Thunderstone Search Appliance
WWW Site Indexer Version 9.1.9

Thunderstone Software

August 7, 2014

Contents

1 Overview **19**

1.1 Features 19

1.2 Technical Support 20

2 Installation **21**

2.1 How to unpack and install the Search Appliance 21

2.1.1 Console Menu 21

2.1.2 Front Panel LCD 24

2.2 Customizing the Search Appliance's Appearance 27

3 Operation **29**

3.1 Running the Administrative Interface 29

3.2 First Time Run: Quick Start 29

3.3 Administrative Interface Overview 31

3.4 Basic Walk Settings 50

3.4.1 Walk Summary 50

3.4.2 Notes 50

3.4.3 Base URL 51

3.4.4 Enterprise 51

3.4.5 Robots 52

3.4.6 Allow Extensions 52

3.4.7 All Extensions 52

3.4.8 Exclude Extensions 53

3.4.9 Exclusions 53

3.4.10 Walk Delay 53
3.4.11 Parallelism 53
3.4.12 Verbosity 53
3.4.13 Disable Walks 54
3.4.14 Rewalk Type 54
3.4.15 Rewalk Schedule 55
3.4.16 Action Buttons 56
3.5 Advanced Walk Settings 56
3.5.1 Watch URL 56
3.5.2 Notify 57
3.5.3 Attach Logs 57
3.5.4 Categories 57
3.5.5 Categories Type 58
3.5.6 DBWalker 58
3.5.7 URL File 58
3.5.8 URL URL 59
3.5.9 Single Page 59
3.5.10 Page File 59
3.5.11 Page URL 59
3.5.12 Strip Queries 60
3.5.13 Keep Query Vars 60
3.5.14 Ignore Query Vars 60
3.5.15 Ignore Case 60
3.5.16 Extra Domains 61
3.5.17 Extra Networks 61
3.5.18 Extra URLs REX 61
3.5.19 Exclusion REX 62
3.5.20 Exclusion Prefix 62
3.5.21 RSS Feeds 63
3.5.22 Exclude by Field 63

3.5.23 Additional Fields 63
3.5.24 Data from Field 64
3.5.25 Required REX 68
3.5.26 Required Prefix 68
3.5.27 Max Page Size 68
3.5.28 Max Pages 68
3.5.29 Max Bytes 68
3.5.30 Max Depth 68
3.5.31 Max URL Size 69
3.5.32 Max Requests 69
3.5.33 Max Connection Lifetime 69
3.5.34 Page Timeout 69
3.5.35 Meta Tags 69
3.5.36 Standard Meta 69
3.5.37 All Meta 70
3.5.38 Storage Charset 70
3.5.39 Source Default Charset 70
3.5.40 XML UTF-8 70
3.5.41 Keep HTML 71
3.5.42 Keep Links 71
3.5.43 Remove Common 71
3.5.44 Ignore Tags 71
3.5.45 Keep Tags 72
3.5.46 Ignore Characters 72
3.5.47 Plugin Split 72
3.5.48 Language Analysis 73
3.5.49 CJK Mode 73
3.5.50 Unknown File Formats 73
3.5.51 PDF Title Action 74
3.5.52 Word Definition 74

3.5.53 Text Search Mode 75
3.5.54 Attribute Compare Mode 76
3.5.55 Index Fields 76
3.5.56 Compound Index Fields 76
3.5.57 Extra Indexes 77
3.5.58 Spell-check Dictionaries 77
3.5.59 Primer Type 77
3.5.60 Primer URLs 78
3.5.61 Login Info 80
3.5.62 Proxy 80
3.5.63 Proxy Login Info 80
3.5.64 Cookie Source Path 80
3.5.65 Cookie Jar 81
3.5.66 Strict Cookie Paths 81
3.5.67 Off-Site Pages 81
3.5.68 Off-Site Components 81
3.5.69 Stay Under 81
3.5.70 Prevent Duplicates 82
3.5.71 Duplicate Check Fields 82
3.5.72 Store Refs 82
3.5.73 Inline Iframes 82
3.5.74 Max Frames 83
3.5.75 Execute JavaScript 83
3.5.76 Fetch JavaScript 83
3.5.77 JavaScript String Links 83
3.5.78 Debug JavaScript 83
3.5.79 JavaScript Memory 83
3.5.80 JavaScript Timeout 84
3.5.81 Protocols 84
3.5.82 HTTP Version 84

3.5.83 SSL Client Protocols 84
3.5.84 Authentication Schemes 84
3.5.85 Embedded Security 85
3.5.86 Entropy Source 85
3.5.87 Multiple Fetches 85
3.5.88 Follow Cross-Site Links 85
3.5.89 Max Redirects 86
3.5.90 Empty Form Redirects 86
3.5.91 Execute Walked Dataload 86
3.5.92 Index Name 86
3.5.93 DNS Mode 86
3.5.94 User Agent 87
3.5.95 Mime Types 87
3.5.96 Custom Headers 87
3.5.97 Respect Expires Header 87
3.5.98 Cache Content 87
3.5.99 Default Refresh Time 88
3.5.100 Minimum Refresh Time 88
3.5.101 Maximum Refresh Time 89
3.5.102 Maximum Process Size 89
3.5.103 Maximum Load Average 89
3.5.104 Replication Settings 89
3.5.105 Send Data 89
3.5.106 Send Settings 90
3.5.107 Batch Rows 90
3.5.108 Batch Size 90
3.5.109 Batch Idle 90
3.5.110 Log Replication 90
3.6 Search Settings 90
3.6.1 Notes 91

3.6.2 Query Logging 91
3.6.3 Rotate Schedule 91
3.6.4 Email 91
3.6.5 Result Order 91
3.6.6 Results Style 92
3.6.7 Allow RSS 92
3.6.8 Format XSL Output 92
3.6.9 XSL Engine 92
3.6.10 XSL File 93
3.6.11 Abstract Style 93
3.6.12 Abstract Length 93
3.6.13 Max Title Length 93
3.6.14 Max URL Display Length 93
3.6.15 Results per Page 94
3.6.16 Max User Results per Page 94
3.6.17 Page Links Shown 94
3.6.18 Results per Site 94
3.6.19 Allow site: syntax 95
3.6.20 Results Width 95
3.6.21 Box Color 95
3.6.22 Show File Icons 95
3.6.23 Show Thunderstone logo on results 95
3.6.24 Show Advanced Search 96
3.6.25 Query Autocomplete 96
3.6.26 Max Completions 96
3.6.27 Results Highlighting 96
3.6.28 Context Highlighting 97
3.6.29 PDF Query Highlighting 97
3.6.30 PDF Highlighting Format 97
3.6.31 Font 97

3.6.32 Display Charset 97
3.6.33 Top HTML and Bottom HTML 98
3.6.34 Enable Sherlock 98
3.6.35 Top Best Bet Title 99
3.6.36 Right Best Bet Title 99
3.6.37 Top Best Bet Group 99
3.6.38 Right Best Bet Group 99
3.6.39 Top Best Bet Box Color 99
3.6.40 Right Best Bet Box Color 99
3.6.41 Top Best Bet Border Style 100
3.6.42 Right Best Bet Border Style 100
3.6.43 Right Best Bet Box Width 100
3.6.44 Authorization Method 100
3.6.45 Login Cookies 100
3.6.46 Login URL 101
3.6.47 Basic/NTLM/file Cookie Type 101
3.6.48 Login Verification URL 102
3.6.49 Unauthorized Result Query 102
3.6.50 Username Fixup 102
3.6.51 Max Docs to Auth-Check 103
3.6.52 Successful Auth Result Limit 103
3.6.53 Total Auth Timeout 104
3.6.54 Allow Authorization URL 104
3.6.55 Authorization Caching 104
3.6.56 Debug Results Authorization 105
3.6.57 Show Authorization Info 105
3.6.58 Enable Spell Check 105
3.6.59 Suggest Time Limit 105
3.6.60 Number of Suggestions 106
3.6.61 Synonyms 106

3.6.62 Main Thesaurus 106
3.6.63 Secondary Thesaurus 106
3.6.64 Translate Boolean 107
3.6.65 Allow the @ Operator 107
3.6.66 Allow Linear 107
3.6.67 Allow NOT Logic 107
3.6.68 Allow Post-Processing 107
3.6.69 Allow Wildcards 108
3.6.70 Allow Leading Wildcards 108
3.6.71 Single-Word Wildcards 108
3.6.72 Allow WITHIN Operators 108
3.6.73 Require All Words 108
3.6.74 Resolve Phrase Noise Words 109
3.6.75 Phrase Word Processing 109
3.6.76 Keep Noise Words 109
3.6.77 Noise List 109
3.6.78 Search Timeout 110
3.6.79 Show Error Messages 110
3.6.80 Debug SQL Level 110
3.6.81 Debug Metamorph Level 111
3.6.82 Fast Result Counts 111
3.6.83 Proximity 111
3.6.84 Language Characters 111
3.6.85 Word Forms 112
3.6.86 Custom Suffix List 112
3.6.87 Custom Suffix Default Removal 112
3.6.88 Custom Suffix Min Length 112
3.6.89 Word Ordering 113
3.6.90 Word Proximity 113
3.6.91 Database Frequency 113

3.6.92 Document Frequency 113
3.6.93 Position in Text 113
3.6.94 Clicks from Home 113
3.6.95 Ranked Rows 114
3.6.96 XML Export Variables 114
3.6.97 File URL Format 114
3.6.98 Redirect Format 115
3.6.99 Phishing Protection 115
3.6.100 Decode Displayed URLs 115
3.6.101 Results Caching 115
3.6.102 Max Cache Entry Age 116
3.6.103 Max Cache Size 116
3.6.104 Min Search Time 116
3.6.105 Visible 117
3.7 System Wide Settings 117
3.7.1 Admin Theme 117
3.7.2 Home Page 117
3.7.3 Enter At Search, Default Profile 117
3.7.4 Favicon.ico 117
3.7.5 Robots.txt 118
3.7.6 Cluster Members 118
3.7.7 API Logging 118
3.7.8 Task Monitor Logging 119
3.7.9 Google Connector Logging 119
3.7.10 Disable All Walks 119
3.7.11 Update Software 119
3.7.12 HTTP Proxy Server 119
3.7.13 Proxy Username 120
3.7.14 Proxy Password 120
3.7.15 System Replication Settings 120

3.7.16 Allow Receiving 120
3.7.17 Log All Replication 120
3.7.18 Enable HTTPS Server 120
3.7.19 Require HTTPS for Direct Admin 121
3.7.20 Require HTTPS for Proxy Admin 121
3.7.21 Admin Access IPs 122
3.7.22 HTTPS/SSL Protocols 122
3.7.23 HTTPS/SSL Ciphers 122
3.7.24 Enable SNMP service 122
3.7.25 SNMP Community Name 123
3.7.26 SNMP Location Value 123
3.7.27 SNMP Contact Value 123
3.7.28 SNMP Access IPs 123
3.7.29 Administration Interface Options 123
3.7.30 <title> order 124
3.7.31 <title> max profile length 124
3.8 Results Authorization 124
3.8.1 Results Authorization Walk Settings 125
3.8.2 Results Authorization Search Settings 125
3.9 Meta Search - Search multiple profiles as one 125
3.9.1 Profile Creation 125
3.9.2 Meta Search Walk Settings 125
3.9.3 Search Settings 126
3.10 Access Control 127
3.10.1 User Groups 127
3.10.2 Object hierarchy 127
3.10.3 Access Control Lists 128
3.10.4 Determining Effective Rights 128
3.10.5 Required Rights for Admin Actions 129
3.11 Running the Search Interface 131

4 Procedures and Examples 133

4.1 Searching your Index . . . 133

4.2 Similarity Searching . . . 134

4.3 Using the Thesaurus Feature . . . 135

4.4 Getting Software Updates . . . 136

4.5 Page Exclusion, Robots.txt, and Meta-robots . . . 136

4.6 Indexing Other Sites . . . 138

4.7 Indexing Individual Pages . . . 138

4.8 Reindexing on a Schedule . . . 139

4.9 Checking for Web Server Errors . . . 139

4.10 Removing Pages from the Database . . . 139

4.11 Erasing the Entire Database . . . 139

4.12 Using Multiple Databases . . . 139

4.13 Integrating Search with your Site . . . 139

4.13.1 Link to the Appliance . . . 140

4.13.2 Embed a search box . . . 140

4.13.3 Request XML search results . . . 141

4.13.4 Invoke the search SOAP API . . . 149

4.14 Search Result RSS Feeds . . . 149

4.15 OpenSearch Support . . . 149

4.16 Using Best Bets . . . 150

4.16.1 Quick Creation . . . 150

4.16.2 Fully Customized . . . 150

4.17 Using Access Control . . . 151

4.17.1 Initial Lockdown . . . 151

4.17.2 Example: User with Complete Control on One Profile . . . 152

4.17.3 Example: User with Look and Feel Control on All Profiles . . . 152

4.18 Indexing File Servers . . . 152

4.19 Replication . . . 153

4.19.1 Replication Overview . . . 153

4.19.2 Procedure 153
4.19.3 DataLoad API 154
4.20 Additional Fields 161
4.20.1 Overview 161
4.20.2 Populating 161
4.20.3 Sorting 162
4.20.4 Searching 162
4.21 DBWalker 162
4.21.1 Overview 162
4.21.2 Configuration Overview 162
4.21.3 DBWalker Output Overview 163
4.21.4 DBWalker Authentication Overview 163
4.21.5 Obtaining DBWalker 164
4.21.6 Managing DBWalker 164
4.21.7 DBWalker Global Options 164
4.21.8 Managing DBWalker Configurations 165
4.21.9 Managing DBWalker Stylesheets 168
4.21.10 Adding Configurations to Profiles 168
4.22 SOAP API 169
4.22.1 SOAP Overview 169
4.22.2 SOAP API vs. XML Output 169
4.22.3 Getting the WSDL 169
4.22.4 Global vs. per-profile WSDLs 170
4.22.5 Configuring the SOAP Interface 170
4.22.6 C# example project 171
4.22.7 SOAP Links for Languages 171
4.22.8 SOAP API search Reference 172
4.22.9 SOAP API admin Reference 174
4.23 Thunderstone ISAPI Proxy Module 180
4.23.1 Overview 180

4.23.2 Requirements 181
4.23.3 Installing the Proxy Module 181
4.23.4 Post-Install Setup 181
4.23.5 Manually Configuring the Proxy Module 184
4.23.6 Troubleshooting the Proxy Module Authentication 187

5 Reference 189

5.1 Supported File Formats 189
5.2 Database and File Usage 192
5.3 Walk Database Tables and Fields 192
5.4 Options Table Fields 195
5.5 Customizing the Search 196
5.6 Customizing the Walker 196

6 Search Interface Help 197

6.1 Forming a Query 197
6.1.1 Query Rules of Thumb 197
6.1.2 Overview of Query Abilities 198
6.1.3 Controlling Proximity 198
6.1.4 Ranking Factors 198
6.1.5 Keywords Phrases and Wild-cards 198
6.1.6 Applying Search Logic 199
6.1.7 Natural Language Query 200
6.1.8 Using the Special Pattern Matchers 200
6.1.9 Invoking Thesaurus Expansion 201
6.2 Using Word Forms 201
6.3 Controlling Proximity 202
6.4 Interpreting Search Results 202
6.4.1 Viewing Match Info 203
6.4.2 Finding Similar Documents 203
6.4.3 Showing Document Parents 203

A Third-Party Software **205**

A.1 Antiword 205

A.2 Aspell 205

A.3 Catdoc xls2csv 206

A.4 Cole library 206

A.5 iconv 206

A.6 libxml2 206

A.7 Libxslt 207

A.8 Libexslt 207

A.9 JDBC drivers 208

A.9.1 Oracle JDBC driver 208

A.9.2 JTDS JDBC driver 212

A.9.3 PostgreSQL JDBC driver 212

A.9.4 MySQL JDBC driver 213

A.10 ppt2html, msg2html 213

A.11 SSL/HTTPS plugin 213

A.12 unrar 216

A.13 unzip 217

A.14 zlib 218

A.15 SpiderMonkey (JavaScript-C) Engine 218

A.16 PDF/anytotx plugin 219

A.17 thttpd - throttling HTTP server 219

A.18 RedHat Linux 219

A.19 CentOS Linux 220

A.20 Webmin 220

A.21 Java 220

A.22 OpenSSL RPM 226

A.23 RAID utilities 226

A.24 LCDpoc 227

A.25 GNU General Public License 227

A.26 GNU Lesser General Public License 234
A.27 GNU Library General Public License 244
A.28 Netscape Public License 253
A.29 UnixUtils 261
A.30 PuTTY 262

Chapter 1

Overview

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

It consists primarily of the Texis binary program and two Vortex scripts that are run by the Texis CGI program on the Search Appliance 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.

1.1 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 the Search Appliance's behavior to your needs.

1.2 Technical Support

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

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

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

If you subscribe to periodic notifications, and at some point in the future no longer wish to receive them, you may select "subscribe" again to enter the administrative area where you may delete your subscriptions.

Chapter 2

Installation

2.1 How to unpack and install the Search Appliance

For basic information about unpacking and installing the Search Appliance, refer to the Getting Started guide. This printed guide was shipped with the Search Appliance. In addition to the instructions it provides, it includes a sticker that lists important information unique to your Search Appliance. This information includes the original password and various network addresses.

Ensure that the Search Appliance is plugged into a UPS or other power conditioner.

Thunderstone is not responsible for damage caused due to environmental problems including, but not limited to, spikes, over/under voltage, outages, static discharge, excessive moisture or condensation, and extreme heat or cold.

Data loss may occur if the Search Appliance loses power without being shut down gracefully.

The Search Appliance is shipped with the network configuration specified when ordering. If no network configuration was specified then the Search Appliance is pre-configured to get its IP information automatically from DHCP. To change the network configuration if properly configured for your network, or to view the IP of the Search Appliance, you will need to use the console menu or front panel LCD.

2.1.1 Console Menu

To use the console menu you need to attach a keyboard and VGA monitor to the standard ports on the rear of the Search Appliance. You will see the following menu:

```
    Thunderstone Search Appliance Summary on YYYY-MM-DD HH:MM:SS
Thunderstone phone: 216-820-2200
MAC Address       : xx:xx:xx:xx:xx:xx
Ethernet 1        : xxxxx
Ethernet 2        : xxxxx
IP Address        : xxx.xxx.xxx.xxx
Index Admin Page  : http://xxx.xxx.xxx.xxx/texis/dowalk
System Admin Page : https://xxx.xxx.xxx.xxx:999

    Thunderstone Appliance Setup and Information
   T) Thunderstone information
   N) Network and system information
   S) Setup network
   F) drop Firewall/NAT (Allow all network access)
   A) drop Admin restrictions (HTTPS,IP,Cipher requirements)
   I) dIagnostics
   R) Reboot system
   D) shutDown system
   H) Help
   L) Login (Thunderstone tech support only)
Enter your choice:
```

Choose an item by entering the letter on the left of the item.

- **T) Thunderstone information**
  will display information about the Search Appliance software version, license information.
- **N) Network and system information**
  will display information about the network and firewall settings.
- **S) Setup network**
  will prompt for network info and change settings accordingly. Entering ”q” for any prompt will abandon changes and return to the menu.
  - **Use DHCP?(y/n/q)**
    Enter ”y” to get IP information automatically from DHCP and not ask further questions. Enter ”n” to specify the IP information manually.
  - **Enter IP address or q:**
    Enter the IP address in the form `#.#.#.#`.
  - **Enter net mask, blank, or q:**
    Enter the IP netmask in the form `#.#.#.#`.
  - **Enter gateway, blank, or q:**
    Enter the network gateway IP address in the form `#.#.#.#`.
  - **Primary nameserver, blank, or q:**
    Enter the IP of the primary nameserver in the form `#.#.#.#`.
  - **Secondary nameserver, blank, or q:**
    Enter the IP of the secondary nameserver in the form `#.#.#.#`. Leave blank for no secondary nameserver.

  - **Enter hostname, blank, or q:**
    Enter the host name of the Search Appliance. Ideally this will match whatever your DNS has for the assigned IP. At the very least it should be a fully qualified name such as ”app.mydomain.com” so that email notifications etc. can work. It should never be a single word like ”mysearch” as that can delay the boot process.
  - **Enter DNS Domain Name, blank, or q:**
    Enter the domain to use on unqualified host names, such ”mydomain.com”. Or leave blank to prevent unqualified names from resolving.

- **F) drop Firewall/NAT (Allow all network access)**
  will drop all firewall ant NAT rules that you may have setup using the webmin web interface. It will also permanently delete the firewall and NAT configs so they will not come back on upon boot. This is most useful when you’ve accidentally created a firewall configuration that locks you out of the admin interface.

- **A) drop Admin restrictions (HTTPS,IP,Cipher requirements)**
  will remove all system-wide settings that restrict the web admin interface. This is useful if you’ve accidentally created a config that locks you out of the admin interface. This is permanent. Any desired restrictions will have to be re-added from the web interface.

- **I) dIagnostics**
  will enter a sub-menu of network diagnostics as follows.

```
    Thunderstone Appliance Diagnostics | YYYY-MM-DD HH:MM:SS
   P) Ping
   T) Traceroute
   L) Lookup hostname
   F) Fetch a URL
   S) System view
   R) RAID info
   M) Main menu
Enter your choice:
```

- Menu items:

  - **P) Ping**
    will request a hostname or IP and send icmp packets to that host. It will run continuously until you press `^C`.
  - **T) Traceroute**
    will request a hostname or IP and trace the network route to that host.
  - **L) Lookup hostname**
    will request a hostname and, optionally, a name server IP and will lookup the IP for the given hostname on the configured, or given, DNS server.
  - **F) Fetch a URL**

will request a URL and attempt to fetch that URL. It will print out all of the received HTTP/HTTPS headers but not the content. You may also enter a FILE URL of the form the appliance takes to test mounted fileservers. See `Network Filesystems & Shares` (3.3).

  - **S) System view**
    will give a self-updating view of the running system showing actively running processes, CPU and memory utilization. Press `q` to quit.
  - **R) RAID info**
    will display the status of the RAID disk array.
  - **M) Main menu**
    will return to the main menu.

- **R) Reboot system**
  will stop and restart the Search Appliance.
- **D) shutDown system**
  will cleanly stop and power down the Search Appliance.
- **H) Help**
  will show a short help message.
- **L) Login (Thunderstone tech support only)**
  will provide low-level system login for extended troubleshooting. This can only be used by Thunderstone technical support staff.

2.1.2 Front Panel LCD

Most machines, except SBE models, shipped since 2009 have a front panel LCD display and operation buttons. The display can show information about the Search Appliance's configuration and activity level. It can also be used to change configurations and shutdown or reboot the Search Appliance.

There are 4 arrows used to scroll through menus and select items. The "check" button is used like a keyboard's "enter" key to finish or confirm choices. The "X" button is used to cancel or back out of a choice.

In the top-right corner of the LCD display there is a heartbeat indicator that should pulse every second. The LCD backlight will dim after a period of inactivity. The LCD backlight will flash when there is a problem reported.

Browse information about the Search Appliance from the main menu using the up and down arrow buttons. Press the "check" button at any time to return to the main menu. Available information:

- Link Speed and Duplex
- IP Config Type (Static or DHCP)
- Primary Ethernet MAC Address
- Current IP
- Current Gateway IP

- License Serial Number
- License Maintenance Expiration Date
- Current Hits per Day
- Current Documents in Database
- Current Searches per Minute
- Number of Running Walks
- Number of Currently Connected Clients (web browsers)
- Current System Load Average for 1, 5, and 15 Minutes.
- Current Network Usage I/O Rates
- Current Memory Usage
- Current Swap Usage
- Current Disk Space Usage
- Current RAID Status
- Host and Domain Names
- Thunderstone Contact Information

From the main menu adjust configuration of the Search Appliance using the ”check” button. Use up and down arrows to navigate through the choices. An asterisk, ”*”, will appear at the end of items that have been modified but not applied. At the end of the config list are the options to apply or trash(discard) the changes. Available configs:

- IP configuration method
  DHCP or Static. Press ”check” or left or right arrow to change the value.
- IP Address
  Only used if IP method is Static. Press ”check” to change the IP. Use up and down arrows to change the digits of each number. Use the left and right arrows to move between digits. Use ”X” to cancel your changes. Use ”check” to keep your changes and move to the next config.
- Netmask
  Only used if IP method is Static. Edit the value as with IP Address.
- Gateway
  Only used if IP method is Static. Edit the value as with IP Address.
- DNS 1
  Primary Nameserver. Only used if IP method is Static. Edit the value as with IP Address.

- DNS 2
  Secondary Nameserver. Only used if IP method is Static. Edit the value as with IP Address.

- Admin PIN
  Set this to non-zero to require this password for making changes via the LCD front panel. Press ”check” to change the PIN. Use up and down arrows to change the digits of each position. Use the left and right arrows to move between digits. Use ”X” to cancel your changes. Use ”check” to keep your changes and move to the next config.

- Drop Firewall
  Remove firewall restrictions etc. See same on console menu 2.1.1.

- Drop Restrct
  Remove admin restrictions etc. See same on console menu 2.1.1.

- Sys
  Shutdown or reboot the system. Use ”check” or left and right arrows to select the value. Use up or down arrows to move to another config.

- Apply Now
  Press ”check” to apply changes. Select ”No” or ”Yes” to confirm then press ”check”.

- Trash Changes
  Press ”check” to discard changes. Select ”No” or ”Yes” to confirm then press ”check”.

The LCD backlight and heartbeat are also configurable. From the main menu press ”X” then ”check” to escape to the ”Options” menu. Use up and down arrows to select the item to adjust. Use ”check” or left and right arrow buttons to cycle through choices. Changes to LCD settings are immediate. Press ”X” twice to return to the main menu.

Under normal circumstances the LCD main menu will alternate between

```
Thunderstone
Search Appliance
```

and

```
Up/Dn for info
Check to config
```

If the display is not alternating or shows ”`Cli:0 Scr:  0`” then the LCD menu is not functioning properly.

2.2 Customizing the Search Appliance's Appearance

You may make common changes to the Search Appliance'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.

Chapter 3

Operation

3.1 Running the Administrative Interface

The Search Appliance's administrative interface is a web application that you access using your web browser. Access it using `http://YOURSERVER/texis/dowalk` where `YOURSERVER` is the name (or IP address) of your Search Appliance.

When you run the administrative interface you will be asked for the login and password. By default there is one login name. It is `admin` 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 the Search Appliance, then you should logout when you are finished. This will help prevent unauthorized configuration.

The Search Appliance 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 the Search Appliance will be lost if JavaScript is turned off in your browser (e.g. to prevent pop-ups on other sites). In this document, the user interface description assumes that JavaScript is enabled.

3.2 First Time Run: Quick Start

Step 1: Create an Account

A password was created by Thunderstone for the default administration account (`admin`), 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 the Search Appliance (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 profile may already have been created by Thunderstone if you requested it when ordering. If so, you may click on the profile name and proceed to the last step, searching. Otherwise, choose a new profile name and 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. Use the `Base URL` setting to specify the starting point of your walk. This is often the homepage of a site, or the sitemap page.

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.

3.3 Administrative Interface Overview

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

```
Settings
  Basic Walk Settings
  All Walk Settings
  Search Settings
Tools
  List/Edit URLs
  List Duplicates
  Test Fetch
  Test Search
  Query Log
  Replication Tools
  Results Cache
  SOAP Tools
  Best Bet Groups
Status
Search
Profiles
System
  Information
    System Information
    Log Viewer
    Test Network and Servers
    Task Monitor
  Modules
    Thesaurus
    Static Content
    DBWalker
    Connectors
    Network Shares
  System Setup
    System Wide Settings
    Update Software
    RAID Array Management
    SSL/HTTPS Certificates
    Webmin System Management
  Backup/Restore Settings
    Backup Appliance Settings
    Restore Appliance Settings
  System Replication
    System Replication Queue
    System Replication Target Status
```

```
  Security
    Accounts & Groups
    Access Control Lists
  Advanced Tools
    Extra Downloads
    Repair Tools
      Check Version Upgrade Actions
      Re-output XSL files
      Re-schedule walks
Docs
```

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 3.4.

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

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

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

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

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

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

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

  The current settings from the form will be saved as is done when you click `Update`. Then a new walk will be started. The new walk will be performed to either a temporary database or the live database, depending on the setting of Rewalk Type (Section 3.4.14). Then you will be shown the walk status page where you may monitor the progress of the walk.

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

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

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

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. Please see section 3.5 for details on the individual settings.

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 3.6.

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

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

List/Edit URLs

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

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

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

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

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

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

Then `Submit`.

All matching URLs will be listed. Clicking on a listed URL opens a page of details about that URL. On that detail page, everything the database knows about that URL is presented. You can also see what pages refer to the selected page by clicking `Parents` and what pages the selected page refers to by clicking `Children`. The `test` link next to the URL can be used to do a live test fetch of the page to find out how the Search Appliance processes it. See `Test Fetch` 3.3.

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

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

Live Search and New Walking Database

These options are presented on the `List/Edit URLs` page (see 3.3) if a walk is active. They allow you to choose which database to query. The “Live” database is the one from a previous successful walk that is what search users see. The “New” database is the database currently being built by the new walk. It is not visible to search users.

List Duplicates

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

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

Test Fetch

This allows for testing Search Appliance fetching of URLs. Be sure to properly encode any entered URL, like space as `%20`.

Several processing options are provided to control how much processing to do. Expand the Options link to show and edit these options. Note that most will only be set at the request of Thunderstone tech support. Some options may produce copious messages.

- Full Processing
  Perform full processing on the fetched file as if it is being prepared for the search database (execute any relevant Primer URLs before-hand, apply rejection rules to its links, etc). Otherwise only perform the basic download of the page.
- Keep Download
  Keep the raw encoded download and decoded data for display. Using this can make the test results page particularly large for large source documents like PDFs etc.
- Fetch verbosity
  A hex or decimal integer value whose bits control verbose messages during the fetch, e.g. printing of HTTP request/response lines and headers. *Note*: a non-zero value may reveal sensitive data in messages, such as **Login Info**.
- Trace authorization level
  A hex or decimal integer value whose bits control authorization trace messages, e.g. for Basic/file/NTLM authentication. *Note*: a non-zero value may reveal sensitive data in messages, such as **Login Info**.
- Trace socket level
  A hex or decimal integer value whose bits control socket trace messages, e.g. data written to or read over the network. *Note*: a non-zero value may reveal sensitive data in messages, such as **Login Info**.
- Trace DNS level
  A hex or decimal integer value whose bits control domain name system messages, i.e. hostname resolution.
- Trace encoding level
  A hex or decimal integer value whose bits control transfer and content coding trace messages, e.g. for `gzip` encodings.
- Trace fetch level
  A hex or decimal integer value whose bits control fetch action trace messages, i.e. when network fetches actually occur.
- Trace calls level
  A hex or decimal integer value whose bits control fetch function call trace messages, e.g. for network settings.
- Trace script level
  A hex or decimal integer value whose bits control JavaScript trace messages.

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

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

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 to the Live Settings 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 Search Appliance 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.

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 result clicks, 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.

Replication Tools

Replication Tools allows you to work with replicating data from this profile.

- **View Replication Status**

  Replication Status shows you an overview of the contents of the replication queue. The data is presented grouped by host/profile, and the next items queued are detailed below.

- **Send Profile Settings**

  Send Profile Settings is used when you want to send this profile's configuration to another Appliance, possibly with a different name. If the specified profile doesn't exist on the target machine, it will be created and given this profile's settings. If it already exists, it will have its settings set to this profile's values.

  Enter a remote machine and profile name & hit `Send Settings` to queue up the Send Settings action in the replication queue, which you can view with the previous link.

- **Send Profile Data**

  Send Profile Data allows you send all the current profile's data to an established replication target. This can be useful when adding a new target to existing sender profile, and you need to load the target with existing content but don't want to perform a full walk on the sender.

  Select one of the machine/profile pairs listed, and hit `Send Data` to queue all this profile's content for replication. Please see the Replication Status page describe above to monitor the progress.

Results Cache

The `Results Caching` profile tool allows management of the results cache for the profile. If enabled (p. 115), the results cache can improve search response time by caching frequently used search results: if a later query is made with the same query string parameters, it may be found in the cache, thereby saving the time needed to run the query again.

On the `Results Caching` profile tool page, the status of the cache manager is reported (whether it is running, and what process ID). The cache manager runs in the background, deleting old entries and refreshing the cache if requested. The size of the cache table (if it exists) is also shown, both number of entries and size in bytes. Old entries are deleted as per the current Results Caching settings (p. 115).

Several actions can be performed on the cache, if results caching is enabled:

- `Clear` - Clears the results cache table. All previously cached results will be removed. This can also be used to free up disk space if needed.
- `Start Refreshing` - Mark all entries for refresh. The cache manager will then start refreshing the cache in the background, starting with the highest-priority entries: queries will be re-executed to ensure the cache entries reflect the latest crawl data. This can increase load on the machine, as the cache manager will now be executing queries (though only serially, one at a time). Existing cache entries marked for refresh – but not yet actually refreshed – are still otherwise valid, i.e. they may still be used to resolve user searches.

- `Stop Refreshing` - Mark all entries as not needing refresh. This will stop the cache manager from executing queries to refresh the cache, and thus may reduce machine load. The manager will still continue to run to expire old entries, etc.
- `Import` - Import cache queries (without results data) from another profile, and mark for refresh. This will copy the results cache from another profile, but since each profile's crawl data is different, the actual search results will not be copied, as they would be invalid. Thus, the entries will not have search data (and cannot service requests), but will be marked for refresh by the cache manager. This option can be used to "prime the pump" when creating a new profile to replace an existing live profile as the default: the existing profile's queries can be copied over and refreshed. When the refresh is complete (indicated by "no entries marked for refresh"), the new profile can be made live, and will already have a largely up-to-date cache.

Caveats: As results caching currently only takes into account the query string to differentiate requests, it should not be enabled under results authorization or other scenarios that utilize additional data (such as cookies or request headers) for search requests. The results cache is stored in the same database as the profile crawl data, so it is deleted and started anew for each New type crawl – just as the crawl data its results are based on starts anew with New type crawls. If a Refresh crawl is done that significantly changes the crawl, the cache should probably be cleared to avoid stale or out-of-date results; alternatively, the **Max Cache Entry Age** setting might be decreased.

SOAP Tools

Soap Tools provides various WSDLs and references for working with Search Appliance via the SOAP API. Please see the SOAP API section for more details (4.22)

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.

Status

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

During an active walk, it indicates a summary of how many pages are to be walked in the next hour, how many were walked in the last hour, and the total number of pages. There is a list of the most-recent URLs

fetched, with number of errors and duplicates found, followed by a list of the next URLs to be walked. Below that is summary information about the walk itself, including walk start time, starting URLs, and some profile settings. The Walk Status page updates automatically every 10 seconds until the walk is complete or another page is selected. (After 10 minutes of user inactivity it will refresh once a minute to save traffic.)

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

Walk Status tabs

If more than one database is available for viewing, tabs will appear at the top of the Walk Status, allowing you to swap between database. This can be caused by a ”New” crawl running (allowing you to switch between `New Walk Database` and `Live Search Database`), or if a ”New” crawl failed and automatically reverted to the previous database (allowing you to switch between `Live Search Database` and `Failed Walk Database`).

If more than one walk has been performed, then an `Archived Logs` tab will be available. This lets you view log files from previous crawls for errors or other unexpected behavior. No database or searchable content from these old walks is retained, only the log files.

Old archived logs are automatically cleaned out if they become too numerous or consume too much space, with the oldest logs removed first.

While a walk is occurring, multiple buttons are available:

- **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 (i.e. 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.

Search

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

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. You should not delete a database that is being actively searched.

You may also create a new profile by entering a new name.

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.

System

The System section contains various links for maintaining and editing operating-system and overall settings.

System Information

This page provides system information, such as network IP addresses, MAC address, kernel version, load, etc. It also allows killing of processes owned by the `texis` user, which includes walks.

Log Viewer

This allows you to view, delete, rotate, and email the Search Appliance logs. It lists every file in the log partition and allows you to manipulate each log individually or many at once. Each item has a check box for selection for mass operations, the date and time of last addition, the size, a link to see the most recent bit of that log, and a list of processes currently using that log. Clicking a column header will sort the list by that column.

There may be multiple versions of each log. The version with no numeric extension (.1, .2, etc.) in the filename is the current log. Those with numeric extensions are older logs. Extension .1 is the most recent old log, .2 is the second most recent, etc.. Logs are automatically rotated once a month or once a day if they exceed 10MB. The need for rotation is checked once a day around 4am. To force a rotation check you can click the `Rotate Logs` button at the top of the manage logs page. If you see a log file with numeric extension that also has process numbers listed in the `In Use By` column you'll probably need to reboot the Search Appliance to free up that log file. That should be a rare occurrence, but has been known to happen.

The log listing is divided into sections. The first, unnamed, section is the system level logs. They contain information about the core operating system of the Search Appliance. That's where hardware, network, and similar events and problems are logged.

The Apache section contains the usage logs for the Apache web server which is used for HTTPS access to the Search Appliance, if enabled.

The texis section contains logs related to Texis, the relational database server that is a major technical component of the Search Appliance. These logs provide detailed information about operational events of Texis.

The webmin section contains usage logs for the Webmin system management interface.

At the bottom of the page is a form which allows you to perform actions on the selected log(s). You may View, Delete, Send, or Download any of the logs. For deleting you will be asked to confirm the deletion before it's carried out so it should be difficult to delete a log by accident. Deleting logs listed as in use may require rebooting to reclaim their disk space.

For viewing, sending, and downloading you can choose how many lines of the log(s) to see and whether to see the newest lines first (reverse chronological order) or natural order with the oldest lines first.

For sending the logs to Thunderstone technical support you should fill in your email address and a ticket number given to you by Thunderstone. Sending the logs via email assumes that your Search Appliance is configured to send mail to the Internet. It is by default in the simple case but if you have outbound SMTP blocked by your firewall or need to use a mail relay you'll need to configure sendmail using the Webmin interface. An alternative to having the Search Appliance email the log is to instead download the log then attach that file to your email to Thunderstone tech support.

Test Network and Servers

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

- **Test URL fetch**

  This is the same functionality as found in `Test Fetch` in a profile's `Tools`. Please see its documentation (p. 35) for more detail.

- **Test Network**

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

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

Task Monitor

Provides an interface to manage the Task Monitor, which is a background/daemon process that automatically handles various Search Appliance tasks, e.g. updating indexes during settings replication. The queue of pending tasks can be viewed and/or deleted (only if needed by tech support).

Thesaurus

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

Each thesaurus may be used by zero or more profiles and should not be deleted if it is in use by a profile. Search options that affect the use of these thesauri are `Synonyms`(3.6.61), `Main Thesaurus`(3.6.62), and `Secondary Thesaurus`(3.6.63).

See section 4.3 for further details.

Static Content

Static Content allows you to upload any files you'd like the Search Appliance to also host. This can be used to serve images, javascript, or CSS files that are referenced by your XSL-customized search interface.

Note that this is not meant for searchable content, which remains on the original website or file server. This is for images or files used by the search interface, rather than file contents searched by the Search Appliance.

Content will be available under the `/custom/` directory, so if you upload `companyLogo.png` you can use `<img src="/custom/companyLogo.png">` in your custom search interface.

DBWalker

Configuration for the DBWalker module. Please see the DBWalker section (4.21) for more details.

Connectors

Google Connectors are 3rd party programs that allow you search data from large, complex data data sources (such as Microsoft Sharepoint). There are different connectors for different types of content - Sharepoint, Livelink, File Servers, etc. The connector is installed on a separate server, and the connector machine pulls content from the remote data source, and pushes it into the Search Appliance.

`Manage Google Connectors` allows you to specify the location of a Connector Manager that's been installed, using an address like `http://otherServer:8080/connector-manager/`. You can then instantiate any connectors installed on that Connector Manager, and configure them appropriately (varies by connect).

Connector Managers must be listed in the Appliance's `Cluster Members` to allow it to push in data. Adding a new Connector Manager automatically adds it to Cluster Members.

For more information on Google Connectors, and to download the Connector Manager and connector packages, please visit http://code.google.com/apis/searchappliance/documentation/connectors/

Network Shares

Use this interface to mount remote file server(s) to the Search Appliance so that it may be indexed into one or more walk profiles.

All created mounts are permanent until manually removed. They will be remounted upon reboot of the Search Appliance.

To mount a remote filesystem or share select the type from the drop-down list and click Add.

NFS filesystems - Unix/Linux/etc. servers

Server: Enter the hostname of the server. CaSe does not matter. (e.g.: `nas1.mycompany.com`)

Directory: Enter the full path of the directory to mount as it is exported from the server. (e.g.: `/documents/internal`)

Reliability: Select `Hard` or `Soft`. Hard will cause the Search Appliance to keep retrying the same file forever in the event of an error reaching the server. Soft will allow files to fail if the NFS server can not be reached.

NFS Version: The highest NFS version to use. Leave this at 3 unless you have problems with old NFS servers.

CIFS - Windows 2000+

Server: Enter the hostname of the server. CaSe does not matter. (e.g.: `nas1.mycompany.com`)

Share: Enter the name of the share as exported by the server. (e.g.: `internal`)

Login Name: Enter the login name for the account to use to access the files on this share. This should be a user that has permission to read all the files that need to be indexed.

Login Password: Enter the password for the selected login name.

Server IP: Rarely used. If the ”Name” of the computer is different than it’s DNS name it may reject mount requests to the ”wrong” name. In that case enter whatever name makes the server happy into the `Server` field and enter the machine’s IP address into this field.

NOTE: When using Windows 2003 server you may need to change a setting on the server to allow mounting from the Search Appliance. If the share won’t mount try setting `control panel->admin tools->domain security policy->security settings->local policy->security options->Microsoft network server: Digitally sign communications (always)` to disabled.

SMB - Windows The `SMB - Windows` mount is an older system, which has been replaced by `CIFS - Windows 2000+`. If your file server doesn’t support CIFS or if you’re having problems with CIFS, you can try using SMB instead, which takes the same options as CIFS.

Current mount list

Under the Add form is the list of currently mounts and their status. Each mount has a `Remove` link to unmount the filesystem and remove it from the list. The options for each mount may be clicked to examine or modify the options and remount the filesystem.

If an entry shows as ”unmounted” there is a problem with the settings and it is not able to be mounted as is. If it was a transient problem with the server click the options then click ”Save changes” without making any changes to retry the mount.

Under the options for each mount is also an example of the minimum `Base URL` you would enter into a profile to index the files on that filesystem.

The `Technical Info` link shows some internal details about the mounts that may be helpful to tech support if you have problems.

Note: This feature appeared in scripts version 5.4.11. Prior to that `Webmin` was the only way to manage remote mounts.

System Wide Settings

This area is for settings that affect the Search Appliance as a whole and/or may be shared by multiple walk profiles. Please see section 3.7 for documentation on the individual settings.

Update Software

This allows you to manually initiate a check for software updates. It provides a list of available updates, allows you to select which updates to download, and allows you to manually initiate the installation of the downloads.

Refer to `Getting Software Updates` (4.4) for the procedure to manually perform updates.

If your Search Appliance has a CD drive, it also allows you to install software updates from a CD.

RAID Array Management

Note: This area only applies to larger models (such as 3000) that include multiple hot-swap disks in a RAID-5 configuration.

Note: The pages in this area may load somewhat slowly as they collect information from the RAID controller.

Overview

Little to no maintenance of the RAID array is required. In the event of a disk failure the hot-spare will automatically take over and the array will be automatically rebuilt. The rebuild process takes several hours. After the array is rebuilt the failed disk will have to be manually removed (no shutdown required) and replaced with a same size or larger disk of the same type. After the disk is replaced it needs to be `Add`ed to the array.

Details

The `RAID Status` page displays a summary of the RAID's state. It's an abbreviated form of the information on the RAID Management page to provide a quick Good/Bad check.

The `RAID Management` page lists information about the overall RAID array as well as each of the hard disks in the system. Each item starts with a Status and is color-coded to indicate it's state. Green is good, red is problem, blue is hot-spare disk, light blue is unused disk, yellow is verifying/testing.

The first line of the Storage table contains information about the overall array with the `Use` column set to `Array`. The remaining lines are individual disks, either `Member`, `Spare`, or `None`. Member disks are part of the RAID array. Spare disks are hot-spares that will take over for a failed member disk.

Each item in the Storage table has associated actions that may be taken.

Rebuild An array that is in a non-optimal state may be forced into a rebuild to become optimal again.

Verify Verifies the integrity of the parity information for the array. This is not generally needed as the array is automatically verified periodically as controlled by the hardware BIOS.

Fail Forces an individual disk into a failed state so that it may be replaced. This is not generally needed as failures will be automatically detected.

Remove This removes a hot-spare disk from the array. It then becomes an unassociated disk with a Use of `None`. All arrays should have a hot spare.

Add This adds an unassociated disk with a Use of `None` to the array as a hot-spare.

The first number in the `Disk:Addr` column (everything up to the :) is the disk number which corresponds to labels on the front panel of the Search Appliance.

Rebuild/Verify Rate

The Rebuild/Verify Rate is how aggressively the RAID will rebuild. A higher rate will rebuild a partially failed array more quickly so that it's in a non-fault-tolerant state for the shortest possible time. The downside of the higher rate is that operations that use the disk such as walks and searches will see slower performance.

Controller

This table shows various model and version information about the RAID controller.

Command

The Command input box should not be used except at the request of Thunderstone technical support. It is for issuing arbitrary commands to the RAID controller. Putting the wrong thing in this field could irrevocably damage the RAID array and render your machine completely unusable! If the "Ok to run this command" checkbox is not also checked anything in the command input will be ignore.

Perform

The Perform button at the bottom will perform all of the actions selected on the form. You must also set "Are you sure you want to perform these actions?" to "Yes" or the actions will not be performed.

Front panel

This provides a rough approximation to the physical front panel of the Search Appliance. It shows the drive arrangement to aid in locating the proper disk when performing maintenance.

SSL/HTTPS Certificates

This allows you to manage the server certificates provided by the Search Appliance when serving pages via HTTPS. The admin interface, including Webmin, and search will use the same certificate. By default the Search Appliance has a self-signed certificate. If you have multiple hosts you may need to regenerate the

self-signed certificate before your browser will allow you to access the second host using HTTPS. If you want to use HTTPS for searches you'll want to obtain a secure certificate from a trusted authority so that end users don't get warnings in their browser.

If you're familiar with requesting and obtaining/creating secure certificates and have a key and certificate pair ready to install you can use the `Enter a premade Private Key/Certificate pair` option at the top of the `Manage SSL/HTTPS Server Certificates` page. You will be presented with 3 large input boxes where you can paste in your Private key, Certificate, and an optional Intermediate Certificate that may be provided by your certificate authority.

You can generate a self-signed certificate or a CSR that can be provided to a certificate authority to request a secure certificate by filling in the boxes on the `Manage SSL/HTTPS Server Certificates` page. If you just want a self-signed certificate to use for encryption but don't care about authoritativeness you can check `Self sign` and enter the number of days you want the certificate to be good for then click the `Install Certificate` button. If you selected `Self sign` then you're finished. Otherwise click the `Generate CSR` button to generate the CSR.

When generating a CSR you will be presented with a block of text beginning with `-----BEGIN CERTIFICATE REQUEST-----` and ending with `-----END CERTIFICATE REQUEST-----`. You need to send everything between, and including, those lines to your certificate authority. The certificate authority may ask what type of server you're using or what format of certificate you need. Tell them you need an Apache compatible certificate.

After the certificate authority has confirmed your CSR they will provide a similar but different block of text bracketed with `-----BEGIN CERTIFICATE-----` and `-----END CERTIFICATE-----`. Paste that entire block, including the BEGIN and END lines, into the `New Certificate` box. They may also provide an "Intermediate Certificate" that you would need to paste into the `New Intermediate Certificate` box. If they don't provide an Intermediate certificate leave the `New Intermediate Certificate` box empty.

Once you generate a CSR the certificate management page will only present the option of installing the new certificate(s) from that CSR. If you need to regenerate the CSR or want to abandon the old CSR for any reason click the 'Cancel CSR' button on the certificate form.

If you have set the Search Appliance to require HTTPS admin and manage to install a certificate that you can't use or somehow prevents HTTPS access you can re-enable HTTP admin by going to the physical console of the Search Appliance and selecting the `drop Admin restrictions (HTTPS,IP,Cipher requirements)` option.

Webmin System Management

This area has its own login and allows for control of various low-level system settings. The login is `admin` using the same password as the `admin` account in the normal interface. If the password gets out of sync somehow it may be reset by setting the admin password from the Accounts area (3.3).

- **Network** Configure the IP address, DNS servers, and routing.
- **Firewall** Restrict access by IP.

- **Clock** Synchronize the Search Appliance to your local time.
- **Email delivery** Configure how to send email for walk notifications etc.
- **Shutdown** Shut the system down cleanly and power off.

Backup Appliance Settings

This allows you to save all of the current profile and most of the system settings from the Search Appliance to an XML file on your local workstation. (Mounted filesystems and IP configurations are not currently saved.) This file can be used to aid in cloning Search Appliances for a cluster and as a backup in the event the machine needs to be restored from scratch.

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

”Internal Settings” includes things that aren’t used when restoring, such as currently running Process IDs. This should only be included at the request of Thunderstone Support.

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

Restore Appliance Settings

Use this option to restore settings that you’ve previously captured using `Save Appliance settings`. Missing profiles will be created, and existing profiles will have their settings set to the values contained in the backup.

System Replication Queue

Provides an interface to view and manage the queue for “non-profile” content when using System Replication Settings. Non-profile content is things that do not apply just one profile, such as `System Wide Settings`.

System Replication Target Status

Allows you to check the status of all the profiles on this sender machine against all System Replication targets at once, indicating which profiles are present and which aren’t. It also allows you to create the missing profiles on the targets.

Accounts & Groups

This section provides information to maintain multiple login accounts for access to the Search Appliance administration. All users are listed on this page. You may add users, delete users, and change individual user passwords. The default user, called `admin`, 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.

Usernames may only contain letters, numbers, and underscores, they must begin with a letter, and they must be 20 characters or less. Names and passwords are case sensitive.

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.

User groups may be created on this page, by clicking the `Add a Group` link. Existing groups may be edited or deleted with the appropriate links. User groups are used to associate administrative users into similar-privilege groups for easier access control maintenance. See the User Groups section for more details (p. 127).

Access Control Lists

The Access Control page allows configuration of administrative users' access to administrative actions (creating profiles, starting walks etc.). In conjunction with user groups, access control can be used to restrict certain users to only certain actions, instead of allowing all users access to all administrative functions. See the Access Control section for more details (p. 127).

Extra Downloads

Provides links to auxillary downloads for the Search Appliance. This includes example programs for interacting with the Search Appliance.

Repair Tools

Provides internal verification (and any necessary fixes) of Appliance systems, including actions automatically taken at upgrades.

These are tools are meant to address situations that the Appliance should not normally find itself in Only use at the request of Tech Support.

Check Version Upgrade Actions

There are internal actions automatically performed when the Search Appliance upgrades between major versions, such as creating new database tables. If something unexpected went wrong in the upgrade process, this section lets you check for and re-apply the upgrade actions. This is usually not needed, and should only be done under the advice of Thunderstone Support.

Re-output XSL files

In the past, when a profile was restored from backup or made as a copy, it was possible for a profile's XSL files on disk to become out of sync with the profile's settings. This has been fixed, but customers that had restored profiles using old software may have profiles in this state.

This tool checks which profiles have XSL settings and files, and will re-write the profile's XSL data to disk if necessary. Not normally needed.

Re-schedule walks

In the past, when a profile was restored from backup, it was not actually scheduled with the profile's rewalk schedule setting. This has been fixed, but customers that had restored profiles using old software may still have profiles in this state.

The `Re-schedule walks` section confirms that all profiles' schedules match their rewalk schedule settings, and allows the re-application all profile's scheduled settings to the walk scheduler. Not normally needed.

Docs

This provides a hyperlink to the online version of this documentation. It also contains a link to download a PDF version of this documentation.

3.4 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.

3.4.1 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.

3.4.2 Notes

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

3.4.3 Base URL

Syntax: one or more URLs, one per line

This is the address where the web walker 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; the Search Appliance'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 (i.e. the one with the longest prefix) will also be fetched. This helps ensure that login cookies are properly restored to allow the Search Appliance 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` 3.5.7, `URL URL` 3.5.8, `Single page` 3.5.9, `Page file` 3.5.10, and `Page URL` 3.5.11 for more ways to specify URLs.

3.4.4 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. The Search Appliance 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` 3.5.16 which is the same but allows more than one domain. These options may

be used together.

3.4.5 Robots

Syntax: select Yes or No buttons

robots.txt

With this set to Yes, the Search Appliance 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` 4.5.

Meta

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

See also `Robots.txt` 4.5.

3.4.6 Allow Extensions

Syntax: one or more file extensions separated by space

A list of the URL extensions that the walker will accept. The default list is empty, i.e. all extensions are allowed. Include the `.` in each extension.

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
```

3.4.7 All Extensions

Syntax: select Yes or No button

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

3.4.8 Exclude Extensions

Syntax: zero or more file extensions separated by space

A list of URL extensions that the walker will reject. The default is empty, i.e. no extensions will be rejected. Include the `.` in each extension.

3.4.9 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` 3.5.19 and `Exclusion prefix` 3.5.20 for more ways to exclude URLs.

3.4.10 Walk Delay

Syntax: a decimal number from 0 to 10

Causes the Search Appliance to wait the specified number of seconds between page fetches. Normally set this to 0, and the Search Appliance will fetch and process pages as quickly as it can. Increase the Walk 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 Walk 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`(3.4.11) to 1. A delay defeats the advantage of multiple threads and large delays could cause unexpected page fetch timeouts.

3.4.11 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.

3.4.12 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 3.1: 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.

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

3.4.13 Disable Walks

When set to `Y`, no walk will launch for this profile for any reason (manually run, schedule, etc). Note that even if set to `N`, walks may still be globally disabled if the System Wide Setting `Disable All Walks` is set.

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

3.4.14 Rewalk Type

Syntax: select from drop down box

This determines how rewalks are performed.

New

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

Refresh

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

Here are other considerations for using `Refresh`. Pages that were referenced but were missing in the initial walk (the walk prior to the `Refresh`), but were added after the initial walk, will be missed by

`Refresh` if their parent page has not been modified. If you change your settings to be more inclusive (i.e. 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 the Search Appliance 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 the Search Appliance whether the page has changed.

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

3.4.15 Rewalk Schedule

Syntax: select from drop down boxes

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

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

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

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

See also `Notify` 3.5.2. If you are using "`On Change`" see also `Watch URL` 3.5.1.

3.4.16 Action Buttons

These buttons tell the Search Appliance 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 (3.3) for details about the operation of these buttons.

3.5 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`.

3.5.1 Watch URL

Syntax: an HTTP URL

The URL specified here will be refreshed every time that The Search Appliance 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.

3.5.2 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.

3.5.3 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. See also `Rotate Schedule` (section 3.6.3).

3.5.4 Categories

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

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

Table 3.2: Example Categories

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

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

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

Note: URL Patterns will not be used to determine categories if any Data From Field rules set `Category`. Please see the **Data From Field** (p. 64) for more details.

3.5.5 Categories Type

Syntax: radio button choice

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

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

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

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

For best search performance, categories that overlap one another – i.e. contain pages in common – should be listed most-commonly-searched first in the **Categories** setting list. Also, the `CatnoLowest` field should be selected as one of the **Compound Index Fields** (p. 76. These guidelines will allow the `Auto-detect` mode to optimize the most searches to the fastest possible speed.

3.5.6 DBWalker

Here you can select one more more database walking configurations to include in this profile. This can be done in addition to specifying any Base URLs (section 3.4.3). To select multiple configurations, hold `Ctrl` while clicking in the select box.

For more information on the database walker module, please see the **DBWalker** (4.21, pg. 162) of the manual.

3.5.7 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 3.4.3. 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.

3.5.8 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 3.4.3. 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 walk 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.

3.5.9 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.

3.5.10 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` 3.5.9.

3.5.11 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` 3.5.9.

3.5.12 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.

3.5.13 Keep Query Vars

Syntax: comma-separated list of query variable names

If set, the Search Appliance will remove all but the these query variables when walking URLs. This can be useful when you know only a few select query variables are important, and others may appear but are irrelevant.

All the named variables are not required to be present, they are simply the ones that will be let through.

For example, setting Keep Query Vars to `id,type` will turn the URL

```
http://www.mysite.com//page.aspx?state=A3N4&id=432&printit=Y
```

into

```
http://www.mysite.com//page.aspx?id=432
```

3.5.14 Ignore Query Vars

Syntax: comma-separated list of query variable names

If set, the Search Appliance will remove these query variables when walking URLs. This can be useful when you can establish a few irrelevant query variables, but anything else would be significant.

URLs with these query variables do not cause any kind of error, they simply have the variables stripped out and continue processing normally.

For example, setting Ignore Query Vars to `printit` will turn the URL

```
http://www.mysite.com//page.aspx?state=A3N4&id=432&printit=Y
```

into

```
http://www.mysite.com//page.aspx?state=A3N4&id=432
```

3.5.15 Ignore Case

Syntax: select Yes or No button

This tells the Search Appliance 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.

3.5.16 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` the Search Appliance will walk all of `www.mysite.com` and any URLs referring to any machine in `othersite.com`.

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

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

3.5.17 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` the Search Appliance 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.

3.5.18 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 walked 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. 61.

3.5.19 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 3.3: Exclusion REX examples

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

See also `Exclusions` 3.4.9, `Exclusion prefix` 3.5.20 and `Exclude by Field` 3.5.22.

3.5.20 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` 3.4.9, `Exclusion REX` 3.5.19 and `Exclude by Field` 3.5.22.

3.5.21 RSS Feeds

Syntax: select from options

`RSS Feeds` determines the behavior taken when a RSS or Atom feed is encountered during a walk, either by directly linking or an embedded `<link rel="alternate">`.

- `Follow Links Only` *(default)* - Links listed in the feed are followed, but the text content feed itself is not indexed.
- `Index Content and Follow Links` - Links listed in the feed are followed, and the feed itself is considered searchable content. The title and description of the feed are indexed, as well as the titles of the entries in the feed.

3.5.22 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 (i.e. a typical search on the Search Appliance) is entered: e.g. 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` 3.4.9 and `Exclusion REX` 3.5.19.

3.5.23 Additional Fields

Syntax: Name, Type, Searchable, Sortable, Output

The additional fields allow you to add up to three additional fields to the index. These fields can be included in the output (if you use the XML output style), sorted on, and searched on. They are populated with the **Data from Field** settings (p. 64).

- **Name** - specifies the name of the additional field. It also specifies element that will hold the field contents if it is output in XML. The name must be a valid XML element name (may contain only `alnum` or `-_` and must start with a letter or _).
- **Type** - specifies the internal storage type for the additional field. Anything can be stored as `Text`, but if you want to do numeric or date comparisons (such as sorting), you have to use an appropriate data type.

- **Searchable** - specifies whether this additional field is directly searchable. This is done with an additional URL parameter that is separate from the normal query. Please see the **Additional Fields** section of **Procedures and Examples**, p. 161, for more details.

- **Sortable** - specifies whether you allow sorting by this additional field. This is done with the `order` search parameter. Please see the **Additional Fields** section of **Procedures and Examples**, p. 161, for more details.

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

Note that once Additional Fields are created and used, changing their order or **Type** may alter other settings that use them, such as **Data from Field** (p. 64), **Index Fields** (p. 76) or **Compound Index Fields** (p. 76).

3.5.24 Data from Field

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

This provides alternate means of setting both the HTML fields (`Modify Date`, `Title`, `Description`) and any Additional Fields. It allows getting page information from non-default places by searching and optionally replacing the data. New blank rows will be provided as rows are used. See below for examples.

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

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

- `HTML`
- `Text`

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

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

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

- `HTML` - the raw HTML source of the page.
- `Text` - the text of the page, after HTML rendering has been applied.
- `Title` - the HTML title of the page
- `All Meta` - the contents of all `meta` headers specified in the HTML page.

- `Meta Field ->` - the contents of a specific `meta` field, specified in the next input box, **From Meta Field**.
- `Keywords` - the contents of the `keywords` meta header.
- `Description` - the contents of the `description` meta header.
- `Mime Type` the MIME type of the page. This may have been derived from the `Content-Type` header, a `<META HTTP-EQUIV>` tag, or the URL extension, depending on what is available.
- `URL` - the URL of the page.
- `URL Decoded` - the decoded version of the URL. Any `%XX` 'URL-safe' sequences in the URL are replaced with their real characters. E.g. `Pre%20%2D%20Expense%20Report.doc` is decoded into `Pre - Expense Report.doc`.
- `URL Protocol` - the URL's protocol, e.g. `http`.
- `URL Host` - the host (without port number) from the URL.
- `URL Host and Port` - the host (and port number if given) from the URL.
- `URL Path` - the file path from the URL.
- `URL Path Decoded` - the file path from the URL, URL-decoded.
- `URL Anchor` - the anchor from the URL (if any), i.e. the part after the # (pound sign). May not be available if already stripped.
- `URL Query` - the query string from the URL (if any), i.e. the part after the `?` (question mark).
- `URL Query Var ->` - the value of the URL query-string variable named in **From Meta Field**, URL-decoded.
- `Referrer's Data` - the value of a referring pages field. Store refs is required for this. The field selected will be the same field being populated.

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

Entering text in this field will force the use of `Meta Field ->`, if From Field is set to anything besides Meta Field or URL Query Var.

To Field - specifies where information should be stored.

- `Modified Date`, `Title`, `Description`, `Keywords`, `Depth`, and `Body` - Override the standard fields extracted from the content.
- `Authorization Url` - Populates the Url used when checking this result for Results Authorization. Please see the `Allow Authorization Url` section (3.6.54) for more details.

- `Category` - To populate the category via Data From Field, all the possible category names must be entered in the Category setting. Using one or more Data From Field rules to set Category will cause the Appliance to ignore the Categories' URL Patterns and instead set category membership based on these Data From Field rules.

  Note: due to the way categories are stored, if categories are added, reordered, or removed after content has been walked, then a New walk will need to be performed to update the content's categories. Renaming categories does not need a rewalk.

- `Subfetch` - This causes the Search Appliance to take the value it finds, and performs a fetch of that URL. Nothing is changed by the subfetch, but any further Data From Field rules will use that fetched document as the source of its content. Please see the Subfetch example below for a situation where this could be used.

- `Additional Fields` - If this profile has any Additional Fields, they will be available as a target `To Field`.

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

Data From Field Example - Using Description for Title

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

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

Data From Field Example - using PublishDate for Modified Date

If you're walking 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 walking 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>`.

Data From Field Example - Subfetch to use PDF contents for a Web Page

`Subfetch`es allow you to use content from other URLs to populate the current URL's record. We may have a site about articles, where each article has a web page describing the article, and a link to a PDF of the actual article. We'd like searches that match article contents to take us to the web page, not the article PDF itself.

If the web page has a meta header called "pdfLink" with a URL to the article PDF, we can use the body of the PDF as a replacement for the web page's body with two DataFromField rules like this:

REX Search - *(Empty)*
Replace - *(Empty)*
From Field - `Meta Field ->`
From Meta Field - pdfLink
To Field - `Subfetch`

REX Search - `.+`
Replace - *(Empty)*
From Field - `Text`
From Meta Field - *(Empty)*
To Field - `Body`

The `Subfetch` DataFromField rule fetches the URL specified in the `pdfLink` header. While this grabs the PDF, it doesn't change anything on its own. We then pull from the PDF's text output, and use that as the `Body` of the current web page.

3.5.25 Required REX

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

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

3.5.26 Required Prefix

Syntax: zero or more URL prefixes, separated by whitespace

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

3.5.27 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.

3.5.28 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.

3.5.29 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.

3.5.30 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.

3.5.31 Max URL Size

Syntax: an integer from 1 through 2033

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

3.5.32 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 (i.e. Keep-Alive requests), if the server and protocol support multiple requests. Multiple requests per connection increases walk speed, and is needed for Windows/NTLM-protected pages. The default is `100`.

3.5.33 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).

3.5.34 Page Timeout

Syntax: a whole number from 1 up

Causes the Search Appliance 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.

3.5.35 Meta Tags

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

This option tells the Search Appliance 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.

3.5.36 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.

3.5.37 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.

3.5.38 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 the Search Appliance 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.

3.5.39 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, e.g. a Windows charset.

3.5.40 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 (e.g. 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.

3.5.41 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.

3.5.42 Keep Links

Syntax: select Yes or No buttons

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

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.

3.5.43 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.

3.5.44 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.

3.5.45 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.

3.5.46 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. E.g. with the setting "`-@`", the text "`part 123-45@6`" would be stored (and searchable as) "`part 123456`" instead.

3.5.47 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 e.g. "`#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 (i.e. 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.

3.5.48 Language Analysis

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

Language

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

Preserve 7-bit

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

3.5.49 CJK Mode

Syntax: select Yes or No

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

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

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

3.5.50 Unknown File Formats

Syntax: select Exclude or Include

Unknown File Formats controls how files in an unknown format (e.g. binary content not identified as PDF, Word, text, etc.) are handled. If set to `Exclude` (the default), such unknown formats' data will be ignored; this avoids bloating the crawl database and query autocomplete dictionary with garbage binary content.

If set to `Include`, the data will be included. This might help find words in otherwise-unsearchable binary files, but is unlikely to succeed: since the file format is unrecognized, all that can be done is a simple

`strings`-like scan for ASCII words in the file. If the file does not store words in an ASCII format, only garbage binary content will be returned.

Note that unlabeled plain-text files – i.e. those not identified by MIME type nor by file extension "`.txt`" – will generally be identified by a natural language scan (if running Texis 7.01 or later), and properly passed as-is. This setting only applies to files that fail that test, i.e. are unlikely to be plain text. Added in Appliance 9 / Webinator 7.

3.5.51 PDF Title Action

Syntax: select option

Controls how titles are handled for PDF files.

- `Automatic (default)` - Use the internal title if one is set, otherwise generate a title.
- `Always Generate` - Always generate a title for the PDF, ignoring any title set in the document.
- `Never Generate` - Always use the internal title from the PDF. If there isn't one, the title is left empty.

Binary files like Word documents can have titles set internally. They're usually unset, and if they are set, it's likely to a good title that should be used.

PDF files often differ from this. Many PDF converters set the PDF's title to the original filename that the PDF came from. This results in a PDF with a title like `expenseReport.doc`, which shows up as the link text in the search result. Users click on what appears to be a Word document, only to arrive at a PDF document.

Setting `PDF Title Action` to `Always Generate` tells the Search Appliance to ignore any internally set title for PDF files, and always generate one based on the content.

3.5.52 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. 111) 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.

3.5.53 Text Search Mode

Syntax: select from options or enter custom mode

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

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

- `Loose` - Ignore case, ignore diacritics (accents), expand ligatures, ignore width differences. **Storage Charset** should be empty or `UTF-8`, though `ISO-8859-1` may sometimes work. With this mode, not only will a lower-case "`e`" match an upper-case "`E`" and vice-versa (ignore case), but "`e`" will match "`é`" (Unicode U+00E9), "`oe`" will match "œ" (U+0153), and full-width will match half-width characters (for ASCII and katakana).
- `Strict` - Ignore case only. "`e`" will match "`E`", but not "`é`". **Storage Charset** should be empty or `UTF-8`, though `ISO-8859-1` may sometimes work.
- `Strict ISO-8859-1` - Ignore case only, and assume **Storage Charset** is `ISO-8859-1`. For back-compatibility. Available only for **Text Search Mode**.
- `Exact` - Match characters exactly, respecting case, diacritics, width etc. Available only for **Attribute Compare Mode**.
- `Custom ->` - Use the custom mode entered in the **Custom Mode** box. This is a comma-separated list composed from the following tokens; consult Thunderstone tech support for advice:
  - `iso-8859-1` - Assume text is ISO-8859-1 encoded. Should only be used if **Storage Charset** is also `ISO-8859-1`. If this flag is not set, text is assumed to be UTF-8, though occasional ISO-8859-1 characters will usually be able to match their UTF-8 equivalents.
  - `ignorediacritics` - Ignore diacritic marks (accents, umlauts, etc.). E.g. "`e`" will match "`é`" (U+00E9) and vice-versa.
  - `expandligatures` - Expand ligature characters. E.g. "`oe`" will match "œ" (U+0153) and vice-versa. Note that with this flag off, certain ligatures may still be expanded if necessary for case-folding with `ignorecase`.
  - `ignorewidth` - Ignore half-/full-width differences, e.g. for ASCII and katakana characters.
  - `ignorecase` - Ignore case differences, e.g. "`e`" matches "`E`" and vice-versa; this is the default. The alternative is `respectcase`.
  - `respectcase` - Case-sensitive search, e.g. "`e`" does *not* match "`E`". The alternative is `ignorecase`.
  - `unicodemulti` - Use Unicode case-compare tables, with multi-character expansions where needed (e.g. for ligatures). The alternative is `ctype` or `unicodemono`.

- `unicodemono` - Use Unicode case-compare tables, but do not expand characters. The alternative is `ctype` or `unicodemulti`.
- `ctype` - Use the operating system's `ctype.h` case-compare tables. Only codepoints U+0001 through U+00FF (i.e. single-byte or ISO-8859-1 range) are supported, though the actual encoding may be ISO-8859-1 or UTF-8 depending on the `iso-8859-1` flag. The alternative is `unicodemulti` or `unicodemono`.

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

3.5.54 Attribute Compare Mode

Syntax: select from options or enter custom mode

Sets the character-matching mode for attribute comparison searches, e.g. equals, less-than, order-by, `IN`. This controls aspects like case-sensitivity, ignoring accents, etc. See **Text Search Mode** (p. 75) for details on what the setting values mean. The default is `Exact`. Note that searches on `Enum` fields are unaffected by this setting, as the `Enum` type is defined to be case-insensitive.

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

3.5.55 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. **Additional Fields** may be selected, if they are to be searched by the user's text query. Note that changing **Additional Fields**' names, types or order later may affect their presence in **Index Fields**.

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.

3.5.56 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. Note the following caveats:

Adding a field to **Compound Index Fields** will not help search performance if there is no main (text) query also, as the compound part of the index can only be used in conjunction with a text query.

Only a fixed-size amount of data can be stored in each row of each of the **Compound Index Fields**, so only fixed-size fields such as dates, integers, numbers, etc. should be chosen. If text data is used, all values for the field in the database should be small (a few characters) for best performance.

Additional Fields `Text` fields are not permitted in **Compound Index Fields**.

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.

3.5.57 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 walks or other database modifications. Non-`Metamorph`/`Metamorph Inverted` indexes should therefore be created *before* the profile is walked or populated with data to avoid this issue, if possible. **Extra Indexes** should only be created when the profile is not actively walking, to minimize load and potential walk impediments.

3.5.58 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.

3.5.59 Primer Type

“Primer URLs” are URLs that are fetched before actually starting a walk. They are not stored in the search database, but instead are used to “prime” the Search Appliance with any necessary credentials (e.g. 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 walked 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 walking.

  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` *(default)* - The URLs listed in **Custom Primer URLs** (if any) are used, as described below.

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

3.5.60 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 (e.g. 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`". E.g.:

```
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.

Note: After version 8.0.4 (2012-07-13) the Search Appliance will set the HTTP(S) `Referer` header for each primer URL to the URL of the previously used primer. So authentication systems that require `Referer`s will work. If the first primer URL also requires a `Referer` add a primer URL before that so it picks up that as the `Referer`. This does not affect the use of `Referer` in the main walk.

Checking for Bad Logins: Bad Login MM Query

Sometimes, the primer URL login may fail, e.g. bad login. However, since the only error indication may be a "Login failure"-type message and not a true HTTP error code, the Search Appliance 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 walk 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 walked. The associated primer URL will only be fetched if it matches.

Following additional links with the !FOLLOW_LINK token

The primer system automatically follows the HTTP redirects `301 Moved Permanently` and `302 Found`. Sometimes login systems produce additional links that must be followed to get the login

cookie, but aren't true HTTP redirects. Examples could be javascript that sets `document.location` or a page that simply says "click here to continue".

The special Custom Primer Url token `!FOLLOW_LINK` can be used instead of a url to follow the first link generated by the previous primer's fetch. This can be added multiple times to follow multiple links.

3.5.61 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 backslash (\), i.e. `MY_DOMAIN\myuser`.

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

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

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

3.5.62 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.

3.5.63 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.

3.5.64 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 the Search Appliance, so it can inherit the browser's state. This is not recommended with the Search Appliance as it would require exporting a filesystem with your cookie file and mounting it to the Search Appliance. Then every walk would be dependent on that machine being online and the cookie file being present.

3.5.65 Cookie Jar

Netscape or Microsoft Internet Explorer format cookie data to load at start up. An inline alternative to `Cookie Source Path`. This allows persistent cookies saved by a browser to be read by the Search Appliance, so it can inherit the browser's state. To easily walk a site that requires a custom login (i.e. not HTTP Basic authentication), and that uses persistent cookies, just login normally using a browser. Then, copy that browser's cookie data into the `Cookie Jar` setting. Then, the Search Appliance will automatically inherit the browser's permissions. Added after version 8.0.4 2012-07-13.

3.5.66 Strict Cookie Paths

If set to `Y`, the `Path` attribute (if present) of any received cookie must be a prefix of the URL setting it, or the cookie will be discarded, as per RFC 2965 3.3.2. This helps prevent one application from altering the cookie(s) of another application on the same server; such isolation may be desired if the applications should be protected from each other. However, typically such cross-path altering is acceptable – e.g. some login systems depend on it – so this setting defaults to `N`, which also aligns with typical browser behavior. Only available with products using Texis version 6.00.1342215000 201210713 and later; earlier versions effectively behave as if this setting were always `Y`.

3.5.67 Off-Site Pages

Syntax: select Yes or No button

Allow retrieval of individual off-site pages. By default the Search Appliance 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.

All other discarding rules don't involve the site except for Stay Under still apply (Extensions, Exclude Prefix, Required Rex, etc).

3.5.68 Off-Site Components

Syntax: select Yes or No button

This option also allows off-site resources embedded within on-site pages to be fetched for processing. This includes javascript sources, embedded frames, and redirects.

3.5.69 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.

3.5.70 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.

3.5.71 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 (e.g. `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 (i.e. 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.

3.5.72 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.

3.5.73 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.

3.5.74 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.

3.5.75 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.

3.5.76 Fetch JavaScript

Syntax: select Yes or No button

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

3.5.77 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.

3.5.78 Debug JavaScript

Syntax: select Yes or No button

Print additional debugging messages for JavaScript errors.

3.5.79 JavaScript Memory

Syntax: numeric memory size e.g. `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.

3.5.80 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, e.g. if "`JavaScript exceeded scripttimeout`" messages are received.

3.5.81 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`, `gopher` and `file`.

3.5.82 HTTP Version

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

3.5.83 SSL Client Protocols

Which SSL protocols to allow for client HTTPS/SSL connections when walking and searching, i.e. for connections from the Search Appliance 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.

Note: To change the *server*-side SSL protocols accepted by the Search Appliance, see **HTTPS/SSL Protocols** under System Wide Settings.

3.5.84 Authentication Schemes

Select which authentication schemes to allow for password-protected URLs. The settable schemes are `Basic`, `File` (for `file://` URLs), `NTLMv1` and `NTLMv2`. NTLMv2 requires Texis version 5.01.1213917000 20080619 or later. Note that the scheme(s) actually *accepted* for a given URL are determined by the server; if none of the server-offered schemes are enabled by this setting, then the

protected URL cannot be walked. This setting can be used to disable less-secure or undesired schemes, such as Basic or NTLMv1 authentication.

3.5.85 Embedded Security

Select the security for embedded objects on a page (e.g. 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, i.e. 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.

3.5.86 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.

3.5.87 Multiple Fetches

Syntax: select Y or N

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

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

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

3.5.88 Follow Cross-Site Links

Syntax: select Y or N

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

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

3.5.89 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.

3.5.90 Empty Form Redirects

Syntax: select Y or N

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

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

3.5.91 Execute Walked Dataload

Syntax: select Y or N

The Dataload system (section 4.19.3, p. 154) allows administrators to load arbitrary content into the Search Appliance by POSTing XML files.

If `Execute Walked Dataload` is set to `Y`, then valid dataload XML files that are encountered during the crawl will be fully processed as if they were uploaded to the appliance.

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

3.5.92 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/`).

3.5.93 DNS Mode

Syntax: choose from drop down list

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

3.5.94 User Agent

Syntax: full user-agent string

Set the User-Agent (browser type) to report to web servers. Normally the Search Appliance 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).

3.5.95 Mime Types

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

These are the Multipurpose Internet Mail Extensions (MIME) types that the Search Appliance 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 (`*/*`).

3.5.96 Custom Headers

Syntax: one or more Name/Value pairs

These are extra headers to send to the webserver when fetching pages. Sometimes a server application will not work right when optional HTTP headers are not present. This allows setting them as needed in such situations.

Header names should not have any spaces or the trailing colon (:).

3.5.97 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".

3.5.98 Cache Content

Syntax: choose from drop down list

Cache Content allows the Appliance to store a copy of the content as it walks. In the search interface, a `View Cached` link will appear by results, which allows users to download the contents directly from the Appliance rather than using to the original location.

This can be useful in situations like providing results when the original location becomes unavailable, or allowing internet users to download `file://` results from a file server when users aren't on the local

network.

The choices for Cache Content are:

- `None` *(default)* - No extra storage of the original content occurs, the Appliance only maintains the plain text representation of the results for searching (as it always has).
- `Results Only` - While walking, the Appliance maintains a copy of the original content. Selecting `View Cached` from a search result will present that cached copy. Links and references from web pages are left unmodified, pointing back to the original site.
- `Site Mirror` - While walking, the Appliance maintains a copy of the original content, and also collects and stores auxiliary content used by web pages (images, css stylesheets, etc) that isn't normally necessary for the Appliance walks.

  When search users select `View Cached`, links and references on cached web pages are rewritten to point back to the Appliance, utilizing the cached versions of these resources. This allows a full cached view of a site to operate when the original site becomes unavailable, instead of presenting a cached web page with lots of broken images and links.

  Note that some complex, dynamic web sites might not be able to be fully mirrored.

Note: Cache Content is a completely separate feature from the similarly named Results Caching.

- Results Caching saves an entire page of results after a search is performed, so additional searches for the same terms can use that cache instead of performing the search again. This is invisible to search users.
- This feature, Cache Content, saves additional information while walking so that users can download an original copy of the result directly from the Appliance instead of the original location. This is done through a `View Cached` link on the search interface.

Please see 3.6.101 for more information on Results Caching.

3.5.99 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.

3.5.100 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 (i.e. constantly refreshing it and loading the web server). Typically set to 1 minute.

3.5.101 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.

3.5.102 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. Note that this is a soft, not hard limit: the process may exceed it briefly, yet will exit gracefully and not crash if so.

3.5.103 Maximum Load Average

Syntax: number, or `-1` for no limit

`Maximum Load Average` helps prevent the Search Appliance from becoming overloaded. If the load goes above the specified `Maximum Load Average`, the walk will pause itself until the load decreases.

You can see the current load of the appliance at any time by going to `System -¿ System Information`, next to `Load`. The 3 numbers represent the average load during the last 1, 5, and 15 minutes.

3.5.104 Replication Settings

Syntax: List of hosts and profiles

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

3.5.105 Send Data

Syntax: select Y or N

Send Data will cause this profile to send changes in searchable content to be sent to this replication target. This includes adding or changing content, and refresh walks removing URLs that are no longer present.

This does not include starting a new database for a “New” walk. If URLs are present in a previous walk and not present in the next “New” walk, this will not remove the URLs from the replication target. This requires `Send Settings`, as described below.

3.5.106 Send Settings

Syntax: select Y or N

Send Settings causes this profile to send any settings changes to this replication target. This includes starting "New" walks, which will cause the replication target to start from a fresh database like the replication sender does.

3.5.107 Batch Rows

Syntax: number

Defines the number of items the replication sender will attempt to accumulate in a single batch. This should not need changed.

3.5.108 Batch Size

Syntax: number (in bytes)

Defines the threshold for size for a sending a batch of replication items. When collecting items to replicate, once the size of the items is over this threshold, the batch will be sent off to the targets. This should not need changed.

3.5.109 Batch Idle

Syntax: number (in seconds)

Defines the idle timeout for sending a batch of replication items. When collecting a batch, if no new items appear for this many seconds, the replication batch will be sent to the targets. This should not need changed.

3.5.110 Log Replication

Writes information for this profile's replication queue processor to `replication.log`.

If both the System-wide `Log All Replication` and this profile's `Log Replication` are set, logging for this profile will be the more verbose of the two.

See also "Replication" 4.19.

3.6 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.

See also "Customizing the Search Appliance's Appearance" 2.2.

3.6.1 Notes

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

3.6.2 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.

3.6.3 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 3.5.3).

3.6.4 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.

3.6.5 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.

3.6.6 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.

3.6.7 Allow RSS

Syntax: select Yes or No button

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

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

3.6.8 Format XSL Output

Syntax: select Yes or No button

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

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

3.6.9 XSL Engine

Syntax: Choose `Internal` *(recommended)* or `Legacy`

the Search Appliance recently changed its XSL processing from an external tool to an internal library. All processing should use the new `Internal` engine.

This switch shouldn't cause any issues, but in case there are quirks, the `XSL Engine` setting allows profiles to change back to the old `Legacy` engine until any issues with the `Internal` engine are ironed out.

This feature will be deprecated in the future, and the `Legacy` engine completely removed.

3.6.10 XSL File

Syntax: Browse local disk for a XSL file

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

3.6.11 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.

3.6.12 Abstract Length

Syntax: enter number in text box

This determines the length in bytes of the document abstract.

3.6.13 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.

3.6.14 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.

3.6.15 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.

3.6.16 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.

3.6.17 Page Links Shown

Syntax: a whole number, defaults to `10`

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

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

3.6.18 Results per Site

Syntax: an integer select box and Yes/No button

The **Max** setting controls the maximum results per site per page to display. For large profiles with many pages (and thus results) per site, setting Results per Site can increase the results variety shown on a single page, by replacing some same-site results with lesser-ranked but different-site results from subsequent pages.

When results are limited to N per site, no more than N results for any given site will be shown on any given page. Results past N for a site are suppressed, and results from new sites (that are not past N yet) added, until the page is full. Once the page is complete, the results are reordered: the second and later results from a site are moved up under the first result from that site, indented, and followed by a `More results for site` link. The next page's results will be obtained by the same process, resuming at the point in raw results left off by the previous page. Note that a given site may appear on subsequent pages (if there are more results for it), but its results that were suppressed for the previous page will not be shown (because they are visible via the previous page's `More results for site` link). Also, since there are now two degrees of navigation through the results – standard pagination, plus the `More results for site` links – not all of the total results counted may be shown via pagination: the rest are shown via `More results for site`.

Note also that setting Results per Site to other than `Unlimited` can increase search time, as potentially much more than one page of raw results must be obtained and suppressed, and results must be regrouped.

The **Allow override** button controls whether the search user can override the profile's Results per Site limit on the Advanced Search form. This can be set to `N` to prevent the potential delay of grouping by site, or `Y` to allow the user to set a custom value. For profiles that are Meta Search back-ends, if the front-end Meta Search is using Results per Site, all the back-ends should have **Allow override** set to `Y` so that the front-end's value can be used, for consistency.

3.6.19 Allow site: syntax

Syntax: a Yes/No button

This controls whether to allow the `site:host` query syntax in a search, to limit results to a single domain. It has no effect on the **From this domain** box on the Advanced Search form, which uses a separate variable (`sq`) instead of embedding in the query. For profiles that are Meta Search back-ends, if the front-end Meta Search is using Results per Site, all the back-ends should have **Allow site: syntax** set to `Y` so that the front-end's value can be passed via the `site:host` syntax.

3.6.20 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.

3.6.21 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.

3.6.22 Show File Icons

Syntax: select Yes or No button

Show file type icons on individual results. These icons appear for non-html types (office files, PDF, images, etc). The exact positioning or usage of the icons can be customized using the XSL Stylesheet (see `XSL File` setting 3.6.10).

3.6.23 Show Thunderstone logo on results

Syntax: select Yes or No button

This controls the display of the Thunderstone logo on the search results page. The logo is displayed on the first and last page of a search.

3.6.24 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.

3.6.25 Query Autocomplete

Syntax: select Yes or No button

Query Autocomplete (also known as "typeahead") lets the Search Appliance suggest full words from the user's partially typed queries as they're being typed. If the user has typed `ense`, query autocomplete may suggest `ensemble`, `enseignement`, etc.

Like Spell Check, the list of words used for Query Autocomplete comes from the current profile's content, so the only suggestions made will be words that occur somewhere in the content.

3.6.26 Max Completions

Syntax: a whole number

This controls the maximum number of completions that will be shown to the user when using Query Autocomplete functionality.

3.6.27 Results Highlighting

Syntax: select None, Classes, Inline or Bold

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

- `None` or `N` - No highlighting will be done in search results.
- `Classes` or `Y` - Terms will be highlighted with `<span>` tags that refer to classes that are defined in a separate CSS file, `/common/search.css` by default. Each term in the query is tagged with a different class, which are each highlighted a different color in the default `search.css`. All these rules can be overridden with your own CSS on the results page.
- `Inline` - Terms will be highlighted with `<span>` tags that directly specify a fixed CSS style. This is not customizable, but is self-contained and does not depend on a separate stylesheet or file. Same visual result as `Classes` with the default CSS.
- `Bold` - Terms will be highlighted with `<b>` tags.

The default is `Bold`.

3.6.28 Context Highlighting

Syntax: select None, Classes, Inline or Bold

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

3.6.29 PDF Query Highlighting

Syntax: select Yes or No button

When making links to PDFs in search results, the Search Appliance 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.

3.6.30 PDF Highlighting Format

Syntax: select "Acrobat 7+" or "Legacy"

Controls the format used to provide query highlighting information for PDF links.

- `Acrobat 7+`: Uses the `search` parameter as defined in the PDF Open Parameter specification. Supported by Adobe Acrobat 7 (January 2005) and all later versions.
- `Legacy`: Uses the `xml` Highlight File Format syntax, which has been deprecated by Adobe. It was disabled by default in Acrobat Reader 9 (July 2008), and completely removed from Acrobat Reader X (Nov 2010) and all later versions.

This is a temporary setting is for compatibility, and will be removed in a future version of the Search Appliance (forcing `Acrobat 7+` behavior).

3.6.31 Font

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

This specifies the font to use throughout the search interface.

3.6.32 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 the Search Appliance 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).

3.6.33 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.

The `<!-- THUNDERSTONE_HEADERS -->` placeholder is replaced at search time with custom information necessary for search. If you customize your Top HTML, make sure `<!-- THUNDERSTONE_HEADERS -->` is somewhere in the Top HTML's `<head>` section.

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

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

CSS Stylesheet

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

3.6.34 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.

3.6.35 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` 4.16 for more details.

3.6.36 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` 4.16 for more details.

3.6.37 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` 4.16 for more details.

3.6.38 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` 4.16 for more details.

3.6.39 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` 4.16 for more details.

3.6.40 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` 4.16 for more details.

3.6.41 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` 4.16 for more details.

3.6.42 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` 4.16 for more details.

3.6.43 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` 4.16 for more details.

3.6.44 Authorization Method

The `Authorization Method` setting controls what Results Authorization method(s) are used by the Search Appliance when verifying user access to search result URLs. See the Results Authorization section (p. 124) for details. The possible settings are:

- `None`: No access verification; return all search results to all users. This is the default. It is also the setting that should be used for a Meta Search profile, even if one or more of its back-end profiles does use Results Authorization: the request and response for credentials will automatically be passed back and forth from front-end Meta Search to back-end profiles, which will handle the authorization (not the front-end).
- `Forward login cookies`: The Search Appliance will forward login cookies from the user to the result URL. This is for custom HTML-form-based single-sign-on systems.
- `Basic/NTLM/file - prompt via form`: The Search Appliance will prompt the user for their credentials with a form, then send them to the result URL via HTTP Basic, NTLM or Windows/SMB file authentication.

3.6.45 Login Cookies

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

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

3.6.46 Login URL

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

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

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

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

3.6.47 Basic/NTLM/file Cookie Type

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

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

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

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

3.6.48 Login Verification URL

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

To differentiate the two and give users a chance to correct mistyped passwords, a `Login Verification URL` may be set. This should be a URL that *all* users have access to, but that is still protected (i.e. anonymous users are denied). It should be an actual file (not a directory), preferably small (a few KB), and permanent (not likely to move, be renamed or have perms changed).

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

3.6.49 Unauthorized Result Query

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

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

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

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

3.6.50 Username Fixup

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

- `Search` - the search expression to match on the incoming username. Unless you're stripping off a domain, this should be left blank to match everything.

- `Replace` - the replacement string used to modify what was matched in the search. Please see examples below, or the **Replacement Strings** of the Vortex manual on our website for the exact syntax.

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

Examples

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

3.6.51 Max Docs to Auth-Check

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

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

3.6.52 Successful Auth Result Limit

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

The minimum (and default if empty) is the same as the `Results per Page` setting (p. 94), which produces a page of results the fastest. The maximum is -1 for no limit, i.e. continue until all results are checked, or `Max Docs to Auth-Check` or `Total Auth Timeout` is reached.

3.6.53 Total Auth Timeout

This setting is the maximum total time in seconds to spend searching and authorizing results, during Results Authorization. The maximum setting value is -1 for no limit, i.e. let `Search Timeout` (p. 110) cancel the search if reached. Any other negative value is relative to `Search Timeout`. Thus the default (if empty) of -5 means stop searching 5 seconds before `Search Timeout`, so that there are a few seconds left to send the results to the user.

3.6.54 Allow Authorization URL

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

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

For this optimization to be effective, the `Authorization URL` field in the database must be populated (see **Data from Field**, p. 64). For example, on systems where the contents of a directory always have the same authorization as the directory itself, `Authorization URL` should be set to the parent dir of each URL. The more files there are (on average) in a given directory, the more effective this optimization will be. Additionally, the **Authorization Caching** setting should be set to `Session`, so that the one-time directory authorization can be reused for each result inside the directory. (Otherwise Results Authorization must repeat the directory authorization for every result in the directory, as normal.)

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

3.6.55 Authorization Caching

Whether and how to cache Results Authorization traffic. The default of `None` does no caching. When set to `Session`, Results Authorization traffic is cached for the duration of the session, i.e. that search alone. Normally caching is of little benefit, because authorization URLs are typically the same as result URLs, and the latter are typically unique in a given search; thus caching will not help. However, if the `Authorization URL` field is populated, and **Allow Authorization URL** is enabled, enabling caching may speed up Results Authorization searches. See **Allow Authorization URL** (p. 104) for details.

3.6.56 Debug Results Authorization

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

3.6.57 Show Authorization Info

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

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

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

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

3.6.58 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, the Search Appliance 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.

3.6.59 Suggest Time Limit

Syntax: choose from drop-down list

This controls the number of seconds the Search Appliance allows for spelling suggestions to be made. See also `Enable Spell Check` 3.6.58 for more information.

3.6.60 Number of Suggestions

Syntax: choose from drop-down list

This controls the number of spelling suggestions offered. See also `Enable Spell Check` 3.6.58 for more information.

3.6.61 Synonyms

Syntax: choose from drop-down list

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

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

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

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

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

See also `Using the Thesaurus` (section 4.3).

3.6.62 Main Thesaurus

Syntax: the symbolic name for the primary thesaurus

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

See also `Using the Thesaurus` (section 4.3).

3.6.63 Secondary Thesaurus

Syntax: the symbolic name for the secondary thesaurus

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

See also `Using the Thesaurus` (section 4.3).

3.6.64 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.

The Search Appliance 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 the Search Appliance's use of set logic, please see the **Using Set Logic to Weight Search Items** of the Texis manual on our website.

3.6.65 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 (e.g. `@0`) may be slower, as they can generate a copious number of hits.

3.6.66 Allow Linear

Syntax: select Yes or No button

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

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

3.6.67 Allow NOT Logic

Syntax: select Yes or No button

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

3.6.68 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, e.g. 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.

3.6.69 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.

3.6.70 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*`".

3.6.71 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. E.g. the query "`con*tion`" will match "`condition`" but not "`consider my position`" nor "`conditionally`".

3.6.72 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.

3.6.73 Require All Words

Syntax: select Yes or No button

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

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

3.6.74 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.

3.6.75 Phrase Word Processing

Syntax: select box

This setting controls how suffix/wildcard processing – as determined by the **Word Forms** setting – is applied to phrases. Single-word terms always have suffix/wildcard processing applied; however phrases – multi-word terms bound by hyphens or in double-quotes – are only processed if this setting allows it.

There are two choices:

- `Last word only`: Only process the last word in the phrase; this is the default. For example, with this value set (and **Word Forms** set to `Any word forms`) the query "`"vacuum cleaner"`" would match "`vacuum cleaner`" as well as "`vacuum cleaners`".
- `None`: Do no processing on phrase words. With this value set, the query "`"vacuum cleaner"`" would only match "`vacuum cleaner`", regardless of **Word Forms**.

Note that a *single* word term is not a phrase – even if double-quoted – and thus **Phrase Word Processing** does not apply to it.

3.6.76 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.

3.6.77 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

3.6.78 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.

3.6.79 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 (e.g. "`Your query was all noise words.`") visibly (in gray 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.

3.6.80 Debug SQL Level

Syntax: integer/hex 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.

3.6.81 Debug Metamorph Level

Syntax: integer/hex number or empty/0 to disable

This enables extra debug messages for certain Metamorph statements. Generally only set at the request of tech support for diagnosing problems. Note that this setting can generate copious amounts of messages.

3.6.82 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).

3.6.83 Proximity

Syntax: choose from drop-down list

Proximity gives the ability to locate answers with greater precision. The Search Appliance 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 (e.g. all searches except `Show Parents`).

3.6.84 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`, i.e. 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. 74) should usually also be listed in **Language Characters**.

3.6.85 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 match: Only exact matches are allowed. (the default)

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

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

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

3.6.86 Custom Suffix List

Syntax: Space-separated list of suffixes

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

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

3.6.87 Custom Suffix Default Removal

Syntax: Y or N

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

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

Default Removal is set to `Y` when using `Any Word Forms`, but not with `Plurals & Possessives`.

3.6.88 Custom Suffix Min Length

Syntax: Number

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

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

3.6.89 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).

3.6.90 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.

3.6.91 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.

3.6.92 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.

3.6.93 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.

3.6.94 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, i.e. do not factor in clicks-from-home for results ranking.

3.6.95 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.

3.6.96 XML Export Variables

Syntax: names separated by newlines

`XML Export Variables` is a list of variables, one per line, that are to be displayed and propagated through XML search results. This can be used to propagate HTTP headers from the client's request, or for including arbitrary extra information.

- `HTTP Headers`

  You can specify the header name in all caps, with underscores for spaces, and with a `HTTP_` prefix. For example, to include "User Agent", specify `HTTP_USER_AGENT`.

- `Arbitrary variables`

  You can specify any named variable, and values passed in to the search query will be propagated in the XML output.

For example, if you use `cbtGroup` and `HTTP_USER_AGENT`, and the search Url includes `...&cbtGroup=user&cbtGroup=backup&...`, then the following block will appear in the XML output:

```
<exportVar>
  <variable name="HTTP_USER_AGENT">Mozilla/5.0 ....</variable>
  <variable name="cbtGroup">user</variable>
  <variable name="cbtGroup">backup</variable>
</exportVar>
```

3.6.97 File URL Format

Syntax: choose from drop down list

Controls how file URLs are formatted. The Max Compatibility setting will format them as `file:////server/share` which both Internet Explorer and Mozilla-based browsers such as Firefox and SeaMonkey support, while the Internet Explorer only setting will format them as `file://server/share`.

Note that for Mozilla-based browsers you will also need to enable permission for HTML pages to open files by creating a `user.js` in your profiles directory (where `prefs.js` is) that contains (note: lines wrapped to fit the printed page):

```
user_pref("capability.policy.policynames", "localfilelinks");
user_pref("capability.policy.localfilelinks.sites",
          "http://HOSTNAME");
user_pref("capability.policy.localfilelinks.checkloaduri.enabled",
          "allAccess");
```

where `HOSTNAME` is the name or IP address of the search results as seen in the browser address bar.

3.6.98 Redirect Format

Syntax: choose from drop down list

Controls how redirects are presented when doing Query Logging, as some combination of HTTP 302 redirects and HTML meta-refresh redirects.

The `Max Compatibility` setting is recommended, which works around known bugs in Internet Explorer by not using a 302 redirect when Internet Explorer is requesting file:// URLs.

3.6.99 Phishing Protection

Phishing Protection prevents the Search Appliance from being used as a tool in a phishing attack.

The Search Appliance 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 the Search Appliance, 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.

3.6.100 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"`.

3.6.101 Results Caching

The **Results Caching** option can be used to enable a search results cache mechanism, which can speed up response time for frequently-used queries. When enabled, the HTML or XML results of searches are stored

in a per-profile cache, and if a later search is made with the same query string parameters, the cache result may be returned, avoiding the need for a full search. Setting **Enabled** to `Y` enables results caching.

Setting **Allow Override** to `Y` allows users to control whether a search uses and/or updates the cache, by specifying a value for the `rescache` query string variable:

- `n` - Do not search or refresh the cache; i.e. search as if caching were disabled.
- `refresh` - Do not search the cache, but update it if rules allow.
- `norefresh` - Search the cache, but do not update it.
- `y` - Search and update the cache normally.

Both **Enabled** and **Allow Override** are considered "appearance" options, i.e. they will not take effect for Live Search unless `Update Live and Test` is used. See the Results Caching section of Profile Tools (p. 37) for more details on when to use results caching, its caveats and management.

Note: Results Caching is a completely separate feature from the similarly named Cache Content.

- Cache Content saves additional information while walking so that users can download an original copy of the result directly from the Appliance instead of the original location. This is done through a `View Cached` link on the search interface.
- This feature, Results Caching, saves an entire page of results after a search is performed, so additional searches for the same terms can use that cache instead of performing the search again. This is invisible to search users.

Please see 3.5.98 for more information on Cache Content.

3.6.102 Max Cache Entry Age

This sets the maximum age, in seconds, for a results cache entry. Entries older than this will not be used for results, and will be purged by the results cache manager. The default (if empty) is 21600 seconds, i.e. six hours.

3.6.103 Max Cache Size

This sets the maximum size, in bytes, of the results cache. This is not a hard limit, but when the cache grows larger than this, the results cache manager will start to remove old/low-priority entries. The default (if empty) is 100000000 (one hundred million bytes).

3.6.104 Min Search Time

This sets the minimum search time, in seconds, that a query must take in order to be considered for results caching. Queries that are faster than this are not cached, because they are considered fast enough to save the space. The value may be an integer or floating-point (decimal) number; the default (if empty) is 2.0.

3.6.105 Visible

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

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

3.7 System Wide Settings

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

3.7.1 Admin Theme

Selects a color theme to use for the Search Appliance's administrative interface. This doesn't affect walking or the search interface in any way.

It can be useful to provide a means of quick visual differentiation between multiple Search Appliances.

3.7.2 Home Page

By default when the Search Appliance is directly accessed, as in `http://hostname_or_ip` it will present a page that allows selection of the admin or search interface. This option allows you to replace that page with any HTML you devise. The HTML you upload should refer to images and such using fully qualified URLs because they can not be uploaded to the Search Appliance for use in relative URLs.

Checking `Default` will revert the Search Appliance home page to its factory behavior.

3.7.3 Enter At Search, Default Profile

By default users accessing the Search Appliance using no particular URL will be given a choice of admin or search. Enabling this option removes that choice and enters at the search for the profile named in the Default Profile setting.

3.7.4 Favicon.ico

The Search Appliance comes with no favicon.ico file. If you wish users' browsers to display your company's favicon when they are accessing the Search Appliance you'll need to upload that icon. If you no

longer wish to have a favicon check `Delete`.

3.7.5 Robots.txt

The Search Appliance comes with no robots.txt file. If you want to control how web walking robots visit your search pages you can upload a robots.txt file. This robots.txt file will apply to all pages returned bythe Search Appliance including admin, documentation, and searches whether accessed by HTTP or HTTPS. See `http://www.robotstxt.org` for the syntax of robots.txt files.

If you no longer wish to have a robots.txt check `Delete`.

Note: You may also control robots for searches on a profile by profile basis by using meta robots within the top HTML of the search settings or the custom XSL if you're using an XSL style sheet for search results. See `http://www.robotstxt.org` for the syntax of the meta robots tag.

3.7.6 Cluster Members

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

3.7.7 API Logging

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

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

3.7.8 Task Monitor Logging

Controls the verbosity of logging for the Task Monitor. Messages are logged to `taskmonitor.log` in the logs directory.

3.7.9 Google Connector Logging

Controls the verbosity of logging for the processing of Google Connectors. Messages are logged to `gcon.log` in the logs directory. This includes Google Feed pushes, and full connector actions.

3.7.10 Disable All Walks

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

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

3.7.11 Update Software

This allows you to configure the system to perform updates automatically. There are three steps to performing an upgrade, and you can select how many steps are performed automatically. The steps are described as follows:

- `Don't check for updates automatically` - the Search Appliance will not perform any actions on its own.
- `Check for updates but don't download or install` - the Search Appliance will discovers whether software that is newer than what is installed is available.
- `Download updates but don't install` - the Search Appliance will download software updates and hold them for you to peruse and install.
- `Install updates automatically` - the Search Appliance automatically applies software updates as they're made available.

3.7.12 HTTP Proxy Server

If you need a HTTP proxy to access the internet while checking for updates, you can specify the server here.

If you don't need a proxy, leave blank.

3.7.13 Proxy Username

If the HTTP proxy used when checking for updates requires credentials, enter your username here.

If you're not using a proxy or it doesn't require credentials, leave blank.

3.7.14 Proxy Password

If the HTTP proxy used when checking for updates requires credentials, enter your password here.

If you're not using a proxy or it doesn't require credentials, leave blank.

3.7.15 System Replication Settings

"Targets" defines machines that will receive system replication from the Appliance. This includes the creation and deletion of profiles, all profile settings changes, and all profile data.

System replication targets must have `Allow Receiving` enabled, as described below.

3.7.16 Allow Receiving

the Appliance can only be a receiver of System Replication (described above) if this Appliance has Allow Receiving set to `Y`. The sender must also be listed in this Appliance's Cluster Members, as described above.

3.7.17 Log All Replication

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

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

3.7.18 Enable HTTPS Server

This enables the Search Appliance's HTTPS web server, allowing web based admin and searches to be accessed via HTTPS in addition to or instead of HTTP. Turn this on to enable encrypted communications. Then access the admin interface using `https` in the URL instead of `http`. (Note: this setting was formerly

called **Enable HTTPS Admin**.)

3.7.19 Require HTTPS for Direct Admin

Set this option to `Y` so that direct access to the admin interface is only permitted via HTTPS and not HTTP. "Direct" means the immediate, direct connection to the Search Appliance from the web browser (or proxy, if any); security of the earlier hops of proxy-forwarded connections is checked by **Require HTTPS for Proxy Admin**.

If you use this you must also turn on **Enable HTTPS Server**.

For safety, `Require HTTPS for Direct Admin` can only be enabled while accessing the admin interface via HTTPS.

If you have set this option `Y` and accidentally configure it such that you can not access the Search Appliance, you can re-enable HTTP admin by going to the physical console of the Search Appliance and selecting the `drop Admin restrictions (HTTPS,IP,Cipher requirements)` option.

(Note: this setting was formerly called **Require HTTPS Admin**.)

3.7.20 Require HTTPS for Proxy Admin

Set this option to `Y` so that proxy-forwarded access to the admin interface is only permitted via HTTPS and not HTTP. Forwarded connections are those hop(s) in the connection chain that are forwarded from the client to a proxy (that then accesses the Search Appliance directly); for control of direct connections to the Search Appliance admin (or the direct last-hop from a proxy to the Search Appliance), see **Require HTTPS for Direct Admin**.

Forwarded connections are checked by examining the `X-Forward-Proto` header value of connections to the admin interface: if all tokens are `https`, the forwarded connection is considered secure/HTTPS, otherwise insecure/HTTP. If no `X-Forwarded-Proto` header is present, the connection is not considered forwarded and this setting does not apply. Note that for this setting to be effective, the network must be secured such that *all* devices with direct access to the Search Appliance can be trusted to set (or clear) the `X-Forwarded-Proto` header properly, as the header is easily forged.

For safety, **Require HTTPS for Proxy Admin** cannot be enabled if you're currently accessing the Search Appliance via an insecure proxies.

If you have set this option `Y` and accidentally configure it such that you can not access the Search Appliance, you can re-enable HTTP admin by going to the physical console of the Search Appliance and selecting the `drop Admin restrictions (HTTPS,IP,Cipher requirements)` option.

3.7.21 Admin Access IPs

This controls what IP addresses are allowed to access the admin interface. You may specify one or more individual IP addresses or networks. Networks may be specified with either address:netmask or address/prefixlen syntaxes. Place each entry on a line by itself. Blank means no IP restriction, the admin interface may be accessed from any IP.

Example. If you have a local class C network of 10.10.1.0 as well as one public IP such as 198.49.220.1 you want to have admin access you would use

```
10.10.1.0/24
198.49.220.1
```

or

```
10.10.1.0:255.255.255.0
198.49.220.1
```

For safety, The Search Appliance will only accept `Admin Access IPs` that allow the IP address it's currently being accessed from.

3.7.22 HTTPS/SSL Protocols

Which protocols to allow for HTTPS/SSL connections to the Search Appliance server from remote clients. Note: `webmin` will only use the first one checked unless all are checked.

For *client* SSL protocols during walks, see **SSL Client Protocols** under All Walk Settings.

3.7.23 HTTPS/SSL Ciphers

Which ciphers to allow in HTTPS/SSL connections to the Search Appliance server from remote clients. This uses the `openssl ciphers` command-line format. e.g. turn off low security ciphers with `DEFAULT:!LOW`.

3.7.24 Enable SNMP service

This enables the SNMP server on the Search Appliance. With this enabled you can use SNMP monitoring tools to monitor the condition of the Search Appliance.

A few items of particular interest might be

What	OID (Object Identifier)
Disk space	.1.3.6.1.4.1.2021.9
System load	.1.3.6.1.4.1.2021.10
Critical processes	.1.3.6.1.4.1.2021.2

3.7.25 SNMP Community Name

This is the community name used to access the SNMP information. We suggest using something unique to your organization rather than ”public”.

3.7.26 SNMP Location Value

This is pretty much anything you want. It has no significance except as a designator for you to identify where or what the Search Appliance is.

3.7.27 SNMP Contact Value

This is pretty much anything you want. It would normally contain some contact information for the admins of the Search Appliance.

3.7.28 SNMP Access IPs

This controls what IP addresses are allowed to access the SNMP interface. You may specify one or more individual IP addresses or networks. Networks may be specified with either address:netmask or address/prefixlen syntaxes. Place each entry on a line by itself. Blank means no IP restriction, the SNMP interface may be accessed from any IP.

Example. If you have a local class C network of 10.10.1.0 as well as one public IP such as 198.49.220.1 you want to have SNMP access you would use

```
10.10.1.0/24
198.49.220.1
```

or

```
10.10.1.0:255.255.255.0
198.49.220.1
```

3.7.29 Administration Interface Options

These settings allow you to tweak the appearance of the pages when you’re working with the Administration Interface. Note that this is the title of the Search Appliance *administration* pages, such as

`All Walk Settings`, `List/Edit URLs`, `System`,etc., not the titles of the actual *search* pages, which are controlled in each profile's `Search Settings`.

3.7.30 <title> order

By default, the title of administration pages show `$profile-$section`, where `profile` is the current profile, and `section` is the section of the profile being viewed. This setting allows you to swap those to so it shows section first.

3.7.31 <title> max profile length

If the name of the current profile is longer than this number, it will be truncated by ellipses when it displays in the title. This can ensure you will always be able to see both the profile and the section in your page titles.

`-1` *(default)* means don't truncate, and `0` removes the profile form page titles.

3.8 Results Authorization

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

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

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

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

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

3.8.1 Results Authorization Walk Settings

The Search Appliance itself needs read access to the entire set of URLs in order to build a search index. Therefore, before walking a protected data set for Results Authorization, it may be necessary to fill out the `Login Info` setting (p. 80) under `All Walk Settings` with a full-access admin type account, so that the Search Appliance can walk the data.

Or it may be necessary to fill out a `Primer URL` (p. 78) containing login info to submit to a site's login form, so that the Search Appliance can obtain the login cookies needed for access to the rest of the site.

3.8.2 Results Authorization Search Settings

After a successful walk, Results Authorization is configured with the `Results Authorization Options` group on the `Search Settings` page. The primary setting is `Authorization Method` (p. 100), which is determined by the authentication system(s) in use by the indexed URLs. If cookie-based, this is set to `Forward login cookies`; for all other systems, it is set to `Basic/NTLM/file - Prompt via form`. Most of the remaining settings depend on which method was selected; see the `Authorization Method` setting (p. 100) for details.

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

3.9 Meta Search - Search multiple profiles as one

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

3.9.1 Profile Creation

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

3.9.2 Meta Search Walk Settings

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

For each profile you want included in the search you list the profile's name in the `Profile Name` column and the host name or IP of the machine where that profile resides in the `Host IP or Name` column. You may use DNS resolvable names or IP addresses in the host column. IP addresses are slightly more efficient because they don't require DNS lookup. But names are more flexible. Only the DNS, not a bunch of profile settings, has to change when machines get replaced or renumbered.

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

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

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

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

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

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

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

`Max Backend Data Size` provides a sanity limit on the information collected from metasearch backends. You can usually leave this untouched, but it may need increased if you're getting `Max Page Size Exceeded` errors in your metasearch results. This can be caused by having an exceptionally large number of results per page (thousands), or very large parametric fields.

3.9.3 Search Settings

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

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

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

3.10 Access Control

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

3.10.1 User Groups

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

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

3.10.2 Object hierarchy

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

same privileges automatically. The access control object hierarchy in the Search Appliance is as follows:

```
/                       Global root object
   Users/                  User accounts
      admin                   admin user
      ...                     Other users
   Groups/                 User groups
   Profiles/               Profiles
      default                 default profile
      ...                     other profiles
   Settings/               Profile settings
   Maintenance/            System page
      Info/
      Updates/
      Logs/
      Settings/
         System Wide
         ACLs
         Thesaurus
         Save, Restore
         Mounts
      System/
         RAID
```

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

3.10.3 Access Control Lists

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

3.10.4 Determining Effective Rights

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

ACEs are examined in the following order[1]:

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

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

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

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

3.10.5 Required Rights for Admin Actions

Certain ACL rights are required for certain administrative actions to be performed. In order to maximize rights-configuration flexibility, some actions require rights on multiple objects. For example, editing settings on a profile requires rights not only on the profile, but also on the setting itself. Note in the object hierarchy (p. 128) that profiles and settings are two "sibling" branches, rather than settings being replicated as descendants of every profile. Thus, profiles and settings can be thought of as a two-dimensional grid for permissions, and a user's rights can be tailored across that grid: access to one setting across all profiles, access to all settings one profile only, etc.

The rights needed for specific actions are listed below. If a user does not have the required rights for an action, either a red `Access denied` message will be displayed, or (if access still granted to other parts) the affected object may simply not appear (read access denied), or may appear grayed out (write access denied). For more information and some example permission schemes, see the Using Access Control section, p. 151.

Walk and Search Settings

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

Starting and stopping a walk

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

Best Bets

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

List/Edit URLs

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

List Duplicates

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

Walk Status

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

Query Log

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

Profiles

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

Accounts

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

User Groups

Write access to `All Groups` is needed to create a new group. Write access to the group, as well as write access to each member being added or removed, is needed to add or remove members to or from a group

(except where the group is only indirectly being modified due to a member itself being deleted). Delete access to the group is needed to delete a group.

Access Control

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

Maintenance

Read access to `Info` under `Maintenance` is needed to read the Information links. Write access to `Updates` under `Maintenance` is needed to install or upgrade software. Read access to `Logs` under `Maintenance` is needed to read the Logs links. Write access is needed to `Kill` in order to kill processes. Write access to `Settings` under `Maintenance` is needed to modify Settings, *with the exception* of Enable or Disable Access Control Lists: these can *only* and *always* be performed by the `admin` user, *regardless* of ACLs (e.g. for emergency reset). **Note:** giving a user write perms on `Settings`, directly or indirectly, can allow them to override anything on the system, e.g. via externally-modified save and restore settings. Also, note that a user with physical access to the machine could overwrite settings, e.g. re-install the software.

3.11 Running the Search Interface

See section 4.1, p. 133.

Chapter 4

Procedures and Examples

4.1 Searching your Index

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

```
http://www.mysite.com/texis/search/
```

The above is a virtual path comprised of 2 parts. "`.../texis`" is the Texis Web Script interpreter and "`/search`" is the path to the search script relative to your installation's `ScriptRoot` (`/usr/local/morph3/texis/scripts`).

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.

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

```
.../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.

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

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

```
.../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 6.

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.

4.2 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.

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

```
ref=http://somesite/somepage.html
```

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

```
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:

```
/texis/search/similar.html
or
/texis/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.

4.3 Using the Thesaurus Feature

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

Here are the steps to use the thesaurus feature.

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

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

- Upload your thesaurus to the Search Appliance. At the main menu click `System` then `Custom Thesaurus` under Manage Modules.

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

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

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

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

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

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

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

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

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

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

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

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

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

- Set the `Main Thesaurus` and `Secondary Thesaurus` fields by using the following information. If you want to use only your thesaurus and not the default one, select yours for the `Main Thesaurus` option and leave verb‘Secondary Thesaurus‘ set to none. If you want the default in addition to your own, leave `Main Thesaurus` set to Built-In and set `Secondary Thesaurus` to yours. The names listed in these options are the symbolic names (`Name` field) you gave your thesauri when uploading them.
- Click `Update` to apply these settings. There is no need to check `Apply Appearance`, and these settings are applied to both `Test Search` and `Live Search`.

4.4 Getting Software Updates

You can obtain software updates manually or automatically. For information about getting them automatically, refer to `Update Software` in System Wide Settings (3.7.11).

Use the following procedure to manually obtain software updates from Thunderstone. You are able to select which updates you want, if any.

Note: if you need to use a proxy to access the internet, please see the `Software Update Settings` (3.7.11) in System Wide Settings.

- At main menu, click `System`, then `Update Software` under Update/Backup.
- A list of available updates is presented. Check the boxes for updates that you want to download from Thunderstone, and click `Yes`. The updates are downloaded to the Search Appliance, but they are not installed yet.
- The selected packages are listed for verification. Click `Yes` to install the updates.
- When the installation is completed, a message indicates that updates are completed.

4.5 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 the Search Appliance. 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 the Search Appliance it would be `ThunderstoneSA`. 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: ThunderstoneSA
Disallow: /text
Disallow: /thunderstonesa

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. The Search Appliance will be blocked from accessing any pages under `/text` and /thunderstonesa. 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. The Search Appliance 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 the Search Appliance. Any HTML may contain a meta tag in the source of the form.

```
<meta name="robots" content="WHAT-TO-DO">
```

`WHAT-TO-DO` may contain any of the following keywords. Multiple keywords may be used by placing a comma(`,`) between them.

Table 4.1: Meta-Robots Flags

Keyword	Meaning
`INDEX`	Index the text of this page
`NOINDEX`	Don’t index the text of this page
`FOLLOW`	Follow hyperlinks on this page
`NOFOLLOW`	Don’t follow hyperlinks on this page
`ALL`	Synonym for `INDEX,FOLLOW`
`NONE`	Synonym for `NOINDEX,NOFOLLOW`

Like `robots.txt` this is not enforced in any way. Robots may or may not use it. The Search Appliance always indexes and follows hyperlinks by default so it only looks for `NOINDEX` and/or `NOFOLLOW` and/or `NONE`.

4.6 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.

4.7 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.

4.8 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.

4.9 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 (3.3) from the menu.

4.10 Removing Pages from the Database

Use the `List/Edit URLs` menu (3.3) to find and delete specific URLs from the the database. You may delete individual pages or many pages at once using wildcards.

4.11 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.

4.12 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.

4.13 Integrating Search with your Site

There are four main techniques to integrate the Search Appliance with your site. The techniques are categorized as follows:

- Link to the Appliance
- Embed a search box
- Request XML search results
- Invoke the search SOAP API

The first two are simple to implement. They involve only static HTML content, and can be used in situations where no dynamic scripting is available.

The later two are more powerful and can be used in dynamically generated web sites. You'll create a "search page" for your site (search.php, search.aspx, etc) which accepts a query from the search user. Your search page makes a search query (HTTP request) to the Search Appliance, and receives result data. Your search page then displays the data in whatever manner you wish to the search user.

The advantage of the later two methods is your site controls all interaction with the search user, making it easy for your search results page to "inherit" the look and feel of your site. They never contact the Search Appliance directly.

4.13.1 Link to the Appliance

When you want to make a HTML link to your profile's search interface, select the appropriate profile in the admin interface and click `Search` in the menu bar.

The address of that search page can be used in a link to search that profile, such as:

`<a href="`*searchAddress*`">Search</a>`

Where *searchAddress* is the URL of the search page.

4.13.2 Embed a search box

It's possible to have a small search box on your pages where the user can type a query and submit searches. The search box submits to the Appliance and users will have the search results served by the Appliance, although the look and feel of the search results page can be customized to look like your site.

To acquire the HTML needed for an embedded search box:

- In the admin interface, click `Profiles` in the menu bar and select the profile you'd like to use in the search box.
- Click `Search` in the menu bar. This opens the search form.
- Use your web browser's "View Page Source" option (`View->Source` in the menu, or `Ctrl+U`) to open a window that contains the HTML source code of the page.
- Copy everything between and including the `<form>` and `</form>`.
- Paste the form into your web page(s).

- Add the Appliance's hostname to the beginning of the form's `action` attribute. For example, if your Appliance is `search.mysite.com`, and the existing attribute looked like:

  ```
  <form action="/texis/...
  ```

  You'd change it to:

  ```
  <form action="http://search.mysite.com/texis...
  ```

4.13.3 Request XML search results

Your own dynamic php/asp.net/etc pages can issue a query to the Appliance, and receive back XML results.

Issuing a Query Programmatically

Here is an example URL for a simple XML search:

`http://`*HOSTNAME*`/texis/search/main.xml?pr=`*profile*`&query=`*query*

Where *HOSTNAME* is the IP/hostname of your Search Appliance, *profile* is the profile to search, and *query* is the user's query.

Search Parameters

The possible parameters that can be used in the query string are:

Queries:

- `cq` - Category query: `cq=1` for 1st category, `cq=2` for 2nd, etc. Can be provided multiple times or as a CSV to include results in any of those categories.
- `cqall` - Setting to `Y` requires results to be in ALL `cq` categories, instead of any one of them
- `dq` - Maximum Depth query
- `mtq` - Mime Type query (e.g. `mtq=text/*`)
- `query` - Main search query
- `sq` - Site query (e.g. `sq=docs.company.com`)
- `tq` - Title-only query
- `uq` - URL query

Search control:

- `dateSource` - What date to use, id or Modified (see below)
- `mdgt` - Modified Date Greater Than: Only results with a modified date less than this will be returned.

- `mdlt` - Modified Date Less Than: Only results with a modified date less than this will be returned.
- `pr` - Specifies the Search Appliance profile
- `prox` - Proximity: Only return results with the query words in the same `line`, `sentence`, `paragraph`, or `page` *(default)*
- `order` - Controls the sort order. Values are `r` (relevance), `dd` (newest first), or `da` (oldest first)
- `rpp` - Results Per Page (see also `Max User Results Per Page` search setting (p. 94)
- `sr` - Results Per Site
- `sufs` - Word forms (suffixes). Values are `0` (Exact match), `1` (Plurals & Possessives), `2` (Any word forms), or `3` (Custom)

Rank Knobs: controls the influence of ranking parameters, from Off (`0`) to Maximum(`1000`). Medium (`500`) is the default.

- `rorder` - Word order: Favors results with query terms in the same order as the query
- `rprox` - Rank Proximity: Favors results with query terms close together
- `rdfreq` Database Frequency: Favors results with query terms more rare across the entire profile
- `rwfreq` - Document Frequency: Favors results with query terms repeated more often
- `rlead` - Position In Text: Favors results with query terms earlier in the document
- `rdepth` - Depth In Site: Results more shallow depth (fewer clicks from the Base URL) are better

Additional Fields: To add search restrictions to the query you can specify form variables with a name constructed as `af#OP`, where # is the number of the additional field (1, 2, or 3), and OP is one of the following operations:

- `eq` - the field is equal to the form variable (e.g. af1eq)
- `gt` - the field is greater than the form variable (e.g. af2gt)
- `gte` - the field is greater or equal to than the form variable
- `lt` - the field is less than the form variable
- `lte` - the field is less or equal to than the form variable
- `like` - the field matches the form variable (text search). This has the same syntax and functionality of the Metamorph query engine used in the main text search.

Examples:

- `af1eq=important` - only results where the first additional field is set to `important`

- `af2lt=100` - only results where the 2nd additional field is less than `100`.
- `af3gte=2010-01-01` - only results where the 3rd additional field is 2010 or newer.
- `af1like=important` - only results where the first additional field contains the word `important`
- `af1like=(critical,important)` - only results where the first additional field contains either `important` or `critical`

dateSource: id vs modified

The `dateSource` parameter allows you to determine which date associated with the URL gets used for display, sorting, etc.

- `Modified` *(default)* - The time the content was last modified
- `id` - The time that Search Appliance last updated its record of the content

If a collection of files that were modified a year ago were picked up by the Search Appliance walk last night, then the `Modified` date would be a year ago, but the `id` date would be last night.

`id` is the default dateSource when requesting an RSS feed of a search.

XML Elements in Search Results

Search results can be sent as XML from the Search Appliance to the host server. This section describes the XML elements.

`<ThunderstoneResults>` Overall container for the search results

- `<XmlOutputVersion>` Defines the version of this xml output
- `<ResultsFromCache>` Set to `Y` if this is from results caching
- `<Query>` Main text search string
- `<TitleQuery>` Query applied only to titles
- `<UrlQuery>` Query applied to Url
- `<DepthQuery>` Maximum Depth
- `<MimeTypeQuery>` Query applied to Mime Type
- `<CategoryQuery>` Numeric index for a category to use. `1` is the first category, etc.
- `<RequireAllCategories>` Set to `Y` if results must be in ALL specified `CategoryQuery` instead of ”one or more.”
- `<ResultsPerSiteQuery>` Max results per site

- `<TextQuery>` Main search query, sans `infield:` prefix. Suitable for concatenation with "`infield:`" and re-insertion in a new query
- `<TextQueryHighlight>` TextQuery with query highlighting (if enabled)
- `<PreviousRefine>` Additional refine queries
- `<SiteQuery>` Site query (from `site:` host in the query, or dedicated `sq` query string variable)
- `<InFieldQueriesAllowed>` Set to `Y` if `infield:` queries are allowed
- `<ModifiedDateLessThan>` Only return results with Modified date earlier than this
- `<ModifiedDateGreaterThan>` Only return results with Modified date greater than this
- `<UrlRoot>` URL root of the search script, for making links
- `<Profile>` Profile used
- `<dropXSL>` Whether to apply or drop the XSL stylesheet
- `<AdvancedSearch>` Set to `1` if the advanced form should be displayed
- `<Proximity>` Proximity used for the search. Possible values:
  - `line` - Must occur on the same line
  - `sentence` - Must occur within the same sentence
  - `paragraph` - Must occur within the same paragraph
  - `page` - must occur within same HTML document (default)
- `<Suffixes>` Suffix processing for the search. Possible values:
  - `0` - Exact Match only
  - `1` - Plurals and Possessives
  - `2` - All Word Forms
  - `3` - Custom
- `<Thesaurus>` Set to `1` if the Thesaurus was used for synonyms
- `<Order>` Ordering of the search. Possible values:
  - `r` - relevance
  - `dd` - newest first
  - `da` - oldest first
- `<RankOrder>` Favors results with query terms in the same order as the query
- `<RankProximity>` Favors results with query terms close together
- `<RankDatabaseFrequency>` Favors results with query terms more rare across the entire profile
- `<RankDocumentFrequency>` Favors results with query terms repeated more often

- `<RankPosition>` Favors results with query terms earlier in the document
- `<RankDepth>` Favors results fewer links away from the starting point
- `<mode>` Set to `admin` if this is a Test Search
- `<opts>` Internal use only
- `<authUser>` User that was authenticated via the Proxy Module
- `<metasearchTarget>` Indicates what backend metasearch targets are available, one element for each target. Currently selected targets will have a `selected="selected"` attribute
- `<AdminUrl>` Url to the admin interface
- `<MakeLiveUrl>` Url to make this Look and Feel live
- `<RssUrl>` Url to RSS version of this search
- `<OpensearchUrl>` Url to the OpenSearch version of this search
- `<OpensearchTitle>` Suggested title for this OpenSearch
- `<QueryAutocomplete>` Set to `Y` if Query Autocomplete is enabled
- `<LogoutUrl>` Url for a 'Logout' link
- `<Category>` Categories available for search
  - `<CatVisible>` Set to `Y` if the category should be selectable in the list of categories
  - `<CatSel>` Set to `Y` if this category is currently selected
  - `<CatVal>` Numeric ID for this category
  - `<CatName>` Display name for this category
- `<TopBestBets>` List of "Best Bets" links
  - `<BBTitle>` Title for this section of BestBets
  - `<BestBet>` Individual Best Bet records
    - `<BBResultNum>` Ordered number for this BestBet
    - `<BBPriority>` Priority for this BestBet, as assigned in the admin interface
    - `<BBLink>` Url for this BestBet
    - `<BBLinkDisplay>` Url that displays for this BestBet. Long Urls are intelligently truncated for display
    - `<BBResult>` Url text for this individual BestBet, as assigned in the admin interface
    - `<BBDescription>` Description for this individual BestBet, as assigned in the admin interface
    - `<BBGroupname>` Name of the BestBet group this BestBet belongs to
    - `<BBGroupid>` id of the BestBet group this BestBet belongs to
    - `<BBKeywords>` Keywords that trigger this BestBet record to display. This is all keywords for this individual record, not just the one that triggered this activation

- `<ProfileInfo>` Encloses some profile summary info
  - `<Profile>` Profile to which this `ProfileInfo` refers to
  - `<ResultDecl>` Declarations of User Fields that will be in `Result` elements, each has a `name` and `type` attribute
  - `<ExitIsEarly>` Set to `Y` if search aborted
  - `<ExitReason>` Set to `ok` if search finished normally, otherwise token indicating reason (see ExitReason table below)
  - `<RedirectUrl>` If present, specifies external Url to redirect the user to for Results Authorization
  - `<LoginUrl>` If present, local Url to login with `rauser`/`rapass` for Results Authorization.
- `<Summary>` Encloses search results summary, only present if a search was actually performed
  - `<Profile>` Profile that this `Summary` element applies to
  - `<Start>` First result item to list
  - `<End>` Last result item to list
  - `<TotalNum>` Total number of result items found, *before* ResAuth etc.
  - `<TotalIsEstimate>` Set to `Y` if `TotalNum` is an estimate
  - `<TotalIsShort>` Set to `Y` if `TotalNum` is known to be short (e.g. early exit)
  - `<UserResultsNum>` Total number of result items found, *after* ResAuth etc.
  - `<UserResultsIsEstimate>` Set to `Y` if `UserResultsNum` is an estimate
  - `<UserResultsIsShort>` Set to `Y` if `UserResultsNum` is an estimate
  - `<ResultsAuthorization>` Set to `Y` if Results Authorization was used
  - `<Total>` Readable text for total number of results, *after* ResAuth etc.
  - `<GroupBySite>` Set to `Y` if Results per Site was used with this query.
  - `<CurOrder>` Text that describes the order by which results are listed
  - `<OrderLink>` Url that provides an alternative sorting order results list
  - `<OrderType>` Text that describes `OrderLink`
  - `<NewSkip>` (Metasearch only) Skip value to use for any further request. Only needed with the SOAP API
  - `<PreviousLink>` Url to the previous page of results
  - `<FirstPage>` Set to `1` if this is the first page of results
  - `<Pages>` Contains data on pages of results
    - `<PageLink>` Url to a certain page of results
    - `<PageNumber>` Page number a page of results
  - `<NextLink>` Url to the next page of results
  - `<LastPage>` Set to `1` if this is the last page of results
  - `<Credit>` Text to introduce the credit image
  - `<CreditImage>` Url of the credit image

- `<Result>` Contains data about a given result
  - `<Profile>` Profile for this `Result`. Note that results from meta-search back-ends are re-labeled to the front-end profile
  - `<Num>` Number of this result item
  - `<Skip>` Internal use: raw skip(s) for result. Valid for Meta Search back-ends
  - `<Id>` Identifier for this result
  - `<ResultTitle>` Title of this result
  - `<Url>` Url of this result
  - `<ClickUrl>` Url for this result item, as should be clicked by the user. Use `Url` if not present. Only sent if Query Logging is enabled, in which case it contains redirect for logging the click-through
  - `<UrlPDFHi>` Url to highlight this PDF in Acrobat Reader, only used with `Legacy` highlighting
  - `<UrlDisplay>` Displayed Url for this result
  - `<UrlCached>` Url to retrieve the cached version of this result
  - `<RawRank>` Raw relevance rank value for this result (0-1000)
  - `<ScaledRank>` Raw rank scaled up for a more-like-this search (0-1000)
  - `<PercentRank>` ScaledRank as a percentage (0-100)
  - `<DocSize>` Size (bytes) of this result
  - `<MimeType>` MimeType for this result
  - `<MimeTypeIcon>` Icon file to use for this MimeType
  - `<Depth>` Number of links walked from Base Url to this Url
  - `<UrlSimilar>` Url to search for pages similar to this result
  - `<UrlInfo>` Url for context of answers within a matching document
  - `<UrlParents>` Url of pages that link to this search result
  - `<Modified>` Date and time this result was last modified
  - `<Visited>` Date and time this result was walked
  - `<Abstract>` Brief text surrounding the matched word or phrase
  - `<Charset>` Character set of the formatted text of the page (typically Storage Charset unless conversion failure)
  - `<SiteName>` Name of the site for this result item
  - `<UrlMoreResultsFromSite>` Url for more results from this site
  - In addition, any Additional Fields that have been selected for Output will be sent as child elements of `Result`, one per field. Each element is named after the field, with a `u:` XML namespace prefix since they are custom fields. The value of the field will be the content of the element.

    For example, an `Integer` field `Quantity` and a `GMLPoint` field `Location` may be given as:

```
<u:Quantity>57</u:Quantity>
<u:Location>47.4500 -122.3000</u:Location>
```

- `<RightBestBets>` List of right "Best Bets" links, see `TopBestBets`
- `<Spelling>` Spelling suggestions
  - `<SuggestWord>` An individual spelling suggestion
    - `<SpellPhrase>` Label for the suggestions
    - `<SpellLink>` Url to search for the suggestion
    - `<SpellWord>` Suggestion content
    - `<SpellCount>` Number of results for this suggestion
- `<exportVar>` Additional exported variables
- `<QueryMessage>` Messages to show to the user
- `<Message>` Additional diagnostic messages

The following table lists the possible value tokens for the `<ExitReason>` element:

Table 4.2: XML `<ExitReason>` Tokens

Token	Description
ok	Normal exit
ResAuth-ExternalLoginRequired	Need Login Cookies: redirect to `<RedirectUrl>`
ResAuth-CredentialsRequired	Need user/pass: send `rauser`/`rapass` to `<LoginUrl>`
ResAuth-LoginIncorrect	User/pass incorrect; re-send to `<LoginUrl>`
ResAuth-SuccessLimit	Successful Auth Result Limit reached
ResAuth-Timeout	Results Authorization timeout
ResAuth-MaxDocsCheck	Max Docs to Auth-Check exceeded
ResAuth-SmbError	SMB error
ResAuth-NoSmb	SMB unavailable/could not be run

Invoking Query Autocomplete

Query Autocomplete can be used in your own custom front end using Javascript similar to that used by the normal search interface. If you want to invoke it arbitrarily, you can request URLs of the form:

`http://`*HOSTNAME*`/texis/search/autocomplete.json?pr=`*profile*`&term=`*term*

Where *HOSTNAME* is the IP/hostname of your Search Appliance, *profile* is the profile to search, and *term* is the user's partially typed term.

Autocomplete returns a JSON array in the OpenSearch format (http://www.opensearch.org/Specifications/OpenSearch/Extensions/Suggestions). Getting completions for `term=sea` would return something like:

```
["sea",["seattle","sears","search"]]
```

Autocomplete also supports JSON-P, so adding `&callback=updateList` to the URL would return:

```
updateList({term: "sea", completions: ["seattle","sears","search"]})
```

Alternatively, you can request `autocomplete.xml` instead of `.json` to get an XML document back:

```
<Completions>
  <Term>sea</Term>
  <Completion score="16">seattle</Completion>
  <Completion score="5">sears</Completion>
  <Completion score="1">search</Completion>
</Completions>
```

4.13.4 Invoke the search SOAP API

Instead of making a HTTP request and parsing the XML response, it's possible for pages to invoke the search SOAP API. This allows interactions with the Appliance to appear as local function calls, automatically handling all the details of HTTP requests and XML parsing.

See the **SOAP API** (p. 169) for details on setting up and using the SOAP API.

4.14 Search Result RSS Feeds

Search result RSS feeds can help you monitor a certain search query, and let you know when new results appear for the query.

All search result pages have an RSS link embedded in them. Recent versions of modern browsers, such as Internet Explorer and Firefox, have built-in features that notify you when an RSS feed you're subscribed to changes.

- IE 7 and 8 - `http://www.microsoft.com/windows/IE/ie7/tour/rss/`
- Firefox - `http://kb.mozillazine.org/Live_Bookmarks_-_Firefox`
- Opera - `http://www.opera.com/mail/rss/`

4.15 OpenSearch Support

The search interface also has an embedded Open Search description. This means that modern browsers can use the Quick Search box (to the right of the address bar) to perform searches on the Search Appliance.

- Bring up the search interface for the profile of your choice

- Hit the ”down” arrow next to the Quick Search box
- Choose “Add Search Provider...” to add the Search Appliance to the list of available searches.

Internet Explorer users can find more detailed instructions at
`http://msdn.microsoft.com/en-us/library/cc848862.aspx`

Adding `strictSpec=Y` to the Open Search Description URL will cause the Search Appliance to truncate various fields as required by the specification. In practice, this isn’t necessary as browsers handle longer names.

4.16 Using Best Bets

The Search Appliance allows you to create links that will appear either at the top or to the right of the search results (or anywhere else, if using an XSL stylesheet) 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.

4.16.1 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 desired URL and click on the URL to get the details on that URL. There is a form on the page that allows keywords to be added to that URL. You can define a priority, title, description, and keywords for the URL (as detailed in the list below, under **Fully Customized**).

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.

4.16.2 Fully Customized

Best Bets can also be created fully customized. The first step is to define a group. This is done from the “Group Settings” tab. You can name the group, and decide which information will be displayed about the group.

After creating a group you can add keywords to specific URLs. From the “List/Edit URLs” page enter the URL you want, and click on the URL to get the details on that URL. Currently you can only use URLs that have been walked and are in the database. The fields on the form are:

- `Priority` *(optional)* - The priority for this Best Bet. If multiple Best Bets match a single query, their `Priority` fields determine what order they’re shown. If you’re only setting one Best Bet per URL, or aren’t particular about the order, you don’t need to set a priority.

- `Title`- The title that will be displayed for the Best Bet on the search results page.
- `Keywords`- A comma separated list of what keywords will trigger this Best Bet. A Best Bet is displayed when the query is completely contained within one of the keywords for that Best Bet.
- `Group`- Which Best Bet Group this Best Bet will be created in. A Best Bet will only have a chance to match if its group is set to display as either `Top Best Bets` or `Right Best Bets`on the Search Settings page.

  If no groups currently exist, `(create new)` will be displayed, and a group will be created for you if you enter keywords and a title for this Best Bet.
- `Description` *(optional)* - The description to display for this Best Bet. The Best Bet Group for this Best Bet might be set to not display the description, so it's optional.

The title and description can contain HTML code. Be careful that it does not disrupt the rest of the page layout. You can create multiple entries for the same URL. Each time you save a new set of blank boxes will be shown.

Once the Best Bets are created you can go to the "Search Settings" page to set up how they are displayed. For the top and right placements you can define which group is shown there, what title if any to display above the links, and the color, size and style of the boxes around the Best Bets.

As with any of the Search Settings these will apply to the "Test Search" first, and then when you apply the settings be copied to the "Live Search", allowing you to test the settings and make sure they are appropriate before going live.

4.17 Using Access Control

The concepts and actions of access control in the Search Appliance are discussed in detail in the Access Control section, p. 127. The following are some general tips on how to setup and maintain access control rights.

4.17.1 Initial Lockdown

Since the default mode for Access Control when created is to allow all rights to all users for back-compatibility, it is recommended that perms be "locked down" first, and only granted as needed. The `admin` user, having the irrevocable ability to reset ACLs, should remain a "superuser" with all access, and other accounts turned into lesser-permission users. Lockdown should happen in this order:

1. Allow superuser: The `admin` user should have an `Allow` entry for all rights to the top-level `Global` object[1].
2. Deny everyone: The group `Everyone` should have a `Deny` entry for all rights to the top-level `Global` object.

[1] In version 5.3.0 and earlier, the admin user should instead be explicitly granted all rights to each of the second-level objects (All Users, All Groups, All Profiles, All Settings, and Maintenance).

With these perms, users other than `admin` – including new users and profiles created in the future – will not be able to see or modify administrative settings. They can be granted perms as needed later, for example, the Read right could be removed from the `Global` deny ACE so that they can read but not modify any admin action/setting.

4.17.2 Example: User with Complete Control on One Profile

To configure a user that has complete access to just one specific profile (but no other profiles, nor the rest of administration such as creating accounts etc.), set up the lockdown settings above, then:

1. Create a Profile ACE on the specific profile, for that user, read and write access, and type `Allow`.
2. Create a Setting ACE for `All Settings`, for that user, read, write and delete access, type `Allow`.

The user will now be able to modify any setting on that profile, as well as start/stop walks on it, but will not be able to edit other profiles.

4.17.3 Example: User with Look and Feel Control on All Profiles

To configure a user that has the ability to change the Top and Bottom HTML on *any* profile, but cannot edit walk settings, nor start nor stop a walk, etc., set up the lockdown settings above, then:

1. Create a Profile ACE on `All Profiles`, for that user, read and write access, and type `Allow`.
2. Create a Setting ACE for `Top HTML`, for that user, read, write and delete access, type `Allow`.
3. Create a Setting ACE for `Bottom HTML`, for that user, read, write and delete access, type `Allow`.

The user will now be able to change the top and bottom HTML for any profile.

4.18 Indexing File Servers

The Thunderstone Search Appliance can index Windows and Unix file servers in addition to web servers. To do so you will first need to mount the file server to the Search Appliance, and then configure the walk. To mount the file server you need to go to the System page and select `Network Filesystems & Shares`3.3.

After you have mounted your server to the Search Appliance the interface will give you the Base URL to use in your profile. You may use that url or anything underneath it.

4.19 Replication

4.19.1 Replication Overview

In replication, a server profile sends walk data to another server profile. The two profiles can be on different machines or they can be on the same one. If the profiles are on different machines, the sending and receiving profiles can have the same or different names. If the profiles are on the same machine, use different profile names.

Here is an example that illustrates the replication process. In this example, the `Sender` profile has been set up as the sender profile and `Receiver` is the receiver profile. After `Sender` performs a walk, it sends the walk data to `Receiver`. The `Receiver` profile accepts the data as-is, without regard to its own profile settings. Only the profile that performed the walk may send the walk data, so in this example `Receiver` cannot replicate (the data it received from `Sender`) to another profile.

To avoid undesired overwriting of replication walk data, you should not allow the receiver profile to perform walks.

Before the receiver will accept replication data, the sender(s) need to be granted permission to send the data. This permission is managed in a cluster member list.

A good use of replication is to set up multiple machines to replicate to a single receiving profile. For example, machines A, B, and C each have a different profile, and they each replicate their walk data to a profile on machine D, which is the receiver. Another use of replication is to send walk data from multiple profiles on a machine to a single receiver profile that is on the same machine. This provides a means of combining walk data into a single profile. Another use of replication is to replicate data from one sender to multiple receivers. This way multiple machines hold the same walk data.

4.19.2 Procedure

The procedure in this section is an example of setting up replication on a single machine. It can be adapted to multiple machine configurations by changing the Replication Settings.

Set up the Sender Profile

- Choose an existing, walkable profile to be the sender. Or go to the `Profiles` menu item and create one, filling in all fields for a normal walk. We'll assume this profile is called `Sender`.
- Go to the `All Walk Settings` menu item for the `Sender` profile.
- Scroll down to `Replication Settings`.
- Enter the information for the receiver. In this example, `Host IP or Name` is `localhost` because we'll be sending data to the same machine, and `Profile Name` is `Receiver`. The page now includes the location of the receiver profile.
- Click `Update and Go` button.

- After a moment, the `Walk Status` page opens. Notice that there are N items in the replication queue. The number N is similar to the number of pages that were walked. The items remain in the queue, because they cannot be sent until the receiver profile is created (below). Normally, when a receiver profile is present, the contents of the queue are automatically sent to the receiver.

Create the Receiver Profile

- Create a new profile called `Receiver` via the `Profiles` menu item. (This matches the receiving profile name we entered on the `Sender` profile.)
- At main menu click `System`, then under `System Settings` heading, click `System Wide Settings`.
- At the `Cluster Members` field, enter the IP address for each server that will send walk data to this machine. Use a separate line for each entry. In this example, there is one sending IP address, and it is 127.0.0.1 (use IP numbers, not the word `localhost`). To enable an entire subnet to send data, use an IP prefix and wildcard, e.g. `10.10.*`.
- Click `Update` button.
- At main menu, click `Profiles`.
- When `Profiles` page opens, click `Sender`. A `Walk Settings` page opens for the `Sender` profile.
- Click `Walk Status` button. The `Walk Status` page for the `Sender` profile opens.
- There are still N items in the replication queue.
- Click the `replication queue` link.
- The items in the replication queue are sent to the `Receiver` profile. On the `Walk Status` page, there are now 0 items in the replication queue, which indicates the items were sent.
- On main menu, click `Profiles`, click `Receiver`, click `Walk Status` and observe that there is a list of pages recently walked. These pages were not walked by `Receiver`, instead they were obtained from `Sender`, which performed the walk.

4.19.3 DataLoad API

The replication system can also be used to load data directly onto the Search Appliance from an outside source, instead of "pulling" it from a URL or its links. This can be used for data that is not permanently stored at its URL (e.g. generated data), and therefore cannot be fetched for indexing; it can instead be pushed to the Search Appliance for indexing. This feature requires version 5.4.19 or later of the `texisScripts` package (see System / Update Software).

Before loading data onto the Search Appliance, it must be configured to accept data from the IP address(es) that will be sending to it. This procedure is the same as for replication; see the Cluster Members setting, p. 154.

Submission Format

Data is submitted to the Search Appliance with an HTTP POST request sent to a similar URL as the admin interface (e.g. `http://.../dowalk`), but with `/recvdata.xml` appended. E.g.:

```
http://www.mysite.com/texis/dowalk/recvdata.xml
```

The following POST variables must be set in the request. Be sure to URL-encode the values:

- `profile`
  Set to the name of the receiving profile.

- `data`
  Set to an XML document containing the data, and what to do with it (insert/delete/etc.). See below for details.

Specifying all fields manually

Below is an example `data` document where all fields are specified. Be sure to HTML-encode values.

```
<?xml version="1.0" encoding="UTF-8"?>
<ThunderstoneReplication
      xmlns:dt="urn:schemas-microsoft-com:datatypes"
>
  <Item>
    <Type>I</Type>
    <Size>150369</Size>
    <Visited>2005-10-25 15:25:18</Visited>
    <Dlsecs>0</Dlsecs>
    <Depth>0</Depth>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Title>Sprocket Specifications</Title>
    <Body>...</Body>
    <Keywords>sprockets, gears, hubs</Keywords>
    <Description>Sprocket details</Description>
    <Meta></Meta>
    <Category>Mechanical</Category>
    <Modified>2005-10-25 11:21:07</Modified>
    <NextCheck>2005-10-25 16:25:18</NextCheck>
    <Views>0</Views>
    <Clicks>0</Clicks>
    <CTR>0.000000</CTR>
    <Pop>0</Pop>
    <MimeType>text/html</MimeType>
    <Charset>UTF-8</Charset>
    <Refs dt:dt="bin.base64">...</Refs>
    <Errors dt:dt="bin.base64">...</Errors>
    <RawData dt:dt="bin.base64"></RawData>
  </Item>
</ThunderstoneReplication>
```

Any element whose text data might not be XML-safe (e.g. binary chars in the `<Body>`) should be base64-encoded, and the attribute `dt:dt="bin.base64"` set in the tag. E.g. the `<Refs>` and `<Errors>` elements' text data are always base64-encoded. Note that the XML namespace prefix `dt` should also then be set to `urn:schemas-microsoft-com:datatypes` in the root `<ThunderstoneReplication>` element.

The elements are:

- `<Type>` The action to take with this data. Text value may be one of:
  - `I` Insert the data (overwrite previous data for URL if any)
  - `D` Delete the URL
  - `DP` Delete the URL as a pattern (e.g. `http://www.mysite.com/dir/*`)

  - `UI` Update search indexes (call after a batch of inserts/deletes)

- `<Size>` The integer size of the original document.
- `<Visited>` When the document was fetched, in `YYYY-MM-DD HH:MM:SS` format.
- `<Dlsecs>` Number of seconds taken to download the document.
- `<Depth>` Depth of URL from a Base URL, e.g. 0 is a Base URL, 1 is one click away, etc.
- `<Url>` The URL of the document.
- `<Title>` The title of the document.
- `<Body>` The formatted body of the document.
- `<Keywords>` Any keywords for the document.
- `<Description>` The description of the document.
- `<Meta>` Any meta data for the document.
- `<Category>` The category the document is in, if any. Must be a category name from the profile's Categories.
- `<Modified>` The `Last-Modified` date of the document, in `YYYY-MM-DD HH:MM:SS` format.
- `<NextCheck>` When the document should be refreshed, in `YYYY-MM-DD HH:MM:SS` format.
- `<Views>` Number of views of the document: how many times it's been shown in search results.
- `<Clicks>` Number of clicks of the document: how many times it's been clicked on in search results.
- `<CTR>` Click-through-ratio: floating-point number ratio of clicks to views.
- `<Pop>` Document popularity: number of references (links) to it.
- `<MimeType>` The MIME type of the content served at the URL, or provided in RawData.
- `<Charset>` Character set of `<Body>` data. Should correspond with `Storage Charset` profile setting (p. 70). If a charset other than the `Storage Charset` is used, it should be a standard IANA charset that the Search Appliance can convert to the `Storage Charset`.
- `<Refs>` Optional element with references (child links) of the document.
- `<Errors>` Optional element with errors of the document.

Uploading a binary file

If you have a binary file, such as a PDF or an Office document, you can send it with the dataload API and let the Search Appliance extract the text from it.

```
<?xml version="1.0" encoding="UTF-8"?>
<ThunderstoneReplication
      xmlns:dt="urn:schemas-microsoft-com:datatypes"
>
    <Item>
        <Type>I</Type>
        <Url>http://www.example.com/dataload.pdf</Url>
        <RawData dt:dt="bin.base64">0M8R4KGxGu....</RawData>
    </Item>
</ThunderstoneReplication>
```

The elements are:

- `<Type>` The action to take with this data. Text value may be one of:
  - `I` Insert the data (overwrite previous data for URL if any)
- `<Url>` The URL of the document.
- `<RawData>` element with the base64 encoding of raw document. It must include the `dt:dt="bin.base64"` attribute.

Combining the two: binary files with custom fields

It is possible to specify both a `<RawData>` document, *and* fields such as `<Title>`, `<Description>`, etc. The binary document will be processed, and any other fields provided will override the values that came from the document.

This can be useful in situations where you have a Content Management System (CMS) that contains metadata about a document that doesn't actually *occur* anywhere in the document. You can do a custom dataload that pushes in the document, and the custom Title/Description/etc.

Additional Fields

Each profile-specific Additional Field is optionally sent in a single element named after the field, with the XML namespace prefix `u`. The value of the field is the content of the XML element. Note that the `u` XML namespace prefix should be declared in the root `<ThunderstoneReplication>` node, as shown earlier.

For example, an `Integer` field `Quantity` and a `Text` field `State` may be given as:

```
<u:Quantity>57</u:Quantity>
<u:State>NY</u:State>
```

Other Details

The optional `<Refs>` element lists the links (references) from the given document, for parent-child linking. Its text value is a base64-encoded XML document with the following format when decoded:

```
<results xmlns:dt="urn:schemas-microsoft-com:datatypes">
  <result>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Ref>http://www.mysite.com/dir/otherpage.html/</Ref>
  </result>
  ...
</results>
```

Each `<Url>` should be the same as the `<Url>` in the above `<Item>` block. The `<Ref>` is a single link from the page. Only one `<Ref>` may be listed per `<result>`; additional links should be sent with additional `<result>` elements.

The optional `<Errors>` element contains any errors to be logged for the document. Note that this may be empty or not present if no errors are to be logged. Its text value is a base64-encoded XML document with the following format when decoded:

```
<results xmlns:dt="urn:schemas-microsoft-com:datatypes">
  <result>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Reason>Document not found: 404 (Not Found)</Reason>
  </result>
  ...
</results>
```

As with the `<Refs>` element, the `<Url>` must correspond with the original `<Item>` `<Url>`, and multiple errors must be listed in separate `<result>` elements.

Reply Format

The response to a DataLoad request is an XML document:

```
<ThunderstoneReplicationResult>
  <ItemResult>
    <rid>000000000</rid>
    <Type>D</Type>
    <DP>1</DP>
    <Status>OK</Status>
    <Info>Not found</Info>
  </ItemResult>
  <Rows>1</Rows>
  <Version>Version 5.01.1234567890 20051010 (...)
           2005-10-10 12:34:56</Version>
</ThunderstoneReplicationResult>
```

The elements are:

- `<rid>` The replication id. Ignored.
- `<Type>` The action type specified in the request.
- `<DP>` The number of URLs deleted by a `<Type>DP</Type>` action. Element is not present for other `<Type>`.
- `<Status>` Result code:
  - `OK` Success
  - `FAIL_UNKNOWNTYPE` The `<Type>` was not recognized
  - `NODATA` No parsable data in request
  - `Not Allowed` Sender is not a `Cluster Member`
  - `No Profile` No `profile` set in request POST
  - `FAIL` Failed, unknown reason
- `<Info>` Optional additional message; e.g. `Not found` if a non-existent URL is deleted
- `<Rows>` How many request `<Item>`s were processed.
- `<Version>` Version and release date of the software.

Once data has been successfully loaded onto the Search Appliance, if the profile has any receiver profiles defined under `Replication Settings`, the data will also be queued for replication to those receivers.

Dataload SOAP API

There is a SOAP API available for dataload, allowing you to use a SOAP library to communicate with the Search Appliance. For an overview of SOAP, Please see the **SOAP API** (p. 169).

The WSDLs for the dataload API can be found on the `Profile Tools` page. Providing these WSDLs to whatever tool your language uses, such as Visual Studio's `wsdl.exe` program, should generate the necessary wrapper class.

The parameters are defined within the WSDL itself, and are generally the same as mentioned above in the **Submission Format** and **Reply Format**s, with a few exceptions:

- The entire transactions are wrapped by SOAP envelops and the top-level elements are called `dataload` and `dataloadResponse` instead of `ThunderstoneReplicationResult`, respectively.
- The `dataload` element contains a `profile` element in addition to all the `Item`s.

C# Example Project

A C# example project is available that demonstrates using both the search and dataload SOAP interfaces. In the `System -¿ Advanced Tools` section of the administration interface, choose `Extra Downloads`, and then `Thunderstone Soap Example`. Instructions are listed on that page and within the zip itself for how to use the project.

4.20 Additional Fields

4.20.1 Overview

The additional fields feature in the Search Appliance allows you to define structured data that can be searched on, sorted by, and included in the results when using an XSL stylesheet. Typical uses might include having prices, dates or ratings associated with the documents.

4.20.2 Populating

To populate Additional Fields they should first be defined in the **Additional Fields** section of **All Walk Settings**. You can specify a name, which is used as the name of the XML element when displaying the results, as well as when using the DataLoad API.

Once the field has been defined it can be populated either via the DataLoad API or through the **Data From Field** settings section. The fields are positionally numbered, and you can load Extra Field 1, 2 and/or 3 from the page that is read. If you are loading from a `<meta>` field you will typically want a **REX Search** of `.+` and the **From Meta Field** you are loading from.

4.20.3 Sorting

To sort the results you can use the `order` form variable. To specify the first field you can set the value to `af1`, for the second `af2` and for the third `af3`. If you want to reverse the sort order you add a `d` to the value, i.e. `af1d`, `af2d`, `af3d`.

4.20.4 Searching

See the **Additional Fields** section in **Issuing a Query Programmatically** (p. 142) for information on specifying additional field searches in the URL.

4.21 DBWalker

4.21.1 Overview

The Texis DBWalker module provides a walkable HTML interface to a remote database. If there is a database server somewhere which has a JDBC driver, DBWalker can serve up that database via HTML, which can then be walked by the Search Appliance or viewed by users. DBWalker can be configured to print all records on a single page, or to provide an "index" page which creates links to individual pages, each of which shows a single record of the table.

DBWalker is different from the normal idea of an "import" in multiple ways:

- DBWalker does not do any actual "importing" at all - it simply enables a way to view parts of a database through a website. It's still up to the Search Appliance to walk the given website and index its content.
- The idea of 'import' implies a one-time action. Because DBWalker provides a HTML interface, it can be used to keep up to date with changes to the remote database. If a single record in the remote database changes, then DBWalker's HTML interface will be different, and a refresh walk by the Search Appliance will see this and change its internal index accordingly.

JDBC connections are cached across HTTP requests. The first time a request for a configuration is received, it establishes a JDBC connection and keeps it for 5 minutes. If another request for the same configuration is received, it will re-use the same connection. This greatly enhances the Search Appliance's walking speed, and keeps from bogging down the remote database with unnecessary connect/disconnect activity. JDBC connections are closed after 5 minutes of inactivity.

4.21.2 Configuration Overview

DBWalker uses multiple individual configurations for the different databases and tables it needs to talk to. Each configuration describes a single group of settings for a single table in a single database. It is possible to have multiple configurations use the same table and databases – for example, you can have one configuration list the entire contents of a table, while another configuration limits the data to a certain range.

Each configuration specifies which database to talk to with a type ("PostgreSQL", "Oracle", etc., which determines what JDBC driver to use), a JDBC connection string (which specifies things like host, port, and database), a username, and a password. The configuration must also specify which table is to be read, and can optionally specify which columns to read (defaults to all), any filter for the data (by way of a `WHERE` clause), and a key field.

If no key field is specified, then DBWalker won't know how to uniquely identify rows, so it will print all the data on a single HTML page. If a key field is specified, then DBWalker will create an index page, which lists only the key field column. Each row's key field is listed as a link back to DBWalker, which will give a page displaying all of the selected fields of only that record. This allows more fine-grained indexing and searching in large tables.

4.21.3 DBWalker Output Overview

The DBWalker's internal libraries produce XML output. This is transformed into HTML via a XSL stylesheet, which is changeable on a per-configuration basis. When a request is received that ends in `.xml`, DBWalker will return an XML document with a reference to an XSL stylesheet on the DBWalker server (which modern browsers will automatically fetch and apply).

However, if a request is received with the extension `.html`, DBWalker will apply the stylesheet server-side before the client ever sees it, and hand the resulting HTML to the client. This is useful for clients that do not apply XSL transformations to XML documents (like the Search Appliance).

4.21.4 DBWalker Authentication Overview

There are two ways to do authentication with the remote database. Authentication information can be stored in the config file, or it can be provided dynamically.

If a username and password are provided in the configuration, then that user/pass will be used for every request for that config. This has the advantage that users never have to input a username/password, but also has the security disadvantage that anyone who opens the website can see the data. Depending on the contents of the database, this may or may not be significant.

The other option is to not include a username or password in the configuration. When DBWalker is invoked by in this situation, it will prompt for a username/password (via Basic authentication). If the remote database accepts the credentials, the page is displayed. For the Search Appliance to walk these pages, it will have to know a valid username/password for the pages. This is supplied in the **Login Info** of **All Walk Settings** ((80).

This integrates well with Results Authorization. if the Search Appliance's search is set to use Authorized Search Results with "Prompt via Form", the credentials will automatically be verified with DBWalker.

To summarize a few key points:

- If you include a username and password in the DBWalker config, anyone will be able to see the results, including searches with `Results Authorization` in use: any user's login will work since the correct user/pass is "built in" to the DBWalker config.

- If no user/pass is included in the config, then users will have to supply their own username/password, and the config can be used properly with `Results Authorization`.
- If no user/pass is included in the config, be sure to put a valid username/password in the`Login Info` (pg. 80) section of any profile that uses it so the Search Appliance is able to index the content. If the Search Appliance can't see the results, then no searches will find it!

4.21.5 Obtaining DBWalker

DBWalker is obtained through `System -¿ Update Software`. The `j2re` update must be installed prior to the `DBWalker` update. Please see the **Getting Software Updates** (pg. 136) for details on installing software updates.

4.21.6 Managing DBWalker

- At main menu, click `System`, then `DBWalker Settings / Status` under Manage Modules.

The main DBWalker Administration interface is divided into 3 sections:

- DBWalker Status

  This shows whether the DBWalker is currently enabled, and if so what the process ID is. A button is provided to enable or disable the DBWalker, whichever is applicable.

  A link is also provided to the `DBWalker Global Options` page.

- DBWalker Configurations

  This section lists all the configurations available for DBWalker. It shows the configuration name, what type of database it uses, its table, and its JDBC connection string. If the DBWalker server is currently running, the name of each configuration will be a link to that configuration's DBWalker page.

  Here you have the option to edit or delete any configuration, and to create new ones with the "Create New Configuration..." button at the bottom. Please see the `Managing DBWalker Configurations` section below for more information on managing configurations.

- DBWalker Stylesheets

  This section allows you to manage the XSL stylesheets that DBWalker uses. You can add, delete, or modify, or view stylesheets here. Please see the **Managing DBWalker Stylesheets** (pg. 168) below for more information on stylesheets and how to manage them.

4.21.7 DBWalker Global Options

The DBWalker Global Options page lists settings that affect all configurations, usually involving the DBWalker environment.

These are considered "advanced" settings and should only need to be changed if advised to by Thunderstone Support.

- `LogLevel`

  This affects how much information is written to `DBWalker.log`. Each level includes all levels above it.

  - `SEVERE` - Errors that cause the DBWalker to fail, giving it no chance to continue. These include errors reading the server config file, server socket errors, errors setting up the JDBC classloader, etc.
  - `WARNING` - errors that cause an individual request to fail, but allow the server to continue on servicing other requests. These include individual connection socket errors, individual configuration errors errors, unexpected SQL errors, etc.
  - `INFO` - This is the default logging level. Logs errors that are probably caused by malformed clients, or other things that we think an admin should know about. Includes clients giving malformed HTTP headers, requesting nonexistent configs, server startup/shutdown notification, etc.
  - `CONFIG` - reports all information being read from configuration files.
  - `FINE` - More fine-tuned book-keeping without going into details about individual classes/methods. Includes worker/socket assignment, worker pool manipulation, cache manipulation, enter/exit of JDBC methods.
  - `FINER` - reports run/stop/resume of individual threads, more details of worker processing of socket.
  - `FINEST` - Kitchen sink and then some. This will cause the log to fill very quickly with superfluous information during normal operation, and is advised not to be used unless requested by Thunderstone Support.

  The default value of `INFO` is fine for normal operation.

- `Max JVM Memory`

  This allows you to increase the maximum amount of memory, in megabytes, that the JVM will allow itself to allocate. If you are working with very large tables and getting `OutOfMemoryException` errors, them you may need to increase this. The default value is 64Mb.

  Note that this value is not the amount of memory that will be immediately allocated by the JVM - it will only allocate as much as it needs. This simply provides an upper limit on how much memory will be used.

- `Accept Debug Connections`

  When working with Thunderstone Software, this can assist in the troubleshooting of DBWalker problems. There's no reason to set this unless Thunderstone Support tells you to.

4.21.8 Managing DBWalker Configurations

Choosing to either edit a configuration or create a new one takes you to a listing page where you can change the facets of a configuration.

- `General Information`

  The General Information section contains things that don't pertain directly to the remote database itself.

  - `Configuration Name`

    If you're creating a new configuration, you will be asked to enter a name. It is used when specifying which group of settings you want to use when DBWalker is invoked, but has no bearing beyond that. Names may contain letters, numbers, dashes, and underscores (no spaces).

  - `Stylesheet`

    Specifies which XSL stylesheet to use. You can only use stylesheets that you've already uploaded. Please see the **Managing DBWalker Stylesheets** (pg. 168) for more information.

  - `Max Rows per Page`

    Sets a maximum number of rows to use on a single index page. If there are more rows than is allowed on a single page, `next` and `back` links are used as necessary to see the rest of the links.

    This is because if a table contains 10 million rows, just generating the index page can take huge amounts of time. DBWalker can be told to only deal with 100 rows at a time, keeping it from getting bogged down.

  - `Appliance Link`

    If you are using an internal interface to access the Search Appliance's administrator interface, this can allow you to force the DBWalker to be walked through an interface that will be visible to external users. Usually the default for this will be fine.

    If administrators are accessing the Search Appliance through an internal-only interface, let's say `internalonly.mysite.com`, then the DBWalker will get walked as `http://internalonly.mysite.com/texis...`. This will work fine for the walk itself, but when external users use search, they will see results referencing `internalonly.mysite.com`, which they won't be able to access.

    By setting Appliance Link to something like `www.mysite.com/texis` (or whatever external users will be able to see), then the DBWalker will get walked with the proper links.

- `Database Information`

  The database information section collects information about how to connect to your remote database.

  - `Type`

    This determines which JDBC driver will be loaded. DBWalker comes with support for `Oracle`, `Microsoft SQL Server`, `Sybase`, `PostgreSQL`, `MySQL`, and `Texis`.

    The `Oracle (dedicated)` type is used to connect to an Oracle database through `dedicated mode` instead of the default `shared mode`. There is a slight performance disadvantage to this, and should only be used when the ordinary `Oracle` type does not work.

    Alternatively, you can select `[jdbcConnect]` as the type, which lets you manually enter the JDBC Connection String. The `Host`, `Port`, and `DB/Service` values are all contained in the

JDBC connection string, so the `Connect String` field replaces all 3 of them.

- `Host` / `Connect String` The contents of this field depend on what `Type` you have selected.

Database Type	field contents
Oracle, Sybase, PostgreSQL, MySQL, or MS SQL Server	the hostname of the machine you're connecting to, or its IP address.
Texis	the hostname and full path to the JDBC script on the remote server, i.e. `host.mysite.com/texis/jdbc`.
[jdbcConnect]	the full JDBC connection string.

  If the type is `[jdbcConnect]`, then this field is `Connect String`, which lets you specify the full JDBC connection string. This is useful if you already know the JDBC connection string for your remote DB and don't want to have to break it down into hostname, port, etc. The exact formatting of this string differs for each remote database type.

- `Port` The port number that the remote database is listening to.

Oracle, Sybase, PostgreSQL, MySQL, or MS SQL Server	the port to use, or leave blank for the default.
Texis	unused, already specified as part of the `Host` field.
[jdbcConnect]	unused, already specified as part of the `Connect String` field.

- `DB/Service` The contents of this field is dependent on your database type.

Sybase, MS SQL Server, PostgreSQL, or MySQL	the name of the database you want to connect to.
Oracle	the name of the service to connect to.
Texis	the full path to the remote database, i.e. `C:\morph3\texis\testdb\` or `/var/db/testdb`.
[jdbcConnect]	unused, already specified as part of the `Connect String` field.

- `Username`

  The username to give to the remote database. If this is left blank, username/password will be asked from the user when a request is made. Please see the "DBWalker Authentication Overview" section (pg. 163) for more information.

  If connecting to a Microsoft SQL Server database, it's possible to enter `DOMAIN\user` as the username to use domain authentication, where `DOMAIN` is the domain that the server belongs to.

- `Password`

  The password to give to the remote database. If this is left blank, username/password will be asked from the user when a request is made. Please see the "DBWalker Authentication Overview" section (pg. 163) for more information.

• `Table Information`

The table information section collects information about the table that you want to access.

- `Table`

  the name of the SQL table you want to retrieve data from.

- `Fields`

  An optional list of fields to retrieve from the table. By default, all fields are retrieved. This is specified as a comma-separated list, as you would use in the beginning of a SQL query.

- `Where clause`

  Allows you to limit the data returned by DBWalker. It is not limited to using the fields specified in the `fields` section. The where clause should not contain the SQL keyword `WHERE`, just the conditional clause. For example, if your table has an id and a name, you could set `Fields` to `name` and `Where clause` to `id>100` to only get names of records where the id is greater than 100.

- `Key Field`

  Specifies the "key" field of the database. This field should be able to uniquely identify each record in the table, allowing DBWalker to create a list of links to each record from a single index page. If no key field is specified, the entire contents of the table will be displayed.

4.21.9 Managing DBWalker Stylesheets

XSL Stylesheets allow you to customize the way the DBWalker results are presented via HTML. They are applied either server-side or client-side, as detailed in the `DBWalker Output Overview` section (pg. 163) above.

Here you can can edit or delete stylesheets, or upload new ones. If the DBWalker server is running, you can view any stylesheet by clicking that stylesheet's name.

- `Uploading XSL Stylesheets`

  Beneath the list of current stylesheets is the **Upload Stylesheet**. Here you can browse to a `.xsl` file and upload it to the Search Appliance.

  If the file already exists, the `Overwrite existing` box must be checked when you upload, to make sure you acknowledge the old file will be overwritten.

- `Editing XSL Stylesheets`

  There are two methods for edit XSL stylesheets. The simplest way is to click on the "Edit" button next to a stylesheet. This provides a page with a large text area that contains the contents of the stylesheet. Make your changes and click "Save" to save the changes.

  If you're doing heavy development to a stylesheet, you'll probably find working in a text area very limiting. Alternatively you can download the stylesheet (by clicking on its name on the main DBWalker page), make the changes you want locally with your preferred XSL editor, and re-upload the file. If further tweaks are necessary, simply change the file and re-upload it as much as necessary.

4.21.10 Adding Configurations to Profiles

To get the contents of a DBWalker config searchable, you add it to one or more profiles, where it will be walked with the rest of the profile.

On the `All Walk Settings` page (pg. 56), there is a multi-select box listing all of DBWalker's configurations. Simply select all the configurations you would like to be included in that profile, and they will be walked.

4.22 SOAP API

4.22.1 SOAP Overview

The Simple Object Access Protocol (SOAP) is a W3C (World Wide Web Consortium) recommendation that essentially allows for Remote Procedure Call (RPC) functionality over HTTP, via XML. (This is a simplification of the 120-page SOAP spec, but it suits our purposes). SOAP web services provides a systematic, defined way of communicating function requests and responses over a network transport.

SOAP interfaces are described by another W3C recommendation, WSDL documents - Web Services Definition Language. WSDL documents are the prototypes for SOAP functions. They define what parameters are expected to the functions, what formats are/aren't allowed, what will be returned, etc. Given the WSDL of a SOAP web service, programs can generate the client code that interacts with the services (as is demonstrated in the C# example project later).

Specifically for the Search Appliance, the SOAP interface provides, when using a language that has a SOAP API, a way to invoke a search and on the Search Appliance and insert data as if it were a local function call.

4.22.2 SOAP API vs. XML Output

The SOAP interface provides functionality very similar to the "XML output" search interface. So why use one as opposed to the other?

Use SOAP if the language you are writing has a SOAP interface available for you. Many languages and environments (including Visual Studio) provide SOAP tools, where you provide the WSDL to the webservice, and it will generate "wrapper" classes for you, allowing you to interact with the Search Appliance as if it were simply a local function.

If whatever development environment you're using doesn't have a real SOAP interface, then use the XML API instead of the SOAP API. All the added information/rules of SOAP that make it easy for programs to exchange data will instead make it more cumbersome to use manually.

4.22.3 Getting the WSDL

The WSDLs can be found on the `Soap Tools` page for the profile. It links to the Dataload WSDLs and a search WSDLs, which lets you choose either a WSDL for this profile, or for all profiles (as explained below).

4.22.4 Global vs. per-profile WSDLs

When viewing search WSDLs, you have the option of requesting a WSDL specific to a single profile, or a global `All Profiles` WSDL, which can be used for any profile.

If you don't make use of Additional Fields, then there will be no difference between per-profile and global WSDLs.

Both per-profile and global WSDLs refer to the same search interface. The same SOAP response is generated for both WSDLs. The only difference is in how specific the WSDLs are - per-profile WSDLs specify which Additional Fields occur in the results, but the global WSDL must use `<xsd:any>` as a catch-all, as the Additional Fields may change from one profile to another.

Which you use is a trade-off that you must decide on.

- **per-profile WSDLs**
  - **Advantage** Additional Fields for the profile are "hard-coded" in the WSDL itself, so a SOAP client consuming the WSDL can make better use of the Additional Fields.

    For example, if your profile has Additional Fields called `price` and `location`, then a per-profile WSDL will specify that each result contains `<price>` and `<location>` elements. WSDL tools can do things like declare `response.price` and `response.location` variables.
  - **Disadvantage** Because the per-profile WSDL is specific to that profile's Additional Fields, a different WSDL must be used for every profile you want to interact with. If you're interacting with many different profiles (or it often changes), an global WSDL may be better suited.
- **Global WSDLs**
  - **Advantage** The `All Profiles` wsdl can be used for any profile. This is better if your application needs to query multiple profiles, or if you don't work with Additional Fields.
  - **Disadvantage** Additional Fields are represented in the `All Profiles` WSDL with `<xsd:any>`, which allows it to not declare which Additional Fields will occur in the XML (as it may change from one profile to another).

    This means that programs consuming the WSDL cannot know which parametric fields will be returned, and will instead do things like offer an array of Additional Field XML elements that you must manually loop over to find the ones you want.

4.22.5 Configuring the SOAP Interface

The WSDL for the Search Appliance is accessible in the following URL path from the Search Appliance:

`/texis/ThunderstoneSearchService/describe.wsdl`

This link is also available from the **Tools -¿ SOAP Tools** menu page in the admin interface.

Dataload SOAP API

The Dataload SOAP API takes the same parameters as the normal dataload API, please see the `Submission Format` (p. 155) and `Reply Format` (p. 160) sections, with a few exceptions:

- The entire transactions are wrapped by SOAP envelopes and the top-level elements are called `dataload` and `dataloadResponse` instead of `ThunderstoneReplicationResult`, respectively.
- The `dataload` element contains a `profile` element in addition to all the `Item`s.

4.22.6 C# example project

A C# example project is available that demonstrates using the search SOAP interface. In the **System -¿ Advanced Tools** of the administration interface, choose `Extra Downloads`, and then `Thunderstone Soap Example`. Instructions are listed on that page and within the zip itself for how to use the project.

4.22.7 SOAP Links for Languages

This section contains links for recommendations of SOAP implementations in other languages. Thunderstone makes no guarantees to the completeness or quality of these projects, we simply provide links for convenience.

- ASP.NET - the same C# API code can be compiled into a .NET assembly and used from an ASP.NET page. Please see the Windows Communication Foundation documentation for more details.
  - `http://msdn.microsoft.com/en-us/library/dd456779.aspx`
- Perl - SOAP::Lite for Perl is a collection of Perl modules which provides a simple and lightweight SOAP interface.
  - `http://www.soaplite.com/`
- Python The Web Services for Python Project provides libraries implementing the various protocols used when writing web services including SOAP, WSDL, and other related protocols.
  - `http://pywebsvcs.sourceforge.net/`
- Java - The Java API for XML Messaging (JAXM) provides a framework for sending and receiving SOAP messages.
  - `http://java.sun.com/webservices/jaxm/`
- C++ - Apache Axis C++ can be used as a client for SOAP servers.
  - `http://axis.apache.org/axis/cpp/clientuser-guide.html`

4.22.8 SOAP API search Reference

Three of the functions for searching (`search`, which performs a normal search, `moreLikeThis`, which finds similar pages, and `showParents`, which shows which pages link to a page) take similar parameters:

- `jump` - number of user-visible results to skip. 0 would return the first page of results, 10 the second page, etc. (assuming 10 results per page).
- `order` - specifies how the results should be ordered. Possible values are: `r` - sort by relevance (default) `dd` - newest pages first `da` - oldest pages first

RankKnobs structure

There's an optional "rankKnobs" parameter for many of the functions that can specify how things should be ranked (each function notes whether it accepts rankKnobs). All of these can be set from 0-1000, where the higher the value, the more heavily that aspect is weighed; 500 is the default. These parameters correspond directly to the "Ranking Factors" settings on the Advanced Search page.

RankKnobs has the following parameters:

- `order` - importance of the words being in the proper order
- `proximity` - importance that the words are close together
- `dbFreq` - importance of the frequency of a word in the database
- `docFreq` - importance of the frequency of a word within the document
- `leadBias` - importance of closeness to the start of the document

search

This performs a normal search based on the query/queries provided.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `query` - Required. The Metamorph query to search for.

The following parameters can be provided to further refine the search:

- `urlQuery` - Used for URL Prefix queries. Corresponds to the default search interface's `uq` query-string variable (p. 141).
- `depthQuery` - the maximum depth that would be allowed. Supplying a value of `3` would only return pages that are no more than 3 clicks away from a Base URL. Corresponds to the default interface's `dq` variable.

- `categoryQuery` - numeric index for a category to use. 1 is the first category, etc. Corresponds to the default interface's `cq` variable.
- `proximity` - specifies a required proximity for the words in the query. Corresponds to the default interface's `prox` variable. Possible values are:
  - `page` - words must occur on the same page (default)
  - `paragraph` - words must occur in the same paragraph
  - `sentence` - words must occur in the same sentence.
  - `line` - words must occur on the same line.

The `search` function may also use the rankKnobs structure (section 4.22.8, p. 172).

The `search` function may also take a number of Additional Field parameters, as described in the `Searching Additional Fields` section (p. 63).

The output of the function is described in the XML Elements in Search Results section (4.13.3, p. 143).

moreLikeThis

`moreLikeThis` returns results that are similar to a result already found.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `id` - Required. the id of a single URL, returned from a previous function.

The `moreLikeThis` function may also use the `<rankKnobs>` structure, as described in section 4.22.8, p. 172.

The output of the function is described in the XML Elements in the Search Results section (4.13.3, p. 143).

showParents

`showParents` lists all the pages that link to a previous retrieved search results.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `id` - Required. the id of a single URL, returned from a previous function.

The output of the function is described in the XML Elements in the Search Results section (4.13.3, p. 143).

getCompletions

`getCompletions` retrieves potential completions for a partially typed query. It can be used by a custom client to implement autocomplete as the user is typing.

Parameters:

- `profile` - Required. Specifies which profile you're working with.
- `term` - Required. The partially typed term to get completions for.

Output: Zero or more `Completion` elements are returned. Their text content is a potential completion, and each completion has attributes:

- `profile` - the profile this completion came from.
- `score` - the score for this completion. Completions are ranked according to score, which can be derived from occurrences in the content, queries, and search user clicks.

4.22.9 SOAP API admin Reference

login

Parameters:

- `username` - the user being logged in
- `password` - the password for the user

Returns:

- `authToken` - an authentication token for use in further requests

`login` logs you in to the appliance, supplying you with an authentication token that will be included in all further requests to show that you're logged in. All other Admin SOAP API calls require an authToken for use. It's sent in further requests via a SOAP Header.

listProfiles

Parameters:

- *none*

Returns:

- `ProfileName` - an array of profile names requests

Returns a list of all profiles that currently exist on the appliance. If no profiles exist, a successful response with no `ProfileName`s is returned.

getProfileStatus

Parameters:

- `Profile` - name of the profile

Returns:

- `IsRunning` - whether or not the walk is currently running, set to `true` or `false`.

Returns information about the profile, currently just whether or not the profile is running.

addProfile

Parameters:

- `Profile` - name of the new profile
- `Type` *(optional)* - the type of profile, `standard` or `metasearch`. Defaults to `standard`.
- `CopyOf` *(optional)* - name of the profile to copy
- `ParametricField` - *unused in the Search Appliance*
- `PrimaryKey` - *unused in the Search Appliance*
- `Dataspace` - *unused in the Search Appliance*

Returns:

- `Success` - will be set to `ok`, indicating the profile was created successfully.

Adds a new profile to the Search Appliance. If there's any problem (already exists, invalid profile name, etc), a SOAP Fault will be thrown.

deleteProfile

Parameters:

- `Profile` - name of the profile to be deleted

Returns:

- `Success` - will be set to `ok`, indicating the profile was deleted successfully.

Deletes a profile from the Search Appliance. If the profile didn't exist, the call will still succeed.

getSettings

Parameters:

- `Profile` - name of the profile. To get System Wide Settings, use `!SYSTEM`
- `Name` *(optional)* - array of setting names to get. If no `Name` is provided, all settings are returned.
- `TestOrLive` *(optional)* - whether to return the test settings, or live settings. Returns live settings by default.

Returns:

- `Setting` - an array of name/value pairs for the requested settings. A `TestOrLive` attribute on the settings indicate whether this applies to test, live, or both.

Gets a list of settings for the requested profile. You can request one or more specific settings by passing in `Name` parameters, or get all settings by not supplying a `Name`.

Some settings have test and live versions. You can request which version you'd like (defaults to live), and the returned settings indicate whether they apply to test or live. "Both" indicates that setting doesn't have different test and live versions.

setSettings

Parameters:

- `Profile` - name of the profile. To set System Wide Settings, use `!SYSTEM`
- `TestOrLive` *(optional)* - whether this should apply to test settings, live settings, or both. Defaults to both.
- `Setting` - multiple name/value pairs of settings that you'd like to set

Returns:

- `Success` - set to `ok`, indicating the settings were set properly.

Applies an array of settings for the given profile.

If there is any problem (such as an invalid setting name) in any one of the settings, a SOAP Fault is returned, and NONE of the settings are applied. This allows you to tweak the problem settings, and re-submit the entire batch again, without having them "partially applied" in between.

getTask

Parameters:

- `Id` - id of the task to get

Returns:

- `TaskInfo` - a structure of information about the task
  - `Id` - id for the task, can be used in further requests
  - `Pid` - Process ID of the task, may be `-1` if task hasn't started
  - `Profile` - Profile for the task, or `!SYSTEM` for system-wide tasks
  - `Action` - the task's main action, defines its behavior (dispatch, updateindex, etc)
  - `Status` - the task's current status, possible values are:
    * `queued` - task is waiting to be run by the task monitor
    * `starting` - task has been launched and is starting up
    * `running` - task is running
    * `finished` - task completed successfully, see `Detail` may contain more information about its exit
    * `incomplete` - task exited before completing everything it could, `Detail` explains the reason for early exit (user canceling a walk, etc)
    * `died` - task ended abruptly and did not unregister itself, or the process was killed by another program.
  - `Db` - notes which of the internal databases the task is operating on for the profile. May be `!NONE` for system wide tasks, or `!UNKNOWN` if the task is still starting and hasn't set a database yet.
  - `DbRole` - currently unused
  - `ParentType` - defines the type of parent recorded for launching this task. See `ParentData` below for values and their meaning.
  - `ParentData` - Additional information about the process that launched this task, its content depends on the `ParentType`:
    * If `ParentType=task`, then `ParentData` contains the id of the task that launched this task.
    * If `ParentType=web`, then `ParentData` contains the name of the Appliance function that was invoked through the web to launch this task.
    * If `ParentType=cli`, then `ParentData` contains the name of the Appliance function that was invoked through the command line to launch this task.
    * If `ParentType=unknown`, then `ParentData` is the process id that launched this task.
  - `Args` - The arguments that were used to launched this task, can occur multiple times for multiple arguments.
  - `Started` - The dateTime this task was started. It will not be present for tasks that are queued or are still starting.

- `Updated` - The dateTime this task last updated its information. Most tasks should update regularly, although may go long periods of time without updating during long operations.
- `NextRun` - currently unused
- `Schedule` - currently unused
- `Description` *(optional)* - Additional information on what the task was launched to do. Some actions don't need a description (like `updateindex`), but others can provide additional information (`walk` lists its Base URL, etc).
- `Detail` - For running tasks, `Detail` may contain additional information about what the task is currently doing (building indexes, etc).

  For finished tasks, `Detail` may contain additional information about the reason the task exited.

  `ProgressInfo` - Some tasks provide additional progress information, such as walks or replication senders.

  * `Attempted` - the number of items attempted (pages attempted to be fetched, items sent for replication, etc).
  * `Saved` - the number of successful items
  * `Errors` - the number of items that encountered an error
  * `Bytes` - the number of bytes transferred or processed
  * `TodoCurrent` - the number of items to do at the current depth
  * `TodoNext` - the number of items to do at the next depth
  * `LoadAttempted1Min` - the average items per minute attempted, weighted towards the last minute
  * `LoadSaved1Min` - the average items per minute saved, weighted towards the last minute
  * `LoadAttempted15Min` - the average items per minute attempted, weighted towards the last 15 minutes
  * `LoadSaved15Min` - the average items per minute saved, weighted towards the last 15 minutes
  * `LoadAttempted1Hour` - the average items per minute attempted, weighted towards the last hour
  * `LoadSaved1Hour` - the average items per minute saved, weighted towards the last hour

`getTask` retrieves information about a specific task that is running or has been run. The id likely comes from a previous `getTasks` call, where checking again can be done more efficiently by giving the id exactly rather than the same criteria to `getTasks`.

getTasks

Parameters:

- `Profile` - the name of the profile to get tasks for, or `!SYSTEM` for system-wide tasks, or `!ALL` for all tasks.
- `Scope` *(optional)* - defines what tasks to get. Possible values are:

  - `running` - return only running tasks. This is the default when using `!SYSTEM` or `!ALL` for Profile.
  - `recent` - return tasks since the most recent crawl or import. This is the default for individual profiles.
  - `!ALL` - all tasks, running and finished.

- `Db` *(optional)* - which profile database to retrieve tasks for. Possible values are:
  - `live` - the profile database currently being used for search.
  - `other` - if a `New` walk is running, this retrieves tasks for the walking database, not the live search database.

  The `Db` parameter is unused with Profile values of `!SYSTEM` and `!ALL`.

- `Max` *(optional)* - the maximum number of tasks to return, defaults to `-1` (unlimited).

Returns:

- `TaskInfo` - an array of structures of information about the task, see `getTask` for details.

`getTasks` retrieves information about tasks that are running or have been run. This can be used to keep tabs on what is running on a machine, or to tell when a specific profile has all of its tasks completed.

getThesauruses

Parameters:

- *none*

Returns:

- `Thesaurus` - an array of Thesaurus information
  - `Name` - name of the thesaurus
  - `Permutations` - what permutations apply to this thesaurus
  - `NumProfileUsing` - the number of profiles that are currently using this thesaurus

Returns information about all thesauruses that exist in the Search Appliance.

setThesaurus

Parameters:

- `Name` - name of the thesaurus. If this thesaurus already exists, it will be replaced.

- `Permutations` - what permutations apply to this thesaurus. Possible values are `Full`, `Single`, or `None`. Defaults to `Single`.
- `Verbose` *(optional)* - If set to `Y`, verbose output of processing the thesaurus content will be included in the response.
- `Content` - the text content that should be used for the thesaurus. See the Thesaurus section for details on the format (4.3, p. 135).

Returns:

- `Output` - output of the thesaurus processing operation. Any errors will be listed in the text.

Creates or updates a thesaurus in the Search Appliance. Once created, a thesaurus can be used in a profile by setting its `SSc_eqprefix` or `SSc_ueqprefix` to this thesaurus' `Name`.

deleteThesaurus

Parameters:

- `Name` - name of the thesaurus to delete

Returns:

- `Output` - output of the thesaurus deletion operation. Any errors will be listed in the text.

Deletes the thesaurus `Name`. Any profiles using this thesaurus will have their setting properly cleared.

4.23 Thunderstone ISAPI Proxy Module

4.23.1 Overview

The Thunderstone ISAPI Proxy Module is an IIS add-on that allows you to proxy requests through another machine.

Users' search requests are not made directly to the Search Appliance, but to the Proxy Module, which then passes the request along to the Search Appliance.

The Proxy Module also has an optional AuthProxy, which allows the use of automatic Active Directory credentials when authenticating the Search Appliance search results. Internet Explorer will be configured to automatically pass along the authentication information, allowing for Single Sign On. The Search Appliance communicates with the Proxy Module to authorize the results.

4.23.2 Requirements

For the Proxy Module:

The Thunderstone Proxy Module must be installed on a machine with IIS.

For the Proxy Module with the authProxy:

The Proxy Module with authProxy machine will need to be added to Internet Explorer's `Local Internet` zone, so if you have a server already listed there, you may want to use it.

This must be a separate machine from the Search Appliance, and the machine the Proxy Module is installed on must be a member of the Active Directory domain.

Ensure that the machine that the authProxy is being installed on is a secure machine. Because the Proxy Module is dealing with authorization functionality, a user with Administrative privileges could potentially tamper with operations.

4.23.3 Installing the Proxy Module

Before installing the proxy module the only thing you need to know is:

- The full hostname of the Search Appliance machine (e.g. `thunderstone.mysite.com`) that the Proxy Module will be communicating with.

You can download the installer that contains the Proxy Module and authProxy from the Search Appliance machine. In the **System** -¿ **Advanced Tools** of the administration interface, choose `Extra Downloads`, then `Thunderstone Proxy Module`, and finally click the `Download setupProxyModule.msi` link for the installer. Once downloaded, the installer must be run on the Windows machine that you wish to make the proxy.

When installing you will be asked for a few items:

- `Destination Location` - This is where the actual DLL for the proxy module and its supporting files are placed.
- `Target` - The full hostname of the Search Appliance machine that this Proxy Module should connect to.

4.23.4 Post-Install Setup

If you're using the authProxy, there are some configuration steps that must be manually performed, as they occur on machines other than the Proxy Module's machine. Please perform these before attempting to use the Proxy Module with authProxy.

Grant "Trust for Delegation" to the proxy machine

The machine that runs the Proxy Module & authProxy must be marked as `trusted for delegation` by the Active Directory domain controller. This is necessary for the proxy to automatically "pass along" the users' authentication to the searched web sites.

- **On the domain controller**, go to `Start` - `Programs` - `Administrative Tools` - `Active Directory Users and Computers`.
- Choose `Computers` on the left.
- Locate the computer that is running the Proxy Module, right-click on it, and choose `Properties`.
- Check `Trust computer for delegation`. A message box warning you that "this is a security-sensitive operation and it should not be done indiscriminately" will pop up. Click `OK`.
- Click `OK` to close the machine's properties, and close the `Active Directory Users and Computers` window.

Configuring Internet Explorer for Passing Credentials

The Proxy Module & authProxy machine must be listed in Internet Explorer's `Local Internet` security zone for all computers using it in order to function properly. If it is not in the `Local Internet`, then credentials will not be automatically provided. Even if the credentials are entered manually, the Proxy Module cannot authenticate with results when not listed in Internet Explorer's `Local Internet`.

If the Proxy Module machine is already in the `Local Internet` settings, you may skip this step.

The following steps adds the Proxy Module machine to Internet Explorer's `Local Internet`:

- Start Internet Explorer.
- Choose `Tools` from the menu, and select `Internet Options`.
- Choose the `Security` tab.
- Choose `Local Internet` from the list of zones.
- Click the `Sites` button to edit the local internet.
- Click `Advanced` to manually add a site.
- *Un*check `Require server verification (https:) for all sites in this zone`
- Enter the full hostname of your proxying machine, for example `proxyMachine.example.com`.
- Click `Add` to add the site to the Trusted Sites.
- Click `Close` to close the `Advanced` window.

- Click `OK` to close the `Local Intranet` window.
- Click `OK` to close the `Internet Options` window.

Internet Explorer is now configured to pass credentials to the proxy machine. This is a per-user configuration, and will need to be configured for any user that is authenticating via the Proxy Module.

Configuring the Search Appliance

There are three things that must be done in the Search Appliance to configure it to accept authentication information from the authProxy, one of them global and two on a per-profile basis.

Add the Proxy Machine to Cluster Members

The IP address of the machine that the authProxy is installed on must be added to the list of `Cluster Members` to tell the Search Appliance to trust the proxy machine.

Go to `System-¿System Wide Settings`, and enter the proxy machine's IP address in the `Cluster Members` field on a new line.

Make the Target Profiles Visible

The profiles that you want to search with the authProxy must be set `Visible`, which enables the profile for things like meta searching and the proxy module.

- Select the profile in the `Profiles` page.
- Choose `Search Settings` on the left.
- Set the `Visible` setting to `Y`.
- Click `Update` at the bottom.

Enable Results Authorization for the Target Profile

Also, Results Authorization must be enabled for the target profile, if it's not already enabled.

- Select the profile in the `Profiles` page.
- Choose `Search Settings` on the left.
- Set the radio button for `Authorization Method` to `Basic/NTLM/file` (occurs beneath `Login Cookies` and `Login URL`).
- Click `Update` at the bottom.

4.23.5 Manually Configuring the Proxy Module

This section describes how to manually configure IIS for use of the Thunderstone Proxy Module. The will be described in more detail in the next section. This is **not** necessary for normal operations - these actions are normally performed automatically by InstallShield upon installation. These steps are only necessary if IIS's configuration gets wiped out and needs to be redone.

The Thunderstone Proxy Module is an ISAPI Extension, two if using the authProxy. They are assigned as Global Application Maps to Virtual Directories in IIS. All requests to the directories are not be served from the file system that the virtual directory points to, but instead go through the Proxy Module `dll`s.

One virtual directory is required per extension: `texis`, which gets assigned `proxyModule.dll`, and `authProxy`, which gets assigned `authProxy.dll`.

If using the authProxy, `texis` must have anonymous access disabled and Integrated Authentication enabled, while `authProxy` must have anonymous access allowed (which is allowed by default).

These are the steps that must be done if you are manually setting up IIS for using the Proxy Module. **Note that these are done automatically by the installer** and do *not* need to be manually done under normal circumstances.

- Open the IIS Configuration
  - Right click on `My Computer` on the desktop.
  - Select `Manage...`
  - Open `Services and Applications` in the tree.
  - Open `Internet Information Services`.
  - Open `Web Sites`.
  - Select the web site you want to add the Proxy Module to (most likely `Default Web Site`).
- Add the `texis` virtual directory
  - Right click on the web site and select `New -> Virtual Directory...`
  - The `Virtual Directory Creation Wizard` opens. Click `Next>`.
  - In the `Alias` box, enter `texis` and click `Next>`.
  - In the `Path` box, enter the real physical path you want the virtual directory to map to, and click `Next`. the Proxy Module uses the directory `<INSTALLDIR>/etc/ISAPI-virtualdir` by default.

    Note that it doesn't matter what directory is selected. This directory will never be used because all requests will be intercepted by the Proxy Module. The only reason a directory must be selected is because IIS insists that *all* virtual directories map to a real physical location.
  - At the `Virtual Directory Access Permissions` screen, just click `Next` to complete the wizard, as we won't be using any of the permissions.
  - Click `Finish` to complete the wizard and return to the `Computer Management` window.
- Apply `proxyModule.dll` as a Wildcard Application Map

  - Right-click on the newly created virtual directory and select `Properties`.
  - The lower half of the properties window is labeled `Application Settings`. Click `Create` to make a custom set of application settings for this virtual directory.
  - After clicking `Create`, the `Configuration` should no longer be disabled. Click `Configuration`.
  - The lower half of the new `Application Configuration` window details Wildcard Application Maps, which is currently empty. Click `Insert`.
  - Next to the `Executable` field, click the `Browse` button and locate `ProxyModule.dll`, which is in the directory you installed the Proxy Module to.
    * (The default location for this file in `C:\windows\system32\inetsrv` on 32bit windows, `C:\windows\SysWOW64\inetsrv` on 64bit windows).
  - **Uncheck** the box next to `Verify that file exists`, and click `OK`.
  - `ProxyModule.dll` will now be in the list of Wildcard Application Maps. Click `OK` to close the `Application Configuration` window.

- Configure `texis` for authentication

  Only necessary if using the authProxy.

  - While still in the `texis Properties` window for the new virtual directory, Select the `Directory Security` tab.
  - In the top section, labeled `Authentication and Access Control`, click the `Edit...` button.
  - **Un**check `Enable Anonymous Access` and ensure that `Integrated Windows Authentication` is **checked**.
  - Click `OK` to close the `Authentication Methods` window.
  - Click `OK` to close the `texis Properties` window.

  Now we need to create the `authProxy` directory in a similar manner, although it doesn't need anonymous access disabled.

  If you're not using the authProxy, please skip to the "Add the Proxy Module files to IIS' list of allowed extensions" section below.

- Add the `authProxy` virtual directory

  - Right click on the website and select `New -> Virtual Directory...`
  - The `Virtual Directory Creation Wizard` opens. Click `Next>`.
  - In the `Alias` box, enter `authProxy` and click `Next>`.
  - In the `Path` box, enter the real physical path you want the virtual directory to map to, and click `Next`. the Proxy Module uses the directory `<INSTALLDIR>/etc/ISAPI-virtualdir` by default.

    Note that it doesn't matter what directory is selected. This directory will never be used because all requests will be intercepted by the Proxy Module. The only reason a directory must be selected is because IIS insists that *all* virtual directories map to a real physical location.

  - At the `Virtual Directory Access Permissions` screen, just click `Next` to complete the wizard, as we won't be using any of the permissions.
  - Click `Finish` to complete the wizard and return to the `Computer Management` window.

- Apply `authProxy.dll` as a Wildcard Application Map

  - Right-click on the newly created `authProxy` virtual directory and select `Properties`.
  - The lower half of the properties window is labeled `Application Settings`. Click `Create` to make a custom set of application settings for this virtual directory.
  - After clicking `Create`, the `Configuration` should no longer be disabled. Click `Configuration`.
  - The lower half of the new `Application Configuration` window details `Wildcard Application Maps`, which is currently empty. Click `Insert`.
  - Next to the `Executable` field, click the `Browse` button and locate `authProxy.dll`, which is in the directory you installed the Proxy Module to.
    * (The default location for this file in `C:\windows\system32\inetsrv` on 32bit windows, `C:\windows\SysWOW64\inetsrv` on 64bit windows).
  - **Uncheck** the box next to `Verify that file exists`, and click `OK`.
  - `authProxy.dll` will now be in the list of Wildcard Application Maps. Click `OK` to close the `Application Configuration` window.

- Add the Proxy Module files to IIS' list of allowed extensions

  By default IIS blocks all ISAPI extensions as a security measure. The Proxy Module must be explicitly allowed in IIS' configuration.

  - Back in the `Computer Management` window, open `Web Service Extensions`, underneath `Internet Information Services`.
  - The right side of the window should now have a list of rules. Right-click beneath the existing rules and select `Add a new web service extension...`
  - In the `Extension Name` field, enter `Thunderstone Proxy Module`.
  - Next to the `Required files` text area, click the `Add...` button.
  - Next to `Path to file:`, click `Browse...` and locate `ProxyModule.dll`, (just as in the previous set of instructions), and click `OK` to close the `Add File` dialog.
  - If using the authProxy, click `Add` again, and this time choose the `authProxy.dll` file.
  - Check the box next to `Set extension status to Allowed`, and click `OK` to close the window.

IIS is now set up properly to use the Proxy Module. Note that if using the authProxy, changes still need to be made to the network and the Search Appliance, as detailed in the **Post-Install Setup** and **Configuring the Search Appliance** sections, on pages 181 and 183, respectively.

4.23.6 Troubleshooting the Proxy Module Authentication

This section details some troubleshooting steps you can go through if Proxy Module Authentication isn't working.

Review Installation Steps

There are a number of steps that must be manually performed after the Proxy Module install (due to them being done on different machines or as different accounts). Please ensure the following steps have been performed:

- Grant "Trust for Delegation" to the proxy machine (p. 182)
- Configuring Internet Explorer for Passing Credentials (p. 182)
- Configuring the Search Appliance (p. 183)

Machine names and SPNs

A Service Principle Name (SPN) is the name by which a client uniquely identifies an instance of a service. By default your IIS machine has SPNs for its hostname, such as `myServer`, and its Fully Qualified Domain Name (FQDN), such as `myServer.branch.example.com`.

If the proxy machine is accessed by a name other than either of these, such as `myServer.example.com`, `otherName.company.com`, or its IP address, then Active Directory authentication will not work. Your choices are:

- Access the machine using either its host name or FQDN.
- Register an additional SPN for the proxy machine on the domain controller.

SPNs can be viewed and changed with the `setspn.exe` tool, which Microsoft provides as part of its `Windows 2003 Support Tools`. This package is available on the Windows 2003 CD, or, at the time of writing, at `http://go.microsoft.com/fwlink/?LinkId=100114`.

DelegConfig Diagnostic Tool

For general Active Directory troubleshooting, Thunderstone has found the `DelegConfig` tool to be handy. It's an ASP.NET application used to help troubleshoot and configure IIS and Active Directory to allow Kerberos and delegating Active Directory credentials. At the time of this writing, it is available at:

`http://www.iis.net/downloads/default.aspx?tabid=34&g=6&i=1434`

Thunderstone did not create DelegConfig, and does not make any guarantees to its accuracy or availability.

Launch IE as a different user

When testing multiple user accounts, you can Right-Click on an Internet Explorer shortcut, select `Run As...`, and enter another user's credentials to launch Internet Explorer as that user. This way you can try out IE as them without needing to go through the full login process.

- Note that the user you run IE as must have logged in at least once in order to properly configure IE, otherwise the `Add to Local Internet` dialog will not function properly.
- If there is no `Run As...` option, create a new shortcut to Internet Explorer and use that - the IE icon on that's on the desktop by default is not a normal shortcut.
- You **cannot** test another user by entering their credentials within the browser when accessing the proxy machine. Credentials MUST be passed to the proxy machine automatically by IE by having the proxy machine in IE's `Local Internet` settings.

  If credentials are entered manually in IE when accessing the proxy machine (even the credentials of the current user), then IE will use NTLM to authenticate with the proxy machine instead of Active Directory, which prevents proper delegation from occurring.

Chapter 5

Reference

5.1 Supported File Formats

- Adobe Acrobat - **.pdf** (Versions 1-10)
- Ami Professional - **.sam** (Versions 1.0-3.1)
- ASCII- **varies** (Any plain text format. Mime text/plain)
- Atom feeds **.atom, .xml**
- Autocad - **.dwg** (2.5-2006 (Appliance Only))
- Bzip2 - **.bz2** (Decompress and process supported sub-files)
- CCMail (All Versions)
- Compress - **.Z** (Decompress and process supported sub-files)
- CTOS DEF
- DG CEOWrite(3.0)
- dBase - **.dbf** (All Versions)
- DCIMARC
- DEC WPS-Plus - **.wpl** (through 4.1)
- Enable - **.wpf** (1.0 through 2.15)
- FoxPro - **.dbf** (dBase workalike)
- FrameWork - (III 1.0, 1.1, IV)
- GIF - **.gif** (textual meta data only)
- Gzip - **.gz** (Decompress and process supported sub-files)

- FrameWork - (III 1.0, 1.1, IV)
- Harris Typesetter
- HTML pages - **.htm, .html, .php, .asp, .cfm, etc.** (All Versions)
- IBM Writing Assistant - **.iwa, .wrt** (All Versions)
- Interleaf (5.2, 6.0)
- JPEG images- **.jpg, .jpeg**
- Legacy - **.leg** (1.x, 2.0)
- Lotus 1-2-3 - **.wks, .wk1-.wk4** (Versions 1A, 2.0 through 5.0)
- MacWrite II - **.mcw, .mw** (1.0, 1.1)
- MacWrite Pro - **.mcw, .mw** (1.0)
- MS Excel Spreadsheets - **.xls .xlsx**
- MS Help files - **.hlp** (only on Windows with ”helpdeco”)
- MS CHM files - **.chm** (only on Windows with ”hh”)
- MS Office
- MS Outlook emails - **.msg** (All Versions)
- MS Powerpoint presentation - **.ppt .pptx** (through 2007)
- MS Transport Neutral Encoding Format - **TNEF** (All Versions)
- MS Word Documents - **.doc .docx**
- MS Write - **.wri** (3.x)
- OfficePower - **.op6, .op7** (6.0, 7.0)
- OfficeWriter - **.ow4, .ow5, .ow6** (4.0, 5.0, 6.0, 6.1)
- Open Document - **.odt, .ods, etc**
- PeachText 5000 - **.pea** (Version 2.12)
- PFS: First Choice - **.pfs** (1.0-3.0)
- PFS: Write - **.pfs** (Version C)
- Plain text - **.txt** (All Versions)
- PostScript - **.ps** (All Versions)
- Professional Write - **.pw** (1.0, 2.1, 2.2)

- Professional Write Plus - **.pw, .pwp** (1.0)
- Q&A for DOS - **.qa, .qw, .dtf** (2.0)
- Q&A Write for Windows - **.dtf** (3.0)
- Rapidfile Memo Writer - **.mmo** (1.0, 1.2)
- Rar - **.rar** (Decompress and process supported sub-files)
- RFC882 Mail (All versions)
- RSS Feeds **.rss, .xml**
- SGML files - **.sgml** (All versions)
- Shockwave/Flash - **.swf** (All Versions)
- Tagged Image File Format (TIFF) - **.tif, .tiff** (Meta data only)
- Tar - **.tar** (Extract and process supported sub-files)
- Total Word - **.tw** (1.2, 1.3)
- Uniplex onGo (v7)
- Usenet News
- Vines Mail
- Volkswriter - **.vw, .vw3** (3.0, 4.0)
- Wang WITA - **.iwp** (through 2.6)
- Wiziword - **.doc** (All Versions)
- Wordpad document - **.rtf** (All Versions)
- Word Perfect - **.wpd** (4.1 through 6.1, +Mail Merge)
- Word Perfect Mac (1.0 through 3.1)
- Wordstar - **.ws, .wsd** (3.3 through 7.0)
- Wordstar 2000 - **.ws2** (3.0, 3.5)
- WriteNow - **.wn** (3.0)
- XML - **.xml** (Process as if HTML)
- XyWrite - **.xy, .xy3, .xy4** (III, III Plus, IV)
- XyWrite for Windows - **.xyw** (4.0)
- Zip - **.zip** (Decompress and process supported sub-files)

5.2 Database and File Usage

The Search Appliance maintains a database that contains text from HTML pages, links to other pages, and a list of categories.

When the Search Appliance 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 called `db1` and `db2`. The Search Appliance 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.

The Search Appliance 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`

Each setting has a record in the `options` table of the default database. See section 5.4 (p. 195) 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`.

5.3 Walk Database Tables and Fields

Table 5.1: Fields in `html` table

Field	Description
`id`	Unique record id
`Hash`	Document hash for duplicate content detection
`Size`	Size of retrieved raw document (i.e. HTML)
`Visited`	The date the page was modified (or fetched if modified not set)
`Dlsecs`	The number of seconds needed to fetch the page
`Depth`	The number of URLs traversed to reach the page
`Url`	The URL of the real HTML page
`Title`	The title of the page
`Body`	The formatted textual content of the page, in Storage Charset (UTF-8)
`Keywords`	The `keywords` meta data from the page
`Description`	The `description` meta data from the page
`Meta`	Other meta data from the page, separated by newlines
`Catno`	List of categories to which the URL belongs
`CatnoLowest`	Lowest `Catno` value
`Modified`	The date the page was modified
`NextCheck`	The date the page should next be refreshed
`Views`	The number of times this URL has been viewed (shown in results)
`Clicks`	The number of times this URL has been clicked (in results)
`CTR`	Click-through ratio
`Pop`	Popularity (number of pages linking to this page)
`MimeType`	MIME type of original page
`Charset`	Character set of page as stored (usually Storage Charset)

Table 5.2: Fields in `refs` table

Field	Description
`Url`	The URL of the HTML page
`Ref`	The URL of a reference (link) on the HTML page

Table 5.3: Fields in `categories` table

Field	Description
`Catno`	The number for the category
`OverlapsLower`	Y if some member(s) also in a lower category
`Url`	The URL pattern for the category
`Category`	The name of the category

Table 5.4: Fields in `error` table

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

Table 5.5: Fields in `querylog` table (if query logging enabled)

Field	Description
`id`	Contains the date and time of the query (unique record id)
`Client`	The hostname of the web client that performed the query
`Query`	The user's query as entered

5.4 Options Table Fields

These are the options table fields (maintained in the default database):

Table 5.6: Fields in `options` table

Field	Description
`id`	Unique id for the record
`Profile`	The name of the profile that the record belongs to
`Name`	The name of the setting
`Type`	The data type of the setting (always `String`)
`String`	The value of the setting
`Int`	Unused
`Float`	Unused
`Strlist`	Unused

5.5 Customizing the Search

You may make common changes to the Search Appliance's search appearance by using `Search Settings` from the administrative interface main menu.

5.6 Customizing the Walker

You may make many changes to the Search Appliance's walk behavior by using `Walk Settings` from the administrative interface main menu.

Chapter 6

Search Interface Help

6.1 Forming a Query

The Search Appliance'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`

6.1.1 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?

6.1.2 Overview of Query Abilities

The Search Appliance is based on Texis and as such it shares its text query abilities with all of Thunderstone's products. Throughout our documentation you will see references to Metamorph or Texis. This is because all of our products share a common text query language. This document provides only a brief overview of this language.

If you'd like to know more see the online manual at `http://www.thunderstone.com/site/texisman/link_mmq.html`.

6.1.3 Controlling Proximity

Mastering the usage of proximity gives the ability to locate answers with greater precision. The Search Appliance 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 (e.g. all searches except `Show Parents`).

6.1.4 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.

6.1.5 Keywords Phrases and Wild-cards

To locate words, just type them in as you would in a word processor. Letter cases will be ignored.

The wild-card character `*` (asterisk) may be used to match just the prefix of a word or to ignore the middle of something.

If the item you wish to locate is more complicated than the simple `*` wild-card can accomplish, try using the regular expression matcher (http://www.thunderstone.com/texis/site/pages/regexp.html).

To locate a number of adjacent words in a specific order, surround them with `"` (double quotation) characters. Putting a `-` (hyphen) between words will also force order and one word proximity.

* see `Word Forms` (6.2, p. 201)

Table 6.1: 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, ... *

6.1.6 Applying Search Logic

Texis and Metamorph use set logic for text queries. Set logic is easier to use and provides more abilities than boolean. The examples below make reference to single keywords, but keep in mind that each keyword can represent an entire list of things or any of the special pattern matchers.

Sets (or lists) of things are specified by placing the elements within parenthesis, separated by commas. Example: *(bob,joe,sam,sue)* . In the examples below, you could replace any of the keywords with a list like this.

The default behavior of the search is to locate an intersection (or 'AND') of every element within a query. This means that the query: *"microsoft bob interface"* is the equivalent to the boolean query: *"microsoft AND bob AND interface"* .

- (without) The `-` (minus) is the most commonly used logic symbol. It means the answer should EXCLUDE references to that item.

+ (mandatory) The `+` (plus) symbol in front of a search item means that the answer MUST INCLUDE that item. This is generally used in conjunction with the permutation operation.

@N (permute) The `@` followed by a number indicates how many intersections to locate of the terms in your query. This may be confusing at first, but it is very powerful.

Table 6.2: 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`

6.1.7 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`

6.1.8 Using the Special Pattern Matchers

These pattern matchers are used to locate hard-to-find items within text:

- Regular expression matching for complex patterns
  `http://www.thunderstone.com/texis/site/pages/regexp.html`
- Approximate pattern matching for fuzzy searches
  `http://www.thunderstone.com/texis/site/pages/xpm.html`
- Numeric pattern matching for finding quantities
  `http://www.thunderstone.com/texis/site/pages/npm.html`

If improperly used these pattern matchers can slow queries. Therefore they require other keyword(s) in the query and are disabled entirely under Page proximity. For more details see the Vortex manual on Query Protection (`http://www.thunderstone.com/site/vortexman/link_qprot.html`).

Table 6.3: Pattern Matcher Examples

Query	Matcher	Finds
`ronald %regan`	Approx	Ronald Raygun, Ronald Re-an, Ronald 8eagan
`%75MYPARTNO9045d/6a`	Approx	Anything within 75% of looking like MYPARTNO9045d/6a
`/19[789][0-9]`	RegEXpr	1970-1999
`/[1-9]{3}\-=[0-9]{4}`	RegEXpr	Phone numbers: 555-1212, 820-2200
`#87`	Numeric	four score and seven, 87
`#>0<1`	Numeric	Fractions like 9/16, 55%, 0.123, 15 nanoseconds

Table 6.4: 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

6.1.9 Invoking Thesaurus Expansion

The Search Appliance 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.

6.2 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 match: Only exact matches will be allowed. (the default)

Plurals & 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.

Custom: Uses the `Custom Suffix List`, `Custom Suffix Default Removal`, and `Custom Suffix Min Length` settings to create your own custom behavior.

We call this morpheme processing, and it is generally smarter than a traditional "stemming" algorithm. It doesn't just rip the end off a word, it actually checks to see if it could be a valid form of the search term. More information is available at
`http://www.thunderstone.com/site/texisman/link_ling.html`.

Notes: Thesaurus terms are also treated in the same manner. Words smaller than 4-5 characters will not be morpheme processed.

6.3 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.

6.4 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)*

6.4.1 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.

6.4.2 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.

6.4.3 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.

Appendix A

Third-Party Software

The Search Appliance 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.

A.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. 227.

A.2 Aspell

The GNU Project's `aspell` package is executed by (but not linked or compiled into) the Search Appliance 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. 244.

A.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. 227.

A.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. 227.

A.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. 244.

A.6 libxml2

`Libxml2` may be used by Thunderstone's products to parse XML documents. It is available at `http://xmlsoft.org/`, and is Copyright (C) 1998-2003 Daniel Veillard. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of `libxml2` and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

A.7 Libxslt

`Libxslt` may be used by Thunderstone's products to apply XSL transforms to XML documents. It is available at `http://xmlsoft.org/XSLT/` and is Copyright (C) 2001-2002 Daniel Veillard. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of `libxslt` and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

A.8 Libexslt

`Libexslt` may be used by Thunderstone's products when applying XSL transforms to XML documents. It is Copyright (C) 2001-2002 Thomas Broyer, Charlie Bozeman and Daniel Veillard. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of `libexslt` and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions

of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of the authors shall not be used in advertising or otherwise to promote the sale, use or other deal- ings in this Software without prior written authorization from them.

A.9 JDBC drivers

Walking external databases with the DBWalker module may use one or more of the following drivers.

A.9.1 Oracle JDBC driver

Walking external Oracle databases with the DBWalker module may use the Oracle driver, subject to the license below:

```
 ORACLE TECHNOLOGY NETWORK DEVELOPMENT AND DISTRIBUTION LICENSE AGREEMENT

"We," "us," and "our" refers to Oracle USA, Inc., for and on behalf of
itself and its subsidiaries and affiliates under common control.
"You" and "your" refers to the individual or entity that wishes to use
the programs from Oracle.  "Programs" refers to the software product
you wish to download and use and program documentation.  "License"
refers to your right to use the programs under the terms of this
agreement.  This agreement is governed by the substantive and
procedural laws of California.  You and Oracle agree to submit to the
exclusive jurisdiction of, and venue in, the courts of San Francisco,
San Mateo, or Santa Clara counties in California in any dispute
arising out of or relating to this agreement.

We are willing to license the programs to you only upon the condition
that you accept all of the terms contained in this agreement.  Read
the terms carefully and select the "Accept" button at the bottom of
the page to confirm your acceptance.  If you are not willing to be
bound by these terms, select the "Do Not Accept" button and the
registration process will not continue.

License Rights

We grant you a nonexclusive, nontransferable limited license to use
the programs for purposes of developing your applications.  You may
also distribute the programs with your applications to your customers.
If you want to use the programs for any purpose other than as
```

expressly permitted under this agreement you must contact us, or an Oracle reseller, to obtain the appropriate license. We may audit your use of the programs. Program documentation is either shipped with the programs, or documentation may accessed online at: http://otn.oracle.com/docs

Ownership and Restrictions

We retain all ownership and intellectual property rights in the programs. You may make a sufficient number of copies of the programs for the licensed use and one copy of the programs for backup purposes.

You may not:
- use the programs for any purpose other than as provided above;
- distribute the programs unless accompanied with your applications;
- charge your end users for use of the programs;
- remove or modify any program markings or any notice of our proprietary rights;
- use the programs to provide third party training on the content and/or functionality of the programs, except for training your licensed users;
- assign this agreement or give the programs, program access or an interest in the programs to any individual or entity except as provided under this agreement;
- cause or permit reverse engineering (unless required by law for interoperability), disassembly or decompilation of the programs;
- disclose results of any program benchmark tests without our prior consent; or,
- use any Oracle name, trademark or logo.

Program Distribution

We grant you a nonexclusive, nontransferable right to copy and distribute the programs to your end users provided that you do not charge your end users for use of the programs and provided your end users may only use the programs to run your applications for their business operations. Prior to distributing the programs you shall require your end users to execute an agreement binding them to terms consistent with those contained in this section and the sections of this agreement entitled "License Rights," "Ownership and Restrictions," "Export," "Disclaimer of Warranties and Exclusive Remedies," "No Technical Support," "End of Agreement," "Relationship Between the Parties," and "Open Source." You must also include a provision stating that your end users shall have no right to distribute the programs, and a provision specifying us as a third party beneficiary of the agreement. You are responsible for obtaining these agreements with your end users.

You agree to: (a) defend and indemnify us against all claims and damages caused by your distribution of the programs in breach of this agreements and/or failure to include the required contractual provisions in your end user agreement as stated above; (b) keep

executed end user agreements and records of end user information including name, address, date of distribution and identity of programs distributed; (c) allow us to inspect your end user agreements and records upon request; and, (d) enforce the terms of your end user agreements so as to effect a timely cure of any end user breach, and to notify us of any breach of the terms.

Export

You agree that U.S. export control laws and other applicable export and import laws govern your use of the programs, including technical data; additional information can be found on Oracle’s Global Trade Compliance web site located at:
http://www.oracle.com/products/export/index.html?content.html
You agree that neither the programs nor any direct product thereof will be exported, directly, or indirectly, in violation of these laws, or will be used for any purpose prohibited by these laws including, without limitation, nuclear, chemical, or biological weapons proliferation.

Disclaimer of Warranty and Exclusive Remedies

THE PROGRAMS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND. WE FURTHER DISCLAIM ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT.

IN NO EVENT SHALL WE BE LIABLE FOR ANY INDIRECT, INCIDENTAL, SPECIAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, OR DAMAGES FOR LOSS OF PROFITS, REVENUE, DATA OR DATA USE, INCURRED BY YOU OR ANY THIRD PARTY, WHETHER IN AN ACTION IN CONTRACT OR TORT, EVEN IF WE HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. OUR ENTIRE LIABILITY FOR DAMAGES HEREUNDER SHALL IN NO EVENT EXCEED ONE THOUSAND DOLLARS (U.S. \$1,000).

No Technical Support

Our technical support organization will not provide technical support, phone support, or updates to you for the programs licensed under this agreement.

Restricted Rights

If you distribute a license to the United States government, the programs, including documentation, shall be considered commercial computer software and you will place a legend, in addition to applicable copyright notices, on the documentation, and on the media label, substantially similar to the following:

NOTICE OF RESTRICTED RIGHTS

"Programs delivered subject to the DOD FAR Supplement are ’commercial computer software’ and use, duplication, and disclosure of the

programs, including documentation, shall be subject to the licensing restrictions set forth in the applicable Oracle license agreement. Otherwise, programs delivered subject to the Federal Acquisition Regulations are 'restricted computer software' and use, duplication, and disclosure of the programs, including documentation, shall be subject to the restrictions in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987). Oracle USA, Inc., 500 Oracle Parkway, Redwood City, CA 94065."

End of Agreement

You may terminate this agreement by destroying all copies of the programs. We have the right to terminate your right to use the programs if you fail to comply with any of the terms of this agreement, in which case you shall destroy all copies of the programs.

Relationship Between the Parties

The relationship between you and us is that of licensee/licensor. Neither party will represent that it has any authority to assume or create any obligation, express or implied, on behalf of the other party, nor to represent the other party as agent, employee, franchisee, or in any other capacity. Nothing in this agreement shall be construed to limit either party's right to independently develop or distribute software that is functionally similar to the other party's products, so long as proprietary information of the other party is not included in such software.

Open Source

"Open Source" software - software available without charge for use, modification and distribution - is often licensed under terms that require the user to make the user's modifications to the Open Source software or any software that the user 'combines' with the Open Source software freely available in source code form. If you use Open Source software in conjunction with the programs, you must ensure that your use does not: (i) create, or purport to create, obligations of us with respect to the Oracle programs; or (ii) grant, or purport to grant, to any third party any rights to or immunities under our intellectual property or proprietary rights in the Oracle programs. For example, you may not develop a software program using an Oracle program and an Open Source program where such use results in a program file(s) that contains code from both the Oracle program and the Open Source program (including without limitation libraries) if the Open Source program is licensed under a license that requires any "modifications" be made freely available. You also may not combine the Oracle program with programs licensed under the GNU General Public License ("GPL") in any manner that could cause, or could be interpreted or asserted to cause, the Oracle program or any modifications thereto to become subject to the terms of the GPL.

Entire Agreement

```
You agree that this agreement is the complete agreement for the
programs and licenses, and this agreement supersedes all prior or
contemporaneous agreements or representations.  If any term of this
agreement is found to be invalid or unenforceable, the remaining
provisions will remain effective.

Last updated: 03/09/05
```

A.9.2 JTDS JDBC driver

Walking external databases with the DBWalker module may use the JTDS driver, for SQL Server(R) and Sybase(R) databases. This driver is governed by the GNU Lesser GPL; see p. 234.

A.9.3 PostgreSQL JDBC driver

Walking external databases with the DBWalker module may use the PostgreSQL driver. The license is reproduced below:

```
Copyright (c) 1997-2005, PostgreSQL Global Development Group
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. Neither the name of the PostgreSQL Global Development Group nor the
   names of its contributors may be used to endorse or promote
   products derived from this software without specific prior written
   permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS 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 COPYRIGHT
OWNER 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.
```

A.9.4 MySQL JDBC driver

Walking external databases with the DBWalker module may use the MySQL driver, governed by the GNU GPL; see p. 227.

A.10 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. 227.

A.11 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 (`eay@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.

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.

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).  This product includes software written by Tim
Hudson (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.]

A.12 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
```

A.13 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.
```

A.14 zlib

The Search Appliance utilizes the `zlib` compression library, with the following license:

```
Copyright (C) 1995-2003 Jean-loup Gailly and Mark Adler.

This software is provided 'as-is', without any express or implied
warranty.  In no event will the authors be held liable for any
damages arising from the use of 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. The origin of this software must not be misrepresented; you must not
   claim that you wrote the original software. If you use this software
   in a product, an acknowledgment in the product documentation would be
   appreciated but is not required.
2. Altered source versions must be plainly marked as such, and must not
   be misrepresented as being the original software.
3. This notice may not be removed or altered from any source
   distribution.

Jean-loup Gailly        Mark Adler
jloup@gzip.org          madler@alumni.caltech.edu

The zlib library data format is described by RFCs (Request for Comments)
1950 to 1952 in the files http://www.ietf.org/rfc/rfc1950.txt (zlib
format), rfc1951.txt (deflate format) and rfc1952.txt (gzip format).
```

A.15 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. 253. Note that the `libtxjs.*` library also contains other (Thunderstone) code.

A.16 PDF/anytotx plugin

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

A.17 thttpd - throttling HTTP server

The Search Appliance's `vhttpd` web server is derived in part from `thttpd`, Copyright ©1995 by Jef Poskanzer ¡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.
```

A.18 RedHat Linux

Some versions of The Search Appliance use the RedHat Linux operating system, version 7.3, which is licensed under the GNU Public License, p. 227. See also `http://www.redhat.com/licenses/thirdparty/eula.html` for more information.

A.19 CentOS Linux

Some versions of The Search Appliance use the CentOS Linux operating system, version 5, which is licensed under the GNU Public License, p. 227. See also `http://www.redhat.com/licenses/thirdparty/eula.html` for more information.

A.20 Webmin

The Search Appliance uses the Webmin web-based system administration system for maintaining and configuring the operating system. Copyright ©Jamie Cameron All rights reserved. Complete source is available at: `http://www.webmin.com/`. The license is reproduced below:

```
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. Neither the name of the developer nor the names of contributors
   may be used to endorse or promote products derived from this software
   without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE DEVELOPER ‘‘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 DEVELOPER 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.
```

A.21 Java

The Search Appliance uses the Java 2 run-time environment developed by Sun Microsystems, Inc. to index third-party databases using JDBC drivers. This product includes code licensed from RSA Security, Inc. Some portions licensed from IBM are available at `http://oss.software.ibm.com/icu4j/` as well. The license agreement is reproduced below:

```
   Sun Microsystems, Inc.  Binary Code License Agreement
```

READ THE TERMS OF THIS AGREEMENT AND ANY PROVIDED SUPPLEMENTAL LICENSE TERMS (COLLECTIVELY "AGREEMENT") CAREFULLY BEFORE OPENING THE SOFTWARE MEDIA PACKAGE. BY OPENING THE SOFTWARE MEDIA PACKAGE, YOU AGREE TO THE TERMS OF THIS AGREEMENT. IF YOU ARE ACCESSING THE SOFTWARE ELECTRONICALLY, INDICATE YOUR ACCEPTANCE OF THESE TERMS BY SELECTING THE "ACCEPT" BUTTON AT THE END OF THIS AGREEMENT. IF YOU DO NOT AGREE TO ALL THESE TERMS, PROMPTLY RETURN THE UNUSED SOFTWARE TO YOUR PLACE OF PURCHASE FOR A REFUND OR, IF THE SOFTWARE IS ACCESSED ELECTRONICALLY, SELECT THE "DECLINE" BUTTON AT THE END OF THIS AGREEMENT.

1. LICENSE TO USE. Sun grants you a non-exclusive and non-transferable license for the internal use only of the accompanying software and documentation and any error corrections provided by Sun (collectively "Software"), by the number of users and the class of computer hardware for which the corresponding fee has been paid.

2. RESTRICTIONS. Software is confidential and copyrighted. Title to Software and all associated intellectual property rights is retained by Sun and/or its licensors. Except as specifically authorized in any Supplemental License Terms, you may not make copies of Software, other than a single copy of Software for archival purposes. Unless enforcement is prohibited by applicable law, you may not modify, decompile, or reverse engineer Software. Licensee acknowledges that Licensed Software is not designed or intended for use in the design, construction, operation or maintenance of any nuclear facility. Sun Microsystems, Inc. disclaims any express or implied warranty of fitness for such uses. No right, title or interest in or to any trademark, service mark, logo or trade name of Sun or its licensors is granted under this Agreement.

3. LIMITED WARRANTY. Sun warrants to you that for a period of ninety (90) days from the date of purchase, as evidenced by a copy of the receipt, the media on which Software is furnished (if any) will be free of defects in materials and workmanship under normal use. Except for the foregoing, Software is provided "AS IS". Your exclusive remedy and Sun’s entire liability under this limited warranty will be at Sun’s option to replace Software media or refund the fee paid for Software.

4. DISCLAIMER OF WARRANTY. UNLESS SPECIFIED IN THIS AGREEMENT, ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT THESE DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

5. LIMITATION OF LIABILITY. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL SUN OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR SPECIAL, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF OR RELATED TO THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In no event will Sun's liability to you, whether in contract, tort (including negligence), or otherwise, exceed the amount paid by you for Software under this Agreement. The foregoing limitations will apply even if the above stated warranty fails of its essential purpose.

6. Termination. This Agreement is effective until terminated. You may terminate this Agreement at any time by destroying all copies of Software. This Agreement will terminate immediately without notice from Sun if you fail to comply with any provision of this Agreement. Upon Termination, you must destroy all copies of Software.

7. Export Regulations. All Software and technical data delivered under this Agreement are subject to US export control laws and may be subject to export or import regulations in other countries. You agree to comply strictly with all such laws and regulations and acknowledge that you have the responsibility to obtain such licenses to export, re-export, or import as may be required after delivery to you.

8. U.S. Government Restricted Rights. If Software is being acquired by or on behalf of the U.S. Government or by a U.S. Government prime contractor or subcontractor (at any tier), then the Government's rights in Software and accompanying documentation will be only as set forth in this Agreement; this is in accordance with 48 CFR 227.7201 through 227.7202-4 (for Department of Defense (DOD) acquisitions) and with 48 CFR 2.101 and 12.212 (for non-DOD acquisitions).

9. Governing Law. Any action related to this Agreement will be governed by California law and controlling U.S. federal law. No choice of law rules of any jurisdiction will apply.

10. Severability. If any provision of this Agreement is held to be unenforceable, this Agreement will remain in effect with the provision omitted, unless omission would frustrate the intent of the parties, in which case this Agreement will immediately terminate.

11. Integration. This Agreement is the entire agreement

between you and Sun relating to its subject matter. It supersedes all prior or contemporaneous oral or written communications, proposals, representations and warranties and prevails over any conflicting or additional terms of any quote, order, acknowledgment, or other communication between the parties relating to its subject matter during the term of this Agreement. No modification of this Agreement will be binding, unless in writing and signed by an authorized representative of each party.

JAVATM 2 RUNTIME ENVIRONMENT (J2RE), STANDARD EDITION, VERSION 1.4.1_X SUPPLEMENTAL LICENSE TERMS

These supplemental license terms ("Supplemental Terms") add to or modify the terms of the Binary Code License Agreement (collectively, the "Agreement"). Capitalized terms not defined in these Supplemental Terms shall have the same meanings ascribed to them in the Agreement. These Supplemental Terms shall supersede any inconsistent or conflicting terms in the Agreement, or in any license contained within the Software.

1. Software Internal Use and Development License Grant. Subject to the terms and conditions of this Agreement, including, but not limited to Section 4 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non- transferable, limited license to reproduce internally and use internally the binary form of the Software complete and unmodified for the sole purpose of designing, developing and testing your Java applets and applications intended to run on the Java platform ("Programs").

2. License to Distribute Software. Subject to the terms and conditions of this Agreement, including, but not limited to Section 4 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to reproduce and distribute the Software, provided that (i) you distribute the Software complete and unmodified (unless otherwise specified in the applicable README file) and only bundled as part of, and for the sole purpose of running, your Programs, (ii) the Programs add significant and primary functionality to the Software, (iii) you do not distribute additional software intended to replace any component(s) of the Software (unless otherwise specified in the applicable README file), (iv) you do not remove or alter any proprietary legends or notices contained in the Software, (v) you only distribute the Software subject to a license agreement that protects Sun’s interests consistent with the terms contained in this Agreement, and (vi) you agree to defend and indemnify Sun and its licensors from and against

any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of any and all Programs and/or Software. (vi) include the following statement as part of product documentation (whether hard copy or electronic), as a part of a copyright page or proprietary rights notice page, in an "About" box or in any other form reasonably designed to make the statement visible to users of the Software: "This product includes code licensed from RSA Security, Inc.", and (vii) include the statement, "Some portions licensed from IBM are available at http://oss.software.ibm.com/icu4j/ ".

3. License to Distribute Redistributables. Subject to the terms and conditions of this Agreement, including but not limited to Section 4 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to reproduce and distribute those files specifically identified as redistributable in the Software "README" file ("Redistributables") provided that: (i) you distribute the Redistributables complete and unmodified (unless otherwise specified in the applicable README file), and only bundled as part of Programs, (ii) you do not distribute additional software intended to supersede any component(s) of the Redistributables (unless otherwise specified in the applicable README file), (iii) you do not remove or alter any proprietary legends or notices contained in or on the Redistributables, (iv) you only distribute the Redistributables pursuant to a license agreement that protects Sun's interests consistent with the terms contained in the Agreement, (v) you agree to defend and indemnify Sun and its licensors from and against any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of any and all Programs and/or Software, (vi) include the following statement as part of product documentation (whether hard copy or electronic), as a part of a copyright page or proprietary rights notice page, in an "About" box or in any other form reasonably designed to make the statement visible to users of the Software: "This product includes code licensed from RSA Security, Inc.", and (vii) include the statement, "Some portions licensed from IBM are available at http://oss.software.ibm.com/icu4j/ ".

4. Java Technology Restrictions. You may not modify the Java Platform Interface ("JPI", identified as classes contained within the "java" package or any subpackages of the "java" package), by creating additional classes within

the JPI or otherwise causing the addition to or modification of the classes in the JPI. In the event that you create an additional class and associated API(s) which (i) extends the functionality of the Java platform, and (ii) is exposed to third party software developers for the purpose of developing additional software which invokes such additional API, you must promptly publish broadly an accurate specification for such API for free use by all developers. You may not create, or authorize your licensees to create, additional classes, interfaces, or subpackages that are in any way identified as "java", "javax", "sun" or similar convention as specified by Sun in any naming convention designation.

5. Notice of Automatic Software Updates from Sun. You acknowledge that the Software may automatically download, install, and execute applets, applications, software extensions, and updated versions of the Software from Sun ("Software Updates"), which may require you to accept updated terms and conditions for installation. If additional terms and conditions are not presented on installation, the Software Updates will be considered part of the Software and subject to the terms and conditions of the Agreement.

6. Notice of Automatic Downloads. You acknowledge that, by your use of the Software and/or by requesting services that require use of the Software, the Software may automatically download, install, and execute software applications from sources other than Sun ("Other Software"). Sun makes no representations of a relationship of any kind to licensors of Other Software. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL SUN OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR SPECIAL, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF OR RELATED TO THE USE OF OR INABILITY TO USE OTHER SOFTWARE, EVEN IF SUN HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Trademarks and Logos. You acknowledge and agree as between you and Sun that Sun owns the SUN, SOLARIS, JAVA, JINI, FORTE, and iPLANET trademarks and all SUN, SOLARIS, JAVA, JINI, FORTE, and iPLANET-related trademarks, service marks, logos and other brand designations ("Sun Marks"), and you agree to comply with the Sun Trademark and Logo Usage Requirements currently located at: http://www.sun.com/policies/trademarks
Any use you make of the Sun Marks inures to Sun’s benefit.

8. Source Code. Software may contain source code that is provided solely for reference purposes pursuant to the terms

```
of this  Agreement.  Source  code  may not be  redistributed
unless expressly provided for in this Agreement.

9.  Termination   for   Infringement.   Either   party   may
terminate  this  Agreement  immediately  should any Software
become, or in either  party's  opinion  be likely to become,
the subject of a claim of infringement  of any  intellectual
property right.

For inquiries please contact:  Sun Microsystems,  Inc., 4150
Network  Circle,  Santa  Clara,  California   95054,  U.S.A.
(LFI#120080/Form ID#011801)
```

A.22 OpenSSL RPM

The Search Appliance uses a Perl module that contains OpenSSL. Copyright ©1996-2002 Sampo Kellomaki `sampo@symlabs.com` All Rights Reserved. See p. 213 for more information.

A.23 RAID utilities

The Search Appliance may use RAID utilities developed by the Adaptec Corporation. These are used by Thunderstone for system maintenance in this product. Usage is governed by the license below:

```
Copyright (c) 1996-2004, Adaptec Corporation
All rights reserved.

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

- Redistributions of source code must retain the above copyright
  notice, this list of conditions and the following disclaimer.
- 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.
- Neither the name of the Adaptec Corporation nor the names of its
  contributors may be used to endorse or promote products derived from
  this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS 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 COPYRIGHT
OWNER 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.
```

A.24 LCDpoc

The Search Appliance may use LCDproc Copyright (C) 1998-2006 William W. Ferrell, Scott Scriven and many other contributors which is licensed under the GNU Public License, p. 227. This is used by Thunderstone for driving the front panel LCD on appliances so equipped.

A.25 GNU General Public License

Some third-party software packages shipped with the Search Appliance are governed by the GNU General Public License, reproduced below. See the Third-Party Software section, p. 205, 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.
```

A.26 GNU Lesser General Public License

Some third-party software packages distributed with the Search Appliance are governed by the GNU Lesser General Public License, reproduced below. See the Third-Party Software section, p. 205, 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!

A.27 GNU Library General Public License

Some third-party software packages distributed with the Search Appliance are governed by the GNU Library General Public License, reproduced below. See the Third-Party Software section, p. 205, 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!
```

A.28 Netscape Public License

Some third-party software packages distributed with the Search Appliance are governed by the Netscape Public License, reproduced below. See the Third-Party Software section, p. 205, 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 (50beneficial 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, MERCHANTABLE, FIT FOR A PARTICULAR PURPOSE OR NON-INFRINGING. 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 declatory 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.]

A.29 UnixUtils

Windows versions of The Search Appliance may include the `UnixUtils` package of Unix utilities ported to Windows, for use in debugging, analyzing and reporting problems, which includes the following

disclaimer:

```
Disclaimer for UnixUtils

THIS SOFTWARE IS PROVIDED "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 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.

Karl M. Syring
```

A.30 PuTTY

Windows versions of The Search Appliance may include the PuTTY SSH connectivity tools, for use in debugging, analyzing and reporting problems, which includes the following license:

```
PuTTY is copyright 1997-2011 Simon Tatham.

Portions copyright Robert de Bath, Joris van Rantwijk, Delian
Delchev, Andreas Schultz, Jeroen Massar, Wez Furlong, Nicolas Barry,
Justin Bradford, Ben Harris, Malcolm Smith, Ahmad Khalifa, Markus
Kuhn, Colin Watson, and CORE SDI S.A.

Permission is hereby granted, free of charge, to any person
obtaining a copy of this software and associated documentation files
(the "Software"), to deal in the Software without restriction,
including without limitation the rights to use, copy, modify, merge,
publish, distribute, sublicense, and/or sell copies of the Software,
and to permit persons to whom the Software is furnished to do so,
subject to the following conditions:

The above copyright notice and this permission notice shall be
included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND,
EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF
MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND
NONINFRINGEMENT.  IN NO EVENT SHALL THE COPYRIGHT HOLDERS BE LIABLE
FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF
CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION
WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
```