggestions

Syntax: choose from drop-down list

This controls the number of spelling suggestions offered. See also *Enable Spell Check 4.6.32* for more information.

### 4.6.35 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).

#### 4.6.36 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 Taxis manual on our website.

#### 4.6.37 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.

#### 4.6.38 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 `alllinear` 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.

#### 4.6.39 Allow NOT Logic

Syntax: select Yes or No button

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

#### 4.6.40 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.

#### 4.6.41 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.

#### 4.6.42 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\*".

#### 4.6.43 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".

#### 4.6.44 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.

#### 4.6.45 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.

#### 4.6.46 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.

#### 4.6.47 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

#### 4.6.48 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.

#### 4.6.49 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.

#### 4.6.50 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.

#### 4.6.51 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).

#### 4.6.52 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).

### 4.6.53 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. 52) should usually also be listed in **Language Characters**.

### 4.6.54 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.

### 4.6.55 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).

### 4.6.56 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.

### 4.6.57 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.

#### **4.6.58 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.

#### **4.6.59 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.

#### **4.6.60 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.

#### **4.6.61 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.

#### **4.6.62 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.

### 4.6.63 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".

## 4.7 Running the Walker by Hand

### 4.7.1 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 (*using gw is discussed later in this section*). 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\Taxis\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.  
`taxis 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. `taxis 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.

```
taxis 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.

### 4.7.2 Using `gw`

**Note:** *Using `gw` to drive Webinator's `dowalk` is no longer recommended or supported. `gw` should only be used to manage old version 2 databases. See your Webinator 2 documentation for how to use `gw`.*

## 4.8 Running the Search Interface

See section 5.1, p. 79.

# 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 `taxis` subdirectory of the installation directory. What you specify for DATABASE is only the portion of the path and name under the `taxis` directory. For example, to search the database `/usr/local/morph3/taxis/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/taxis/` portion of the following URLs will be something like `/scripts/taxis.exe/` but may vary depending upon your installation.*

```
http://www.mysite.com/cgi-bin/taxis/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/taxis/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 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.

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`.

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

## 5.4 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.

## 5.5 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.

## 5.6 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.

## 5.7 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 (4.3.6) from the menu.

## 5.8 Removing Pages from the Database

Use the `List/Edit URLs` menu (4.3.6) to find and delete specific URLs from the the database. You may delete individual pages or many pages at once using wildcards.

## 5.9 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.

## 5.10 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.

## 5.11 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.

### 5.11.1 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.

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.

### **5.11.2 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. There is a form on the page that allows you to add keywords to that URL. You can define a priority, title, description, group and keywords for the URL.

If the users query matches the keywords then the Best Bet will be shown. If several Best Bets match the query the highest priority is shown first. The title, description and URL are shown according to the group setting. 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.



# 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 Taxis, also has the notion of a global “default” database. This database resides in the installation directory. On Unix it is called `INSTALLDIR/taxis/testdb`. On Windows it is called `INSTALLDIR\taxis\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. 90) 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
<code>id</code>	Unique record id
<code>Hash</code>	Document hash for duplicate content detection
<code>Size</code>	Size of retrieved raw document (ie. HTML)
<code>Visited</code>	The date the page was modified (or fetched if modified not set)
<code>Dlsecs</code>	The number of seconds needed to fetch the page
<code>Depth</code>	The number of URLs traversed to reach the page
<code>Url</code>	The URL of the real HTML page
<code>Title</code>	The title of the page
<code>Body</code>	The formatted textual content of the page, in Storage Charset (UTF-8)
<code>Keywords</code>	The keywords meta data from the page
<code>Description</code>	The description meta data from the page
<code>Meta</code>	Other meta data from the page, separated by newlines
<code>Catno</code>	List of categories to which the URL belongs
<code>Modified</code>	The date the page was modified
<code>NextCheck</code>	The date the page should next be refreshed
<code>Views</code>	The number of times this URL has been viewed (shown in results)
<code>Clicks</code>	The number of times this URL has been clicked (in results)
<code>CTR</code>	Click-through ratio
<code>Pop</code>	Popularity (number of pages linking to this page)
<code>MimeType</code>	MIME type of original page
<code>Charset</code>	Character set of page as stored (usually Storage Charset)

Table 6.2: Fields in `refs` table

Field	Description
<code>Url</code>	The URL of the HTML page
<code>Ref</code>	The URL of a reference (link) on the HTML page

Table 6.3: Fields in `categories` table

Field	Description
<code>Catno</code>	The number for the category
<code>Url</code>	The URL pattern for the category
<code>Category</code>	The name of the category

Table 6.4: Fields in `error` table

Field	Description
<code>Url</code>	The URL of an HTML page that could not be retrieved
<code>Reason</code>	The reason it could not be retrieved
<code>id</code>	Unique record id (includes timestamp info).

Table 6.5: Fields in `querylog` table (if query logging enabled)

Field	Description
<code>id</code>	Contains the date and time of the query (unique record id)
<code>Client</code>	The hostname of the web client that performed the query
<code>Query</code>	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
<code>id</code>	Unique id for the record
<code>Profile</code>	The name of the profile that the record belongs to
<code>Name</code>	The name of the setting
<code>Type</code>	The data type of the setting (always <code>String</code> )
<code>String</code>	The value of the setting
<code>Int</code>	Unused
<code>Float</code>	Unused
<code>Strlist</code>	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 Taxis 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/taxis/webinator/search/similar.html?~>
  ↪pr=default&ref=http://somesite/somepage.html">~>
  ↪Find pages similar to somepage.html</a>
```

or

```
<a href="/scripts/taxis.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 Taxis 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 `dowalk` script is not compatible with old-style `gw` databases.

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 is 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 `updateindex` 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 Taxis when the database is created. The schema may not be changed. If you want to modify `dowalk` to work with `gw` style databases, you will need to create the database and tables with `gw` before running `dowalk`.

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 Taxis ISAPI

### 6.6.1 Overview

Taxis 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 Taxis 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 Taxis 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 Taxis ISAPI software acts as a pass-through. IIS is configured to give requests to Taxis ISAPI, which in turn passes it along to Webinator. Taxis ISAPI receives Webinator's response, and passes that response back to the web browser.

There are two types of ISAPI programs: filters and extensions. Taxis 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 Taxis ISAPI filter and Taxis 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 `/taxis`. If the request does, Taxis 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 Taxis ISAPI uses an ISAPI Extension mapped as a Wildcard Application Mapping. The installer accesses the default site and creates a new virtual directory, `taxis`. It creates custom Application Settings for that virtual directory, and adds Taxis ISAPI's DLL file as a Wildcard Application Map. This means that any request that comes to that virtual directory (i.e. `/taxis/...`) will not map to a real file location, but will instead be handed off to Taxis ISAPI. The installer also adds Taxis 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 Taxis ISAPI

Taxis 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.

Taxis ISAPI will first attempt to read a port number from a locally installed Webinator's `taxis.cnf`.

#### Reading values from `taxis.cnf`

Taxis 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 `taxis.cnf`:

**Port:** This setting serves double-duty: it tells monitor web server what port it should listen on, and it tells Taxis 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 `taxis.cnf`, 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 `texis.cnf` 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 `texis.cnf` 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 Taxis ISAPI Filter. You should have an entry in the Event Log detailing the results of Taxis ISAPI's attempt to read settings.

### IIS 6 or later

In IIS 6 and later, Taxis 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 Taxis ISAPI.

For IIS 6 to use Taxis ISAPI, there are two separate things that need to be done. A virtual directory needs to be set up to use ProxyModule.dll, and Taxis 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 Taxis ISAPI to (most likely Default Web Site).
- Add a new virtual directory
  - Right click on the website you want to add Taxis ISAPI to, and select New -> Virtual Directory...
  - The Virtual Directory Creation Wizard opens. Click Next>.

- In the `Alias` box, enter `taxis`, 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 Taxis 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\inet_srv` on 32bit windows, `C:\windows\SysWOW64\inet_srv` 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 Taxis ISAPI to IIS' list of allowed extensions on IIS 6:

By default IIS blocks all ISAPI extensions as a security measure. Taxis 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 `Taxis 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 Taxis 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 Third-Party Software

Webinator may contain and utilize the following third-party software to enhance its functionality, depending on the version purchased. Note that your usage and rights to such third-party software may be governed by the appropriate licenses originating with that software, in addition to your License Agreement with Thunderstone - EPI for Thunderstone software.

### 6.7.1 Antiword

The `antiword` package is used by Thunderstone's `anytotx` plugin to handle Microsoft(R) Word files. It has been modified to work within `anytotx`'s installation and to extract meta information. Thunderstone's modified source may be obtained from

`ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the modified Antiword source. Sending a CD will require payment of shipping and handling charges by the requestor. `antiword` is governed by the terms of the GNU GPL, which is reproduced on p. 107.

### 6.7.2 Aspell

The GNU Project's `aspell` package is executed by (but not linked or compiled into) Webinator for spell-checking and "Did you mean..." queries. Complete source code and documentation is available at `ftp://ftp.thunderstone.com/pub/epi-gpl/aspell-0.50.3.tar.gz` or `ftp://ftp.thunderstone.com/pub/epi-gpl/aspell-0.60.4.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the source. Sending of a CD will require payment of shipping and handling charges by the requestor. `aspell` is governed by the terms of the GNU Lesser GPL, which is reproduced on p. 124.

### 6.7.3 Catdoc xls2csv

Catdoc's `xls2csv` program is used by Thunderstone's `anytotx` plugin to handle Microsoft(R) Excel(R) spreadsheet files. It has been modified to work within `anytotx`'s installation and to extract meta information. Thunderstone's modified source may be obtained from

`ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the modified Catdoc source. Sending a CD will require payment of shipping and handling charges by the requestor. Catdoc is governed by the terms of the GNU GPL, which is reproduced on p. 107.

### 6.7.4 Cole library

The `cole` library is used by Thunderstone's versions of `catdoc` and `antiword`. It has been modified to prevent extraneous printing. Thunderstone's modified source may be obtained from

`ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the modified `cole` source. Sending a CD will

require payment of shipping and handling charges by the requestor. The `cole` library is governed by the terms of the GNU GPL, which is reproduced on p. 107.

### 6.7.5 `iconv`

GNU `libiconv` may be used by Thunderstone's HTML processor to convert documents in certain character sets. GNU `libiconv` is not incorporated into Thunderstone's products but is a separate standalone program, called via `exec()` and writing/reading standard input/output. You may obtain complete source code and documentation for `libiconv` at `ftp://ftp.thunderstone.com/pub/epi-gpl/libiconv-1.9.2.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the GNU `libiconv` source. Sending a CD will require payment of shipping and handling charges by the requestor. GNU `libiconv` is governed by the terms of the GNU Library GPL, which is reproduced on p. 124.

### 6.7.6 `ppt2html`, `msg2html`

`ppt2html` and `msg2html` may be used by Thunderstone's `anytotx` document filter to convert Microsoft(R) PowerPoint and `.msg` files. Source is available at:

```
ftp://ftp.thunderstone.com/pub/epi-gpl/ppt2html.c
ftp://ftp.thunderstone.com/pub/epi-gpl/msg2html.c
ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz
```

or by contacting Thunderstone tech support and requesting a CD containing the source. Sending a CD will require payment of shipping and handling charges by the requestor. `ppt2html` and `msg2html` are governed by the terms of the GNU GPL, which is reproduced on p. 107.

### 6.7.7 SSL/HTTPS plugin

This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (<http://www.openssl.org/>). Copyright ©1998-2002 The OpenSSL Project. All rights reserved. This product includes cryptographic software written by Eric Young ([ey@cryptsoft.com](mailto:ey@cryptsoft.com)). Copyright ©1995-1998 Eric Young. All rights reserved.

The OpenSSL toolkit stays under a dual license, i.e. both the conditions of the OpenSSL License and the original SSLeay license apply to the toolkit. See below for the actual license texts. Actually both licenses are BSD-style Open Source licenses. In case of any license issues related to OpenSSL please contact [openssl-core@openssl.org](mailto:openssl-core@openssl.org).

OpenSSL License

-----

Copyright (c) 1998-2002 The OpenSSL Project. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. All advertising materials mentioning features or use of this software must display the following acknowledgment:  
"This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit. (<http://www.openssl.org/>)"
4. The names "OpenSSL Toolkit" and "OpenSSL Project" must not be used to endorse or promote products derived from this software without prior written permission. For written permission, please contact [openssl-core@openssl.org](mailto:openssl-core@openssl.org).
5. Products derived from this software may not be called "OpenSSL" nor may "OpenSSL" appear in their names without prior written permission of the OpenSSL Project.
6. Redistributions of any form whatsoever must retain the following acknowledgment:  
"This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (<http://www.openssl.org/>)"

THIS SOFTWARE IS PROVIDED BY THE OpenSSL PROJECT ``AS IS'' AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE OpenSSL PROJECT OR ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

---

This product includes cryptographic software written by Eric Young ([eay@cryptsoft.com](mailto:eay@cryptsoft.com)). This product includes software written by Tim Hudson ([tjh@cryptsoft.com](mailto:tjh@cryptsoft.com)).

Original SSLeay License

-----  
Copyright (C) 1995-1998 Eric Young (eay@cryptsoft.com)  
All rights reserved.

This package is an SSL implementation written  
by Eric Young (eay@cryptsoft.com).  
The implementation was written so as to conform with Netscapes SSL.

This library is free for commercial and non-commercial use as long as  
the following conditions are aheared to. The following conditions  
apply to all code found in this distribution, be it the RC4, RSA,  
lhash, DES, etc., code; not just the SSL code. The SSL documentation  
included with this distribution is covered by the same copyright terms  
except that the holder is Tim Hudson (tjh@cryptsoft.com).

Copyright remains Eric Young's, and as such any Copyright notices in  
the code are not to be removed. If this package is used in a product,  
Eric Young should be given attribution as the author of the parts of  
the library used. This can be in the form of a textual message at  
program startup or in documentation (online or textual) provided with  
the package.

Redistribution and use in source and binary forms, with or without  
modification, are permitted provided that the following conditions  
are met:

1. Redistributions of source code must retain the copyright  
notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright  
notice, this list of conditions and the following disclaimer in the  
documentation and/or other materials provided with the distribution.
3. All advertising materials mentioning features or use of this software  
must display the following acknowledgement:  
"This product includes cryptographic software written by  
Eric Young (eay@cryptsoft.com)"  
The word 'cryptographic' can be left out if the rouines from the  
library being used are not cryptographic related :-).
4. If you include any Windows specific code (or a derivative thereof)  
from the apps directory (application code) you must include an  
acknowledgement: "This product includes software written by  
Tim Hudson (tjh@cryptsoft.com)"

THIS SOFTWARE IS PROVIDED BY ERIC YOUNG ``AS IS'' AND ANY EXPRESS OR  
IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED  
WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE  
DISCLAIMED. IN NO EVENT SHALL THE AUTHOR OR CONTRIBUTORS BE LIABLE  
FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR  
CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF  
SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR  
BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY,  
WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE  
OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN

IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

The licence and distribution terms for any publically available version or derivative of this code cannot be changed. i.e. this code cannot simply be copied and put under another distribution licence [including the GNU Public Licence.]

### 6.7.8 unrar

The Thunderstone file converter plugin (anytotx) may utilize Alexander L. Roshal's unrar utility to unpack RAR archive files (\*.rar). The unrar utility is governed by the unRAR license reproduced below:

```

*****      *****      *****      unRAR - free utility for RAR archives
**  **  **  **  **  **  ~~~~~
*****      *****      *****      License for use and distribution of
**  **  **  **  **  **  ~~~~~
**  **  **  **  **  **  FREE portable version
                               ~~~~~

```

The source code of unRAR utility is freeware. This means:

1. All copyrights to RAR and the utility unRAR are exclusively owned by the author - Alexander Roshal.
2. The unRAR sources may be used in any software to handle RAR archives without limitations free of charge, but cannot be used to re-create the RAR compression algorithm, which is proprietary. Distribution of modified unRAR sources in separate form or as a part of other software is permitted, provided that it is clearly stated in the documentation and source comments that the code may not be used to develop a RAR (WinRAR) compatible archiver.
3. The unRAR utility may be freely distributed. No person or company may charge a fee for the distribution of unRAR without written permission from the copyright holder.
4. THE RAR ARCHIVER AND THE UNRAR UTILITY ARE DISTRIBUTED "AS IS". NO WARRANTY OF ANY KIND IS EXPRESSED OR IMPLIED. YOU USE AT YOUR OWN RISK. THE AUTHOR WILL NOT BE LIABLE FOR DATA LOSS, DAMAGES, LOSS OF PROFITS OR ANY OTHER KIND OF LOSS WHILE USING OR MISUSING THIS SOFTWARE.
5. Installing and using the unRAR utility signifies acceptance of these terms and conditions of the license.
6. If you don't agree with terms of the license you must remove unRAR files from your storage devices and cease to use the utility.

Thank you for your interest in RAR and unRAR.

Alexander L. Roshal

### 6.7.9 unzip

The Thunderstone file converter plugin (anytotx) may utilize Info-ZIP's unzip utility to unpack ZIP archive files (\*.zip). The unzip software is governed by the Info-ZIP license reproduced below:

This is version 2002-Feb-16 of the Info-ZIP copyright and license. The definitive version of this document should be available at <ftp://ftp.info-zip.org/pub/infozip/license.html> indefinitely.

Copyright (c) 1990-2002 Info-ZIP. All rights reserved.

For the purposes of this copyright and license, "Info-ZIP" is defined as the following set of individuals:

Mark Adler, John Bush, Karl Davis, Harald Denker, Jean-Michel Dubois, Jean-loup Gailly, Hunter Goatley, Ian Gorman, Chris Herborth, Dirk Haase, Greg Hartwig, Robert Heath, Jonathan Hudson, Paul Kienitz, David Kirschbaum, Johnny Lee, Onno van der Linden, Igor Mandrichenko, Steve P. Miller, Sergio Monesi, Keith Owens, George Petrov, Greg Roelofs, Kai Uwe Rommel, Steve Salisbury, Dave Smith, Christian Spieler, Antoine Verheijen, Paul von Behren, Rich Wales, Mike White

This software is provided "as is," without warranty of any kind, express or implied. In no event shall Info-ZIP or its contributors be held liable for any direct, indirect, incidental, special or consequential damages arising out of the use of or inability to use this software.

Permission is granted to anyone to use this software for any purpose, including commercial applications, and to alter it and redistribute it freely, subject to the following restrictions:

1. Redistributions of source code must retain the above copyright notice, definition, disclaimer, and this list of conditions.
2. Redistributions in binary form (compiled executables) must reproduce the above copyright notice, definition, disclaimer, and this list of conditions in documentation and/or other materials provided with the distribution. The sole exception to this condition is redistribution of a standard UnZipSFX binary as part of a self-extracting archive; that is permitted without inclusion of this license, as long as the normal UnZipSFX banner has not been removed from the binary or disabled.
3. Altered versions--including, but not limited to, ports to new operating systems, existing ports with new graphical interfaces, and dynamic, shared, or static library versions--must be plainly

marked as such and must not be misrepresented as being the original source. Such altered versions also must not be misrepresented as being Info-ZIP releases--including, but not limited to, labeling of the altered versions with the names "Info-ZIP" (or any variation thereof, including, but not limited to, different capitalizations), "Pocket UnZip," "WiZ" or "MacZip" without the explicit permission of Info-ZIP. Such altered versions are further prohibited from misrepresentative use of the Zip-Bugs or Info-ZIP e-mail addresses or of the Info-ZIP URL(s).

4. Info-ZIP retains the right to use the names "Info-ZIP," "Zip," "UnZip," "UnZipSFX," "WiZ," "Pocket UnZip," "Pocket Zip," and "MacZip" for its own source and binary releases.

### 6.7.10 **zlib**

Webinator utilizes the `zlib` compression library. Copyright ©1995-2003 Jean-loup Gailly and Mark Adler.

### 6.7.11 **SpiderMonkey (JavaScript-C) Engine**

The `libtxjs.*` library (Thunderstone JavaScript plugin) contains and utilizes the SpiderMonkey engine, as well as additional functionality.

The `txjs.tar` file contains context diffs (patches) to the Mozilla Project's SpiderMonkey (JavaScript-C) engine, version 1.5-rc4. Complete documentation and source code to the SpiderMonkey Engine is available at <http://www.mozilla.org/js/spidermonkey/>.

The patches in `txjs.tar` were created by Thunderstone Software LLC and apply to the core SpiderMonkey engine. They are provided for compliance with the Netscape Public License, which governs usage of the SpiderMonkey engine. A copy of the Netscape Public License is on p. 133. Note that the `libtxjs.*` library also contains other (Thunderstone) code.

### 6.7.12 **PDF/anytotx plugin**

Portions of this product Copyright 1996-2000 Glyph & Cog, LLC.

### 6.7.13 **thttpd - throttling HTTP server**

Webinator's `vhttpd` web server is derived in part from `thttpd`, Copyright ©1995 by Jef Poskanzer [jef@acme.com](mailto:jef@acme.com). All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR AND CONTRIBUTORS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

### 6.7.14 prngd

Webinator may also contain prngd developed by Lutz Jaenicke. Available at [http://ftp.aet.TU-Cottbus.DE/personen/jaenicke/postfix\\_tls/prngd.html](http://ftp.aet.TU-Cottbus.DE/personen/jaenicke/postfix_tls/prngd.html).

### 6.7.15 GNU General Public License

Some third-party software packages shipped with Webinator are governed by the GNU General Public License, reproduced below. See the Third-Party Software section, p. 100, for a list of applicable packages.

GNU GENERAL PUBLIC LICENSE  
Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.  
675 Mass Ave, Cambridge, MA 02139, USA  
Everyone is permitted to copy and distribute verbatim copies  
of this license document, but changing it is not allowed.

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for

this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

GNU GENERAL PUBLIC LICENSE  
TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the

Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.
- b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.
- c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

- a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
- b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
- c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the

original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

#### NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

## Appendix: How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the program's name and a brief idea of what it does.>
Copyright (C) 19yy <name of author>
```

```
This program is free software; you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation; either version 2 of the License, or
(at your option) any later version.
```

```
This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
GNU General Public License for more details.
```

```
You should have received a copy of the GNU General Public License
along with this program; if not, write to the Free Software
Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
```

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

```
Gnomovision version 69, Copyright (C) 19yy name of author
Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type
'show w'.
This is free software, and you are welcome to redistribute it
under certain conditions; type 'show c' for details.
```

The hypothetical commands 'show w' and 'show c' should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than 'show w' and 'show c'; they could even be mouse-clicks or menu items--whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

```
Yoyodyne, Inc., hereby disclaims all copyright interest in the
program 'Gnomovision' (which makes passes at compilers) written by
```

James Hacker.

<signature of Ty Coon>, 1 April 1989  
Ty Coon, President of Vice

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.

### 6.7.16 GNU Lesser General Public License

Some third-party software packages distributed with Webinator are governed by the GNU Lesser General Public License, reproduced below. See the Third-Party Software section, p. 100, for a list of applicable packages.

GNU LESSER GENERAL PUBLIC LICENSE  
Version 2.1, February 1999

Copyright (C) 1991, 1999 Free Software Foundation, Inc.  
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA  
Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

[This is the first released version of the Lesser GPL. It also counts as the successor of the GNU Library Public License, version 2, hence the version number 2.1.]

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public Licenses are intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users.

This license, the Lesser General Public License, applies to some specially designated software packages--typically libraries--of the Free Software Foundation and other authors who decide to use it. You can use it too, but we suggest you first think carefully about whether this license or the ordinary General Public License is the better strategy to use in any particular case, based on the explanations below.

When we speak of free software, we are referring to freedom of use, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish); that you receive source code or can get it if you want it; that you can change the software and use pieces of it in new free programs; and that you are informed that you can do these things.

To protect your rights, we need to make restrictions that forbid distributors to deny you these rights or to ask you to surrender these rights. These restrictions translate to certain responsibilities for you if you distribute copies of the library or if you modify it.

For example, if you distribute copies of the library, whether gratis or for a fee, you must give the recipients all the rights that we gave you. You must make sure that they, too, receive or can get the source code. If you link other code with the library, you must provide complete object files to the recipients, so that they can relink them with the library after making changes to the library and recompiling it. And you must show them these terms so they know their rights.

We protect your rights with a two-step method: (1) we copyright the library, and (2) we offer you this license, which gives you legal permission to copy, distribute and/or modify the library.

To protect each distributor, we want to make it very clear that there is no warranty for the free library. Also, if the library is modified by someone else and passed on, the recipients should know that what they have is not the original version, so that the original author's reputation will not be affected by problems that might be introduced by others.

Finally, software patents pose a constant threat to the existence of any free program. We wish to make sure that a company cannot effectively restrict the users of a free program by obtaining a restrictive license from a patent holder. Therefore, we insist that any patent license obtained for a version of the library must be consistent with the full freedom of use specified in this license.

Most GNU software, including some libraries, is covered by the ordinary GNU General Public License. This license, the GNU Lesser General Public License, applies to certain designated libraries, and is quite different from the ordinary General Public License. We use this license for certain libraries in order to permit linking those libraries into non-free programs.

When a program is linked with a library, whether statically or using a shared library, the combination of the two is legally speaking a combined work, a derivative of the original library. The ordinary General Public License therefore permits such linking only if the entire combination fits its criteria of freedom. The Lesser General Public License permits more lax criteria for linking other code with the library.

We call this license the "Lesser" General Public License because it does Less to protect the user's freedom than the ordinary General Public License. It also provides other free software developers Less of an advantage over competing non-free programs. These disadvantages are the reason we use the ordinary General Public License for many

libraries. However, the Lesser license provides advantages in certain special circumstances.

For example, on rare occasions, there may be a special need to encourage the widest possible use of a certain library, so that it becomes a de-facto standard. To achieve this, non-free programs must be allowed to use the library. A more frequent case is that a free library does the same job as widely used non-free libraries. In this case, there is little to gain by limiting the free library to free software only, so we use the Lesser General Public License.

In other cases, permission to use a particular library in non-free programs enables a greater number of people to use a large body of free software. For example, permission to use the GNU C Library in non-free programs enables many more people to use the whole GNU operating system, as well as its variant, the GNU/Linux operating system.

Although the Lesser General Public License is Less protective of the users' freedom, it does ensure that the user of a program that is linked with the Library has the freedom and the wherewithal to run that program using a modified version of the Library.

The precise terms and conditions for copying, distribution and modification follow. Pay close attention to the difference between a "work based on the library" and a "work that uses the library". The former contains code derived from the library, whereas the latter must be combined with the library in order to run.

#### GNU LESSER GENERAL PUBLIC LICENSE

##### TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License Agreement applies to any software library or other program which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Lesser General Public License (also called "this License"). Each licensee is addressed as "you".

A "library" means a collection of software functions and/or data prepared so as to be conveniently linked with application programs (which use some of those functions and data) to form executables.

The "Library", below, refers to any such software library or work which has been distributed under these terms. A "work based on the Library" means either the Library or any derivative work under copyright law: that is to say, a work containing the Library or a portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Source code" for a work means the preferred form of the work for making modifications to it. For a library, complete source code means

all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the library.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Library is not restricted, and output from such a program is covered only if its contents constitute a work based on the Library (independent of the use of the Library in a tool for writing it). Whether that is true depends on what the Library does and what the program that uses the Library does.

1. You may copy and distribute verbatim copies of the Library's complete source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Library.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Library or any portion of it, thus forming a work based on the Library, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) The modified work must itself be a software library.
- b) You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.
- c) You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.
- d) If a facility in the modified Library refers to a function or a table of data to be supplied by an application program that uses the facility, other than as an argument passed when the facility is invoked, then you must make a good faith effort to ensure that, in the event an application does not supply such function or table, the facility still operates, and performs whatever part of its purpose remains meaningful.

(For example, a function in a library to compute square roots has a purpose that is entirely well-defined independent of the application. Therefore, Subsection 2d requires that any application-supplied function or table used by this function must be optional: if the application does not supply it, the square root function must still compute square roots.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Library, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Library, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Library.

In addition, mere aggregation of another work not based on the Library with the Library (or with a work based on the Library) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may opt to apply the terms of the ordinary GNU General Public License instead of this License to a given copy of the Library. To do this, you must alter all the notices that refer to this License, so that they refer to the ordinary GNU General Public License, version 2, instead of to this License. (If a newer version than version 2 of the ordinary GNU General Public License has appeared, then you can specify that version instead if you wish.) Do not make any other change in these notices.

Once this change is made in a given copy, it is irreversible for that copy, so the ordinary GNU General Public License applies to all subsequent copies and derivative works made from that copy.

This option is useful when you wish to copy part of the code of the Library into a program that is not a library.

4. You may copy and distribute the Library (or a portion or derivative of it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange.

If distribution of object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place satisfies the requirement to distribute the source code, even though third parties are not compelled to copy the source along with the object code.

5. A program that contains no derivative of any portion of the Library, but is designed to work with the Library by being compiled or

linked with it, is called a "work that uses the Library". Such a work, in isolation, is not a derivative work of the Library, and therefore falls outside the scope of this License.

However, linking a "work that uses the Library" with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a "work that uses the library". The executable is therefore covered by this License. Section 6 states terms for distribution of such executables.

When a "work that uses the Library" uses material from a header file that is part of the Library, the object code for the work may be a derivative work of the Library even though the source code is not. Whether this is true is especially significant if the work can be linked without the Library, or if the work is itself a library. The threshold for this to be true is not precisely defined by law.

If such an object file uses only numerical parameters, data structure layouts and accessors, and small macros and small inline functions (ten lines or less in length), then the use of the object file is unrestricted, regardless of whether it is legally a derivative work. (Executables containing this object code plus portions of the Library will still fall under Section 6.)

Otherwise, if the work is a derivative of the Library, you may distribute the object code for the work under the terms of Section 6. Any executables containing that work also fall under Section 6, whether or not they are linked directly with the Library itself.

6. As an exception to the Sections above, you may also combine or link a "work that uses the Library" with the Library to produce a work containing portions of the Library, and distribute that work under terms of your choice, provided that the terms permit modification of the work for the customer's own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the work that the Library is used in it and that the Library and its use are covered by this License. You must supply a copy of this License. If the work during execution displays copyright notices, you must include the copyright notice for the Library among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

- a) Accompany the work with the complete corresponding machine-readable source code for the Library including whatever changes were used in the work (which must be distributed under Sections 1 and 2 above); and, if the work is an executable linked with the Library, with the complete machine-readable "work that uses the Library", as object code and/or source code, so that the user can modify the Library and then relink to produce a modified executable containing the modified Library. (It is understood

that the user who changes the contents of definitions files in the Library will not necessarily be able to recompile the application to use the modified definitions.)

b) Use a suitable shared library mechanism for linking with the Library. A suitable mechanism is one that (1) uses at run time a copy of the library already present on the user's computer system, rather than copying library functions into the executable, and (2) will operate properly with a modified version of the library, if the user installs one, as long as the modified version is interface-compatible with the version that the work was made with.

c) Accompany the work with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 6a, above, for a charge no more than the cost of performing this distribution.

d) If distribution of the work is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.

e) Verify that the user has already received a copy of these materials or that you have already sent this user a copy.

For an executable, the required form of the "work that uses the Library" must include any data and utility programs needed for reproducing the executable from it. However, as a special exception, the materials to be distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of other proprietary libraries that do not normally accompany the operating system. Such a contradiction means you cannot use both them and the Library together in an executable that you distribute.

7. You may place library facilities that are a work based on the Library side-by-side in a single library together with other library facilities not covered by this License, and distribute such a combined library, provided that the separate distribution of the work based on the Library and of the other library facilities is otherwise permitted, and provided that you do these two things:

a) Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities. This must be distributed under the terms of the Sections above.

b) Give prominent notice with the combined library of the fact

that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.

8. You may not copy, modify, sublicense, link with, or distribute the Library except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Library is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

9. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Library or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Library (or any work based on the Library), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Library or works based on it.

10. Each time you redistribute the Library (or any work based on the Library), the recipient automatically receives a license from the original licensor to copy, distribute, link with or modify the Library subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties with this License.

11. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Library at all. For example, if a patent license would not permit royalty-free redistribution of the Library by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Library.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply, and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that

system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

12. If the distribution and/or use of the Library is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Library under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

13. The Free Software Foundation may publish revised and/or new versions of the Lesser General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Library does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

14. If you wish to incorporate parts of the Library into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

#### NO WARRANTY

15. BECAUSE THE LIBRARY IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LIBRARY, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LIBRARY "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE LIBRARY IS WITH YOU. SHOULD THE LIBRARY PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LIBRARY AS PERMITTED ABOVE, BE LIABLE TO YOU

FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LIBRARY (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LIBRARY TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

#### How to Apply These Terms to Your New Libraries

If you develop a new library, and you want it to be of the greatest possible use to the public, we recommend making it free software that everyone can redistribute and change. You can do so by permitting redistribution under these terms (or, alternatively, under the terms of the ordinary General Public License).

To apply these terms, attach the following notices to the library. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the library's name and a brief idea of what it does.>
Copyright (C) <year> <name of author>
```

```
This library is free software; you can redistribute it and/or
modify it under the terms of the GNU Lesser General Public
License as published by the Free Software Foundation; either
version 2.1 of the License, or (at your option) any later version.
```

```
This library is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU
Lesser General Public License for more details.
```

```
You should have received a copy of the GNU Lesser General Public
License along with this library; if not, write to the Free Software
Foundation, Inc.,
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA
```

Also add information on how to contact you by electronic and paper mail.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the library, if necessary. Here is a sample; alter the names:

```
Yoyodyne, Inc., hereby disclaims all copyright interest in the library
`Frob' (a library for tweaking knobs) written by James Random Hacker.
```

```
<signature of Ty Coon>, 1 April 1990
```

Ty Coon, President of Vice

That's all there is to it!

### 6.7.17 GNU Library General Public License

Some third-party software packages distributed with Webinator are governed by the GNU Library General Public License, reproduced below. See the Third-Party Software section, p. 100, for a list of applicable packages.

GNU LIBRARY GENERAL PUBLIC LICENSE  
Version 2, June 1991

Copyright (C) 1991 Free Software Foundation, Inc.  
59 Temple Place - Suite 330, Boston, MA 02111-1307, USA  
Everyone is permitted to copy and distribute verbatim copies  
of this license document, but changing it is not allowed.

[This is the first released version of the library GPL. It is  
numbered 2 because it goes with version 2 of the ordinary GPL.]

#### Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public Licenses are intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users.

This license, the Library General Public License, applies to some specially designated Free Software Foundation software, and to any other libraries whose authors decide to use it. You can use it for your libraries, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the library, or if you modify it.

For example, if you distribute copies of the library, whether gratis or for a fee, you must give the recipients all the rights that we gave you. You must make sure that they, too, receive or can get the source code. If you link a program with the library, you must provide complete object files to the recipients so that they can relink them

with the library, after making changes to the library and recompiling it. And you must show them these terms so they know their rights.

Our method of protecting your rights has two steps: (1) copyright the library, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the library.

Also, for each distributor's protection, we want to make certain that everyone understands that there is no warranty for this free library. If the library is modified by someone else and passed on, we want its recipients to know that what they have is not the original version, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that companies distributing free software will individually obtain patent licenses, thus in effect transforming the program into proprietary software. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

Most GNU software, including some libraries, is covered by the ordinary GNU General Public License, which was designed for utility programs. This license, the GNU Library General Public License, applies to certain designated libraries. This license is quite different from the ordinary one; be sure to read it in full, and don't assume that anything in it is the same as in the ordinary license.

The reason we have a separate public license for some libraries is that they blur the distinction we usually make between modifying or adding to a program and simply using it. Linking a program with a library, without changing the library, is in some sense simply using the library, and is analogous to running a utility program or application program. However, in a textual and legal sense, the linked executable is a combined work, a derivative of the original library, and the ordinary General Public License treats it as such.

Because of this blurred distinction, using the ordinary General Public License for libraries did not effectively promote software sharing, because most developers did not use the libraries. We concluded that weaker conditions might promote sharing better.

However, unrestricted linking of non-free programs would deprive the users of those programs of all benefit from the free status of the libraries themselves. This Library General Public License is intended to permit developers of non-free programs to use free libraries, while preserving your freedom as a user of such programs to change the free libraries that are incorporated in them. (We have not seen how to achieve this as regards changes in header files, but we have achieved it as regards changes in the actual functions of the Library.) The hope is that this will lead to faster development of free libraries.

The precise terms and conditions for copying, distribution and modification follow. Pay close attention to the difference between a "work based on the library" and a "work that uses the library". The former contains code derived from the library, while the latter only works together with the library.

Note that it is possible for a library to be covered by the ordinary General Public License rather than by this special one.

#### GNU LIBRARY GENERAL PUBLIC LICENSE

##### TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License Agreement applies to any software library which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Library General Public License (also called "this License"). Each licensee is addressed as "you".

A "library" means a collection of software functions and/or data prepared so as to be conveniently linked with application programs (which use some of those functions and data) to form executables.

The "Library", below, refers to any such software library or work which has been distributed under these terms. A "work based on the Library" means either the Library or any derivative work under copyright law: that is to say, a work containing the Library or a portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Source code" for a work means the preferred form of the work for making modifications to it. For a library, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the library.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Library is not restricted, and output from such a program is covered only if its contents constitute a work based on the Library (independent of the use of the Library in a tool for writing it). Whether that is true depends on what the Library does and what the program that uses the Library does.

1. You may copy and distribute verbatim copies of the Library's complete source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Library.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Library or any portion of it, thus forming a work based on the Library, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) The modified work must itself be a software library.
- b) You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.
- c) You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.
- d) If a facility in the modified Library refers to a function or a table of data to be supplied by an application program that uses the facility, other than as an argument passed when the facility is invoked, then you must make a good faith effort to ensure that, in the event an application does not supply such function or table, the facility still operates, and performs whatever part of its purpose remains meaningful.

(For example, a function in a library to compute square roots has a purpose that is entirely well-defined independent of the application. Therefore, Subsection 2d requires that any application-supplied function or table used by this function must be optional: if the application does not supply it, the square root function must still compute square roots.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Library, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Library, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Library.

In addition, mere aggregation of another work not based on the Library with the Library (or with a work based on the Library) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may opt to apply the terms of the ordinary GNU General Public License instead of this License to a given copy of the Library. To do this, you must alter all the notices that refer to this License, so that they refer to the ordinary GNU General Public License, version 2, instead of to this License. (If a newer version than version 2 of the ordinary GNU General Public License has appeared, then you can specify that version instead if you wish.) Do not make any other change in these notices.

Once this change is made in a given copy, it is irreversible for that copy, so the ordinary GNU General Public License applies to all subsequent copies and derivative works made from that copy.

This option is useful when you wish to copy part of the code of the Library into a program that is not a library.

4. You may copy and distribute the Library (or a portion or derivative of it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange.

If distribution of object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place satisfies the requirement to distribute the source code, even though third parties are not compelled to copy the source along with the object code.

5. A program that contains no derivative of any portion of the Library, but is designed to work with the Library by being compiled or linked with it, is called a "work that uses the Library". Such a work, in isolation, is not a derivative work of the Library, and therefore falls outside the scope of this License.

However, linking a "work that uses the Library" with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a "work that uses the library". The executable is therefore covered by this License. Section 6 states terms for distribution of such executables.

When a "work that uses the Library" uses material from a header file that is part of the Library, the object code for the work may be a derivative work of the Library even though the source code is not. Whether this is true is especially significant if the work can be linked without the Library, or if the work is itself a library. The threshold for this to be true is not precisely defined by law.

If such an object file uses only numerical parameters, data structure layouts and accessors, and small macros and small inline functions (ten lines or less in length), then the use of the object

file is unrestricted, regardless of whether it is legally a derivative work. (Executables containing this object code plus portions of the Library will still fall under Section 6.)

Otherwise, if the work is a derivative of the Library, you may distribute the object code for the work under the terms of Section 6. Any executables containing that work also fall under Section 6, whether or not they are linked directly with the Library itself.

6. As an exception to the Sections above, you may also compile or link a "work that uses the Library" with the Library to produce a work containing portions of the Library, and distribute that work under terms of your choice, provided that the terms permit modification of the work for the customer's own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the work that the Library is used in it and that the Library and its use are covered by this License. You must supply a copy of this License. If the work during execution displays copyright notices, you must include the copyright notice for the Library among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

- a) Accompany the work with the complete corresponding machine-readable source code for the Library including whatever changes were used in the work (which must be distributed under Sections 1 and 2 above); and, if the work is an executable linked with the Library, with the complete machine-readable "work that uses the Library", as object code and/or source code, so that the user can modify the Library and then relink to produce a modified executable containing the modified Library. (It is understood that the user who changes the contents of definitions files in the Library will not necessarily be able to recompile the application to use the modified definitions.)
- b) Accompany the work with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 6a, above, for a charge no more than the cost of performing this distribution.
- c) If distribution of the work is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.
- d) Verify that the user has already received a copy of these materials or that you have already sent this user a copy.

For an executable, the required form of the "work that uses the Library" must include any data and utility programs needed for reproducing the executable from it. However, as a special exception, the source code distributed need not include anything that is normally

distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of other proprietary libraries that do not normally accompany the operating system. Such a contradiction means you cannot use both them and the Library together in an executable that you distribute.

7. You may place library facilities that are a work based on the Library side-by-side in a single library together with other library facilities not covered by this License, and distribute such a combined library, provided that the separate distribution of the work based on the Library and of the other library facilities is otherwise permitted, and provided that you do these two things:

a) Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities. This must be distributed under the terms of the Sections above.

b) Give prominent notice with the combined library of the fact that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.

8. You may not copy, modify, sublicense, link with, or distribute the Library except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Library is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

9. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Library or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Library (or any work based on the Library), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Library or works based on it.

10. Each time you redistribute the Library (or any work based on the Library), the recipient automatically receives a license from the original licensor to copy, distribute, link with or modify the Library subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

11. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Library at all. For example, if a patent license would not permit royalty-free redistribution of the Library by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Library.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply, and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

12. If the distribution and/or use of the Library is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Library under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

13. The Free Software Foundation may publish revised and/or new versions of the Library General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Library does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

14. If you wish to incorporate parts of the Library into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

#### NO WARRANTY

15. BECAUSE THE LIBRARY IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LIBRARY, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LIBRARY "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE LIBRARY IS WITH YOU. SHOULD THE LIBRARY PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LIBRARY AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LIBRARY (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LIBRARY TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

#### END OF TERMS AND CONDITIONS

#### Appendix: How to Apply These Terms to Your New Libraries

If you develop a new library, and you want it to be of the greatest possible use to the public, we recommend making it free software that everyone can redistribute and change. You can do so by permitting redistribution under these terms (or, alternatively, under the terms of the ordinary General Public License).

To apply these terms, attach the following notices to the library. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the library's name and a brief idea of what it does.>
Copyright (C) <year> <name of author>
```

```
This library is free software; you can redistribute it and/or
```

modify it under the terms of the GNU Library General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This library is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Library General Public License for more details.

You should have received a copy of the GNU Library General Public License along with this library; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA

Also add information on how to contact you by electronic and paper mail.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the library, if necessary. Here is a sample; alter the names:

Yoyodyne, Inc., hereby disclaims all copyright interest in the library 'Frob' (a library for tweaking knobs) written by James Random Hacker.

<signature of Ty Coon>, 1 April 1990  
Ty Coon, President of Vice

That's all there is to it!

### 6.7.18 Netscape Public License

Some third-party software packages distributed with Webinator are governed by the Netscape Public License, reproduced below. See the Third-Party Software section, p. 100, for a list of applicable packages.

Netscape Public License version 1.1

**AMENDMENTS The Netscape Public License Version 1.1 ("NPL") consists of the Mozilla Public License Version 1.1 with the following Amendments, including Exhibit A-Netscape Public License. Files identified with "Exhibit A-Netscape Public License" are governed by the Netscape Public License Version 1.1.**

#### **Additional Terms applicable to the Netscape Public License.**

##### **I. Effect.**

These additional terms described in this Netscape Public License – Amendments shall apply to the Mozilla Communicator client code and to all Covered Code under this License.

**II. "Netscape's Branded Code"** means Covered Code that Netscape distributes and/or permits others to distribute under one or more trademark(s) which are controlled by Netscape but which are not licensed for use under this License.

**III. Netscape and logo.** This License does not grant any rights to use the trademarks "Netscape", the

”Netscape N and horizon” logo or the ”Netscape lighthouse” logo, ”Netcenter”, ”Gecko”, ”Java” or ”JavaScript”, ”Smart Browsing” even if such marks are included in the Original Code or Modifications.

**IV. Inability to Comply Due to Contractual Obligation.** Prior to licensing the Original Code under this License, Netscape has licensed third party code for use in Netscape’s Branded Code. To the extent that Netscape is limited contractually from making such third party code available under this License, Netscape may choose to reintegrate such code into Covered Code without being required to distribute such code in Source Code form, even if such code would otherwise be considered ”Modifications” under this License.

**V. Use of Modifications and Covered Code by Initial Developer.**

**V.1. In General.** The obligations of Section 3 apply to Netscape, except to the extent specified in this Amendment, Section V.2 and V.3.

**V.2. Other Products.** Netscape may include Covered Code in products other than the Netscape’s Branded Code which are released by Netscape during the two (2) years following the release date of the Original Code, without such additional products becoming subject to the terms of this License, and may license such additional products on different terms from those contained in this License.

**V.3. Alternative Licensing.** Netscape may license the Source Code of Netscape’s Branded Code, including Modifications incorporated therein, without such Netscape Branded Code becoming subject to the terms of this License, and may license such Netscape Branded Code on different terms from those contained in this License.

**VI. Litigation.** Notwithstanding the limitations of Section 11 above, the provisions regarding litigation in Section 11(a), (b) and (c) of the License shall apply to all disputes relating to this License.

**EXHIBIT A-Netscape Public License.**

”The contents of this file are subject to the Netscape Public License Version 1.1 (the ”License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.mozilla.org/NPL/> Software distributed under the License is distributed on an ”AS IS” basis, WITHOUT WARRANTY OF ANY KIND, either express or implied. See the License for the specific language governing rights and limitations under the License.

The Original Code is Mozilla Communicator client code, released March 31, 1998.

The Initial Developer of the Original Code is Netscape Communications Corporation. Portions created by Netscape are Copyright (C) 1998-1999 Netscape Communications Corporation. All Rights Reserved.

Contributor(s): \_\_\_\_\_.

Alternatively, the contents of this file may be used under the terms of the — license (the ”[—] License”), in which case the provisions of [—] License are applicable instead of those above. If you wish to allow use of your version of this file only under the terms of the [—] License and not to allow others to use your version of this file under the NPL, indicate your decision by deleting the provisions above and replace them with the notice and other provisions required by the [—] License. If you do not delete the provisions above, a recipient may use your version of this file under either the NPL or the [—] License.”

**MOZILLA PUBLIC LICENSE**

**Version 1.1**

**1. Definitions.**

**1.0.1. "Commercial Use"** means distribution or otherwise making the Covered Code available to a third party.

**1.1. "Contributor"** means each entity that creates or contributes to the creation of Modifications.

**1.2. "Contributor Version"** means the combination of the Original Code, prior Modifications used by a Contributor, and the Modifications

made by that particular Contributor.

**1.3. "Covered Code"** means the Original Code or Modifications or the combination of the Original Code and Modifications, in each case including portions thereof.

**1.4. "Electronic Distribution Mechanism"** means a mechanism generally accepted in the software development community for the electronic transfer of data.

**1.5. "Executable"** means Covered Code in any form other than Source Code.

**1.6. "Initial Developer"** means the individual or entity identified as the Initial Developer in the Source Code notice required by **Exhibit**

**A.**

**1.7. "Larger Work"** means a work which combines Covered Code or portions thereof with code not governed by the terms of this License.

**1.8. "License"** means this document.

**1.8.1. "Licensable"** means having the right to grant, to the maximum extent possible, whether at the time of the initial grant or subsequently acquired, any and all of the rights conveyed herein.

**1.9. "Modifications"** means any addition to or deletion from the substance or structure of either the Original Code or any previous Modifications. When Covered Code is released as a series of files, a Modification is:

**A.** Any addition to or deletion from the contents of a file containing Original Code or previous Modifications.

**B.** Any new file that contains any part of the Original Code or previous Modifications.

**1.10. "Original Code"** means Source Code of computer software code which is described in the Source Code notice required by **Exhibit A** as Original Code, and which, at the time of its release under this License is not already Covered Code governed by this License.

**1.10.1. "Patent Claims"** means any patent claim(s), now owned or hereafter acquired, including without limitation, method, process, and apparatus claims, in any patent Licensable by grantor.

**1.11. "Source Code"** means the preferred form of the Covered Code for making modifications to it, including all modules it contains, plus any associated interface definition files, scripts used to control compilation and installation of an Executable, or source code differential comparisons against either the Original Code or another well known, available Covered Code of the Contributor's choice. The Source Code can be in a compressed or archival form, provided the appropriate decompression or de-archiving software is widely available for no charge.

**1.12. "You" (or "Your")** means an individual or a legal entity exercising rights under, and complying with all of the terms of, this License or a future version of this License issued under Section 6.1. For legal

entities, "You" includes any entity which controls, is controlled by, or is under common control with You. For purposes of this definition, "control" means (a) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (b) ownership of more than fifty percent (50) beneficial ownership of such entity.

## **2. Source Code License.**

### **2.1. The Initial Developer Grant.**

The Initial Developer hereby grants You a world-wide, royalty-free, non-exclusive license, subject to third party intellectual property claims:

(a) under intellectual property rights (other than patent or trademark) Licensable by Initial Developer to use, reproduce, modify, display, perform, sublicense and distribute the Original Code (or portions thereof) with or without Modifications, and/or as part of a Larger Work; and

(b) under Patents Claims infringed by the making, using or selling of Original Code, to make, have made, use, practice, sell, and offer for sale, and/or otherwise dispose of the Original Code (or portions thereof).

(c) the licenses granted in this Section 2.1(a) and (b) are effective on the date Initial Developer first distributes Original Code under the terms of this License.

(d) Notwithstanding Section 2.1(b) above, no patent license is granted: 1) for code that You delete from the Original Code; 2) separate from the Original Code; or 3) for infringements caused by: i) the modification of the Original Code or ii) the combination of the Original Code with other software or devices.

### **2.2. Contributor Grant.**

Subject to third party intellectual property claims, each Contributor hereby grants You a world-wide, royalty-free, non-exclusive license

(a) under intellectual property rights (other than patent or trademark) Licensable by Contributor, to use, reproduce, modify, display, perform, sublicense and distribute the Modifications created by such Contributor (or portions thereof) either on an unmodified basis, with other Modifications, as Covered Code and/or as part of a Larger Work; and

(b) under Patent Claims infringed by the making, using, or selling of Modifications made by that Contributor either alone and/or in combination with its Contributor Version (or portions of such combination), to make, use, sell, offer for sale, have made, and/or otherwise dispose of: 1) Modifications made by that Contributor (or portions thereof); and 2) the combination of Modifications made by that Contributor with its Contributor Version (or portions of such combination).

(c) the licenses granted in Sections 2.2(a) and 2.2(b) are effective on the date Contributor first makes Commercial Use of the Covered Code.

(d) Notwithstanding Section 2.2(b) above, no patent license is granted: 1) for any code that Contributor has deleted from the Contributor Version; 2) separate from the Contributor Version; 3) for infringements caused by: i) third party modifications of Contributor Version or ii) the combination of Modifications made by that Contributor with other software (except as part of the Contributor Version) or other devices; or 4) under Patent Claims infringed by Covered Code in the absence of Modifications made by that Contributor.

## **3. Distribution Obligations.**

**3.1. Application of License.**

The Modifications which You create or to which You contribute are governed by the terms of this License, including without limitation Section 2.2. The Source Code version of Covered Code may be distributed only under the terms of this License or a future version of this License released under Section 6.1, and You must include a copy of this License with every copy of the Source Code You distribute. You may not offer or impose any terms on any Source Code version that alters or restricts the applicable version of this License or the recipients' rights hereunder. However, You may include an additional document offering the additional rights described in Section 3.5.

**3.2. Availability of Source Code.**

Any Modification which You create or to which You contribute must be made available in Source Code form under the terms of this License either on the same media as an Executable version or via an accepted Electronic Distribution Mechanism to anyone to whom you made an Executable version available; and if made available via Electronic Distribution Mechanism, must remain available for at least twelve (12) months after the date it initially became available, or at least six (6) months after a subsequent version of that particular Modification has been made available to such recipients. You are responsible for ensuring that the Source Code version remains available even if the Electronic Distribution Mechanism is maintained by a third party.

**3.3. Description of Modifications.**

You must cause all Covered Code to which You contribute to contain a file documenting the changes You made to create that Covered Code and the date of any change. You must include a prominent statement that the Modification is derived, directly or indirectly, from Original Code provided by the Initial Developer and including the name of the Initial Developer in (a) the Source Code, and (b) in any notice in an Executable version or related documentation in which You describe the origin or ownership of the Covered Code.

**3.4. Intellectual Property Matters****(a) Third Party Claims.**

If Contributor has knowledge that a license under a third party's intellectual property rights is required to exercise the rights granted by such Contributor under Sections 2.1 or 2.2, Contributor must include a text file with the Source Code distribution titled "LEGAL" which describes the claim and the party making the claim in sufficient detail that a recipient will know whom to contact. If Contributor obtains such knowledge after the Modification is made available as described in Section 3.2, Contributor shall promptly modify the LEGAL file in all copies Contributor makes available thereafter and shall take other steps (such as notifying appropriate mailing lists or newsgroups) reasonably calculated to inform those who received the Covered Code that new knowledge has been obtained.

**(b) Contributor APIs.**

If Contributor's Modifications include an application programming interface and Contributor has knowledge of patent licenses which are reasonably necessary to implement that API, Contributor must also include this information in the LEGAL file.

**(c) Representations.**

Contributor represents that, except as disclosed pursuant to Section 3.4(a) above, Contributor believes that Contributor's Modifications are Contributor's original creation(s) and/or Contributor has sufficient rights to

grant the rights conveyed by this License.

### **3.5. Required Notices.**

You must duplicate the notice in **Exhibit A** in each file of the Source Code. If it is not possible to put such notice in a particular Source Code file due to its structure, then You must include such notice in a location (such as a relevant directory) where a user would be likely to look for such a notice. If You created one or more Modification(s) You may add your name as a Contributor to the notice described in **Exhibit A**. You must also duplicate this License in any documentation for the Source Code where You describe recipients' rights or ownership rights relating to Covered Code. You may choose to offer, and to charge a fee for, warranty, support, indemnity or liability obligations to one or more recipients of Covered Code. However, You may do so only on Your own behalf, and not on behalf of the Initial Developer or any Contributor. You must make it absolutely clear than any such warranty, support, indemnity or liability obligation is offered by You alone, and You hereby agree to indemnify the Initial Developer and every Contributor for any liability incurred by the Initial Developer or such Contributor as a result of warranty, support, indemnity or liability terms You offer.

### **3.6. Distribution of Executable Versions.**

You may distribute Covered Code in Executable form only if the requirements of Section **3.1-3.5** have been met for that Covered Code, and if You include a notice stating that the Source Code version of the Covered Code is available under the terms of this License, including a description of how and where You have fulfilled the obligations of Section **3.2**. The notice must be conspicuously included in any notice in an Executable version, related documentation or collateral in which You describe recipients' rights relating to the Covered Code. You may distribute the Executable version of Covered Code or ownership rights under a license of Your choice, which may contain terms different from this License, provided that You are in compliance with the terms of this License and that the license for the Executable version does not attempt to limit or alter the recipient's rights in the Source Code version from the rights set forth in this License. If You distribute the Executable version under a different license You must make it absolutely clear that any terms which differ from this License are offered by You alone, not by the Initial Developer or any Contributor. You hereby agree to indemnify the Initial Developer and every Contributor for any liability incurred by the Initial Developer or such Contributor as a result of any such terms You offer.

### **3.7. Larger Works.**

You may create a Larger Work by combining Covered Code with other code not governed by the terms of this License and distribute the Larger Work as a single product. In such a case, You must make sure the requirements of this License are fulfilled for the Covered Code.

## **4. Inability to Comply Due to Statute or Regulation.**

If it is impossible for You to comply with any of the terms of this License with respect to some or all of the Covered Code due to statute, judicial order, or regulation then You must: (a) comply with the terms of this License to the maximum extent possible; and (b) describe the limitations and the code they affect. Such description must be included in the LEGAL file described in Section **3.4** and must be included with all distributions of the Source Code. Except to the extent prohibited by statute or regulation, such description must be sufficiently detailed for a recipient of ordinary skill to be able to understand it.

## **5. Application of this License.**

This License applies to code to which the Initial Developer has attached the notice in **Exhibit A** and to

related Covered Code.

## **6. Versions of the License.**

### **6.1. New Versions.**

Netscape Communications Corporation ("Netscape") may publish revised and/or new versions of the License from time to time. Each version will be given a distinguishing version number.

### **6.2. Effect of New Versions.**

Once Covered Code has been published under a particular version of the License, You may always continue to use it under the terms of that version. You may also choose to use such Covered Code under the terms of any subsequent version of the License published by Netscape. No one other than Netscape has the right to modify the terms applicable to Covered Code created under this License.

### **6.3. Derivative Works.**

If You create or use a modified version of this License (which you may only do in order to apply it to code which is not already Covered Code governed by this License), You must (a) rename Your license so that the phrases "Mozilla", "MOZILLAPL", "MOZPL", "Netscape", "MPL", "NPL" or any confusingly similar phrase do not appear in your license (except to note that your license differs from this License) and (b) otherwise make it clear that Your version of the license contains terms which differ from the Mozilla Public License and Netscape Public License. (Filling in the name of the Initial Developer, Original Code or Contributor in the notice described in **Exhibit A** shall not of themselves be deemed to be modifications of this License.)

## **7. DISCLAIMER OF WARRANTY.**

COVERED CODE IS PROVIDED UNDER THIS LICENSE ON AN "AS IS" BASIS, WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, WARRANTIES THAT THE COVERED CODE IS FREE OF DEFECTS, MERCHANTABILITY, FIT FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE COVERED CODE IS WITH YOU. SHOULD ANY COVERED CODE PROVE DEFECTIVE IN ANY RESPECT, YOU (NOT THE INITIAL DEVELOPER OR ANY OTHER CONTRIBUTOR) ASSUME THE COST OF ANY NECESSARY SERVICING, REPAIR OR CORRECTION. THIS DISCLAIMER OF WARRANTY CONSTITUTES AN ESSENTIAL PART OF THIS LICENSE. NO USE OF ANY COVERED CODE IS AUTHORIZED HEREUNDER EXCEPT UNDER THIS DISCLAIMER.

## **8. TERMINATION.**

**8.1.** This License and the rights granted hereunder will terminate automatically if You fail to comply with terms herein and fail to cure such breach within 30 days of becoming aware of the breach. All sublicenses to the Covered Code which are properly granted shall survive any termination of this License. Provisions which, by their nature, must remain in effect beyond the termination of this License shall survive.

**8.2.** If You initiate litigation by asserting a patent infringement claim (excluding declaratory judgment actions) against Initial Developer or a Contributor (the Initial Developer or Contributor against whom You file such action is referred to as "Participant") alleging that:

(a) such Participant's Contributor Version directly or indirectly infringes any patent, then any and all rights

granted by such Participant to You under Sections 2.1 and/or 2.2 of this License shall, upon 60 days notice from Participant terminate prospectively, unless if within 60 days after receipt of notice You either: (i) agree in writing to pay Participant a mutually agreeable reasonable royalty for Your past and future use of Modifications made by such Participant, or (ii) withdraw Your litigation claim with respect to the Contributor Version against such Participant. If within 60 days of notice, a reasonable royalty and payment arrangement are not mutually agreed upon in writing by the parties or the litigation claim is not withdrawn, the rights granted by Participant to You under Sections 2.1 and/or 2.2 automatically terminate at the expiration of the 60 day notice period specified above.

(b) any software, hardware, or device, other than such Participant's Contributor Version, directly or indirectly infringes any patent, then any rights granted to You by such Participant under Sections 2.1(b) and 2.2(b) are revoked effective as of the date You first made, used, sold, distributed, or had made, Modifications made by that Participant.

**8.3.** If You assert a patent infringement claim against Participant alleging that such Participant's Contributor Version directly or indirectly infringes any patent where such claim is resolved (such as by license or settlement) prior to the initiation of patent infringement litigation, then the reasonable value of the licenses granted by such Participant under Sections 2.1 or 2.2 shall be taken into account in determining the amount or value of any payment or license.

**8.4.** In the event of termination under Sections 8.1 or 8.2 above, all end user license agreements (excluding distributors and resellers) which have been validly granted by You or any distributor hereunder prior to termination shall survive termination.

## **9. LIMITATION OF LIABILITY.**

UNDER NO CIRCUMSTANCES AND UNDER NO LEGAL THEORY, WHETHER TORT (INCLUDING NEGLIGENCE), CONTRACT, OR OTHERWISE, SHALL YOU, THE INITIAL DEVELOPER, ANY OTHER CONTRIBUTOR, OR ANY DISTRIBUTOR OF COVERED CODE, OR ANY SUPPLIER OF ANY OF SUCH PARTIES, BE LIABLE TO ANY PERSON FOR ANY INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF GOODWILL, WORK STOPPAGE, COMPUTER FAILURE OR MALFUNCTION, OR ANY AND ALL OTHER COMMERCIAL DAMAGES OR LOSSES, EVEN IF SUCH PARTY SHALL HAVE BEEN INFORMED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION OF LIABILITY SHALL NOT APPLY TO LIABILITY FOR DEATH OR PERSONAL INJURY RESULTING FROM SUCH PARTY'S NEGLIGENCE TO THE EXTENT APPLICABLE LAW PROHIBITS SUCH LIMITATION. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OR LIMITATION OF INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THIS EXCLUSION AND LIMITATION MAY NOT APPLY TO YOU.

## **10. U.S. GOVERNMENT END USERS.**

The Covered Code is a "commercial item," as that term is defined in 48 C.F.R. 2.101 (Oct. 1995), consisting of "commercial computer software" and "commercial computer software documentation," as such terms are used in 48 C.F.R. 12.212 (Sept. 1995). Consistent with 48 C.F.R. 12.212 and 48 C.F.R. 227.7202-1 through 227.7202-4 (June 1995), all U.S. Government End Users acquire Covered Code with only those rights set forth herein.

## **11. MISCELLANEOUS.**

This License represents the complete agreement concerning subject matter hereof. If any provision of this License is held to be unenforceable, such provision shall be reformed only to the extent necessary to make it enforceable. This License shall be governed by California law provisions (except to the extent applicable law, if any, provides otherwise), excluding its conflict-of-law provisions. With respect to disputes in which at least one party is a citizen of, or an entity chartered or registered to do business in the United States of America, any litigation relating to this License shall be subject to the jurisdiction of the Federal Courts of the Northern District of California, with venue lying in Santa Clara County, California, with the losing party responsible for costs, including without limitation, court costs and reasonable attorneys' fees and expenses. The application of the United Nations Convention on Contracts for the International Sale of Goods is expressly excluded. Any law or regulation which provides that the language of a contract shall be construed against the drafter shall not apply to this License.

## **12. RESPONSIBILITY FOR CLAIMS.**

As between Initial Developer and the Contributors, each party is responsible for claims and damages arising, directly or indirectly, out of its utilization of rights under this License and You agree to work with Initial Developer and Contributors to distribute such responsibility on an equitable basis. Nothing herein is intended or shall be deemed to constitute any admission of liability.

## **13. MULTIPLE-LICENSED CODE.**

Initial Developer may designate portions of the Covered Code as "Multiple-Licensed". "Multiple-Licensed" means that the Initial Developer permits you to utilize portions of the Covered Code under Your choice of the NPL or the alternative licenses, if any, specified by the Initial Developer in the file described in Exhibit A.

### **EXHIBIT A -Mozilla Public License.**

"The contents of this file are subject to the Mozilla Public License Version 1.1 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.mozilla.org/MPL/> Software distributed under the License is distributed on an "AS IS" basis, WITHOUT WARRANTY OF ANY KIND, either express or implied. See the License for the specific language governing rights and limitations under the License. The Original Code is \_\_\_\_\_ . The Initial Developer of the Original Code is \_\_\_\_\_. Portions created by \_\_\_\_\_ are Copyright (C) \_\_\_\_\_. All Rights Reserved. Contributor(s): \_\_\_\_\_. Alternatively, the contents of this file may be used under the terms of the \_\_\_\_\_ license (the "[ ] License"), in which case the provisions of [ ] License are applicable instead of those above. If you wish to allow use of your version of this file only under the terms of the [ ] License and not to allow others to use your version of this file under the MPL, indicate your decision by deleting the provisions above and replace them with the notice and other provisions required by the [ ] License. If you do not delete the provisions above, a recipient may use your version of this file under either the MPL or the [ ] License."

[NOTE: The text of this Exhibit A may differ slightly from the text of the notices in the Source Code files of the Original Code. You should use the text of this Exhibit A rather than the text found in the Original Code Source Code for Your Modifications.]



# 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 Taxis 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 Taxis. 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/taxisman/link\\_mmq.html](http://www.thunderstone.com/site/taxisman/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/taxis/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. 147)

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

Taxis 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](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 Reagan
%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](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.