Contents

Figures, Tables, and Listings ........................................................................................................ 13

Preface ......................................................................................................................................... 17
Using This Book ............................................................................................................................ 18
   Version ....................................................................................................................................... 18
   Organization of This Book ......................................................................................................... 18
   Stylistic Conventions ................................................................................................................. 19
Related Documentation ................................................................................................................. 21
Release Notes and Document Updates .......................................................................................... 21
Verity Technical Support ................................................................................................................. 22

PART I MANAGING ULTRASEEK

1 Getting Started .............................................................................................................................. 25
   Ultraseek Features ....................................................................................................................... 25
      Key Features ........................................................................................................................... 26
      Add-in Modules ....................................................................................................................... 26
   Starting Ultraseek ......................................................................................................................... 27
   Accessing the Administrative Interface ....................................................................................... 28
   Accessing Additional Utilities ...................................................................................................... 31
      Obtaining System Information ............................................................................................... 31
      Tuning the System File Cache ............................................................................................... 31
      Verifying the Server Installation ............................................................................................ 32
   Getting Help ............................................................................................................................... 32
      Accessing Online Help ............................................................................................................ 32
      Accessing Application Help .................................................................................................... 32
      More Services ......................................................................................................................... 33
2 Creating Collections ................................................................. 35

About Indexing Documents .......................................................... 36
  HTTP Spidering ........................................................................... 36
  MIME Types .............................................................................. 37
  Filters ...................................................................................... 37
  File Scanning ........................................................................... 38
  File Types ............................................................................... 38

About Creating Collections ......................................................... 39

Spidering the Network ............................................................... 40
  Creating a Spider Collection .................................................. 40
  Adding Blog Entries to Spider Collections ............................. 43

Scanning the File System .......................................................... 44

Following Usenet Groups ......................................................... 46

Indexing Microsoft Exchange Public Folders ......................... 47

Indexing Databases ................................................................. 48
  Supported Databases ........................................................... 49
  Configuring Windows NT/2000/2003 Drivers ....................... 49
  UNIX Drivers ......................................................................... 50
  Creating Database Collections ............................................. 50

Creating Mirrored Collections ................................................ 56

Merging Collections ............................................................... 58

Creating Collections Using the XPA API ................................. 59

Defining Filters ........................................................................ 59
  Specifying Patterns ............................................................. 59
    Disallowing Directories ..................................................... 61
    Allowing Directories ......................................................... 61
    Sequence of Allows and Disallows .................................... 62
    Using Wildcard and Regular Expression Patterns .............. 63
  Allowing or Disallowing Sites .............................................. 64
    Otherwise Allowed Sites .................................................. 64
    Always Disallowed Sites ................................................... 65
  Setting Filter Options for Spider Collections ....................... 65
  Setting Allowable Languages .............................................. 67

Importing Collection Configurations ....................................... 68
  Importing from Ultraseek ...................................................... 68
3 Managing Collections ................................................................. 73

Basic Collection Management .................................................. 74
Editing Collections ...................................................................... 75
Deleting Collections ................................................................... 75
Clearing Collections and Reindexing Documents ......................... 76
Suspending and Resuming Indexing ........................................... 76
Revisiting Sites and Reindexing Documents ................................. 76
Viewing Sites ............................................................................ 77
Viewing Log Files ....................................................................... 78
Merging Indexes ......................................................................... 79
Managing URLs .......................................................................... 80
Getting URL Status .................................................................... 80
Adding URLs ............................................................................. 82
Deleting URLs ........................................................................... 82
Filtering Duplicate URLs ............................................................ 83
Duplicate Determination ............................................................. 83
Duplicate URL Preference .......................................................... 83
Duplicate URL Filter Specification ........................................... 84
Improving Document Relevance Scores ..................................... 84
Using Page Types ....................................................................... 85
Creating Page Types ................................................................... 86
Implementing Page Types .......................................................... 88
Testing Page Types .................................................................... 88
Adjusting Quality Factors ........................................................... 89
Setting Network Parameters ...................................................... 92
User Agent Specification .......................................................... 93
Authentication ........................................................................... 93
Form-Based Authentication ....................................................... 94
Additional HTTP Header Entries ............................................... 96
Spider Throttle ........................................................................... 97
Proxy Server Specification ....................................................... 98
Direct File Access ...................................................................... 98
Automatic URL Adder (Cooperative Spidering) ......................... 99
Tuning Collections ....................................................................................................................... 99
  Setting Tuning Parameters ........................................................................................................... 100
Scheduling Activities .................................................................................................................. 106
Setting Activity Curfews ............................................................................................................ 107
Setting Indexer Weights .............................................................................................................. 108
Specifying HTML META Tag Names ........................................................................................ 109
Getting External Metadata from a Database ............................................................................ 111
  Configuring External Metadata from a Database ................................................................. 111
  Editing the Configuration ........................................................................................................ 112
  Deleting the Configuration ...................................................................................................... 112
Replacing Document Titles Automatically .............................................................................. 113
Setting Word Spam Detection Thresholds .............................................................................. 114
Customizing Title Records ........................................................................................................ 114
Checking the Internal and Display Names ............................................................................. 115

4 Customizing Ultraseek Search Forms .............................................................................. 117
  Customizing Styles ................................................................................................................. 118
  Editing Style Sheets ................................................................................................................ 119
Setting Server Parameters ......................................................................................................... 124
  Setting Style Parameters ....................................................................................................... 124
  Setting Query Parameters ..................................................................................................... 126

5 Generating Ultraseek Reports ....................................................................................... 133
  Generating Reports ................................................................................................................ 134
  Generating Top Queries Reports .......................................................................................... 136
  Generating No-Results Reports ......................................................................................... 137
  Generating No-Clickthrough Reports ............................................................................... 138
  Generating Usage Summaries ............................................................................................... 139
  Generating Most Requested Documents Reports ............................................................... 140
  Generating Trend Reports ................................................................................................... 141
  Generating Reports Per Topic .............................................................................................. 142
  Monitoring Recent Queries ................................................................................................. 143
6 Monitoring Activity ................................................................. 145
  Collection Status ................................................................. 146
  Performance Information ....................................................... 148
  HTTP Server Threads ............................................................ 149

7 Server Administration .......................................................... 151
  Managing the Server ............................................................. 152
    Viewing Ultraseek Status .................................................. 152
    Restarting the Server ....................................................... 152
  Shutting Down and Restarting the Server .............................. 153
    Shutting Down the Server ................................................. 153
    Restarting the Server ..................................................... 153
  Setting Server Parameters .................................................. 154
    Setting Main Server Parameters ..................................... 154
    Setting Advanced Server Parameters .............................. 158
  Managing Administrative Users and Client Access .................. 166
    Creating Administrative User Accounts ........................... 166
    Deleting Administrative User Accounts ......................... 166
    Changing Passwords ...................................................... 167
    Additional Password Security ......................................... 167
      Password Length ......................................................... 168
      Password Longevity .................................................... 168
      Password Uniqueness .................................................. 169
      Failed Administrator Logins ........................................ 169
  Managing Client Access ....................................................... 170
    Allowing HTTPS Administrative Access ............................ 172
  Managing Document Types ................................................ 172
    Document Type Specification ........................................... 173
      Supported MIME Types .............................................. 174
      Supported Highlight Types .......................................... 176
    Document Type Parsing .................................................. 177
    Creating XML Mappings ................................................. 177
      XML Generic Mappings ............................................... 178
      Document-Specific Mappings ....................................... 179
      Specifying Attributes ................................................ 180
8 Secure Spidering ............................................................................. 183
  Authentication .......................................................................................................................... 184
  Managing Digital Certificates .................................................................................................... 187
    Using keyfm ............................................................................................................................ 188
      Options (Functions) .............................................................................................................. 188
      Options (Modifiers) ............................................................................................................ 189
    Adding a Trusted Root to the mykeys.kdb File ............................................................. 190
    Adding a Client Certificate ............................................................................................... 192
    Changing the Password ................................................................................................. 193
    Removing a Key or Certificate ...................................................................................... 194
  Managing Keys ..................................................................................................................... 195
    Creating Public or Private Key Pairs and Installing Certificates ........................................ 195
    Displaying Distinguished Names for Public or Private Key Pairs ...................................... 198
    Using PKCS#12 to Export Public or Private Key Pairs and Certificates ........................ 199
    Using PKCS#12 to Import Public or Private Key Pairs and Certificates .......................... 200
    Deleting Public or Private Key Pairs and Certificates ..................................................... 200
  Managing Trusted Roots ........................................................................................................ 201
    Displaying Trusted Roots ................................................................................................. 201
    Deleting Trusted Roots ................................................................................................. 201
    Installing Trusted Roots .............................................................................................. 202
    Using the ca Tool for Quick Testing ............................................................................... 202
    Installing an UnknownCAHook ......................................................................................... 203

9 Using Content Assistants .............................................................................. 205
  Understanding Content Assistants .......................................................................................... 206
    Evaluating Documents ......................................................................................................... 206
    Example Uses ....................................................................................................................... 207
  About the Content Assistant Interface .................................................................................... 207
  Configuring Your Service ......................................................................................................... 209
  Creating a Topic Tree .............................................................................................................. 211
    Sample Topic Tree ............................................................................................................... 212
    Sample XML Definition ...................................................................................................... 213
Contents

Cross-References.................................................................................................................. 214
Pushing the Topic Tree ........................................................................................................ 214
   CAS Identification Block ................................................................................................ 214
   HTTP PUT Details ......................................................................................................... 215
Document Evaluation ......................................................................................................... 216
   URLDesc Evaluation .................................................................................................... 216
   HdrDesc Evaluation .................................................................................................... 217
   FullDesc Evaluation .................................................................................................... 217
   IndexAsstDesc ............................................................................................................ 219
   IndexAsstDesc Contents ............................................................................................ 220
Modifying Document Contents .......................................................................................... 221
Example GetFullDesc Message ....................................................................................... 224
DTD for Topics.xml .......................................................................................................... 225

10 Search Syntax .................................................................................................................. 227
Search Strategies ................................................................................................................ 228
   Spell Checking .............................................................................................................. 228
   Phrases and Terms ....................................................................................................... 229
   Commas and Separating Phrases ................................................................................. 230
   Capitalization ............................................................................................................... 230
Plus and Minus Operators ............................................................................................... 231
   Plus .............................................................................................................................. 231
   Minus ............................................................................................................................ 231
Vertical Bar and “Search these results” ........................................................................... 232
   The Difference Between Vertical Bar and Plus ......................................................... 232
Double-Vertical bars ......................................................................................................... 233
Wildcard Search ................................................................................................................. 234
   Supported Operators ................................................................................................... 234
   Wildcard Position ....................................................................................................... 235
   Rare Term Exclusion ................................................................................................... 235
   Using Modifiers in Wildcard Searches ...................................................................... 235
Field Searches .................................................................................................................... 236
   Searching by Document Title ..................................................................................... 237
   The above example yields the most narrow pool of search results. ......................... 237
Contents

Searching by URL ................................................................. 237
Searching by Description or Keywords ........................................ 238
Searching by Document Type .................................................. 238
Searching by Language ......................................................... 239
Searching by Link ................................................................. 240
Searching by Image Link ........................................................ 240
Searching by Metadata ........................................................... 240
Searching by Dublin Core Metadata ........................................... 241
Searching by Site ................................................................. 241
Searching by Topic ................................................................... 242
Searching by Encoding ........................................................... 242

PART II  CONTENT CLASSIFICATION ENGINE

11  Getting Started with CCE .................................................. 245
   About CCE ............................................................................. 246
      Manage Topics with Flexible Automation Features .......... 246
      Maintain Control With Flexible Automation Features ....... 246
   Obtaining a CCE License Key .............................................. 247
   Starting CCE ......................................................................... 247
   Sample CCE Topics Pages ................................................... 249
   Setting CCE Display Parameters ........................................ 253
      CCE Initial Page ............................................................... 254
      CCE Topic Browse Pages ................................................ 255
      CCE Search Pages ........................................................... 255
   Customizing CCE in the Search Interface ......................... 256

12  Creating and Editing Topics .............................................. 257
   Creating Topics ..................................................................... 258
   Creating Cross-References ................................................ 260
   Importing Topics ............................................................... 261
   Changing the Top-Level Name and Blurb ......................... 262
   Export Topics as Verity TAX File ....................................... 263
Contents

Editing Topics ......................................................................................................................................................... 263
  Browse a Topic ..................................................................................................................................................... 265
  Delete a Topic .................................................................................................................................................... 265
  Create a Subtopic ............................................................................................................................................... 265
  Create a Cross-reference Subtopic ................................................................................................................... 265
  Import Subtopics ............................................................................................................................................... 265
  Make a Copy of a Topic ....................................................................................................................................... 266
  Create New Super Topic ................................................................................................................................... 266

Understanding the Topic Definition File ........................................................................................................ 266

13 Creating Topic Rules ........................................................................................................................................ 267

Getting Started with Rules .................................................................................................................................... 268

Specifying Rules ............................................................................................................................................... 268
  Rule Terms ....................................................................................................................................................... 269
  Rule Syntax ....................................................................................................................................................... 269
  Field Rules ....................................................................................................................................................... 270
    Specifying Rules by URL .................................................................................................................................. 271
    Specifying Rules by Exact URL ....................................................................................................................... 271
    Specifying Rules by Site ................................................................................................................................. 271
    Specifying Rules by Collection ...................................................................................................................... 272
    Specifying Rules by Document Title ........................................................................................................... 272
    Specifying Rules by Raw Terms .................................................................................................................... 272
    Specifying Rules by Document Type ........................................................................................................... 273
    Specifying Rules by Metadata ...................................................................................................................... 274

Testing Rules ......................................................................................................................................................... 275

Sample Rules ......................................................................................................................................................... 276

14 Viewing Reports ............................................................................................................................................... 279

Getting Started with Reports .......................................................................................................................... 280

Viewing Documents per Topic, Sorted by
Hierarchy ............................................................................................................................................................... 281

Viewing Documents per Topic, Sorted by
Population ............................................................................................................................................................. 282

Viewing Queries, Sorted by Frequency ............................................................................................................... 283

Viewing Topics, Sorted by Browse Frequency .................................................................................................. 284
Contents

Viewing Query Terms, Sorted by Frequency ................................................................. 285

15 Mirroring .................................................................................................................. 287

Enabling and Disabling Topic Mirroring ................................................................. 288
  Enabling Topic Mirroring ...................................................................................... 288
  Disabling Topic Mirroring .................................................................................... 288
Multiple Host Issues ................................................................................................. 289
  Mirroring with Multiple Spidering Hosts ............................................................. 289
  Mirroring on a Multiple-Host System ................................................................. 289

Glossary ...................................................................................................................... 291

Index ......................................................................................................................... 297
Figures, Tables, and Listings

Figure 1-1  Administrative Interface ................................................................. 29

Table 2-1  Ultraseek Collection Types................................................................. 39
Figure 2-1  Creating New Collections ................................................................. 41
Figure 2-2  Root Specification Page for Spidering Networks ............................... 42
Figure 2-3  Updating collections with blog entries ............................................. 43
Figure 2-4  Root URL Specification Page ............................................................ 45
Figure 2-5  Root Specification Page for Usenet Groups ...................................... 46
Figure 2-6  Root Specification Page for Microsoft Exchange Public Folders ....... 47
Figure 2-7  Root Specification for Databases ...................................................... 51
Figure 2-8  New Database Table Page ................................................................. 53
Figure 2-9  Edit Table Keys Page ...................................................................... 54
Figure 2-10 Index Columns Page ....................................................................... 55
Figure 2-11 Table Column Usage Page ............................................................... 55
Figure 2-12 Merging Collections Status Page .................................................... 58
Figure 2-13 Collections | Filters Page ............................................................... 60
Figure 2-14 Incorrect Entry for Disallowing Directories ...................................... 62
Figure 2-15 Correct Entry for Disallowing Directories ........................................ 62
Figure 2-16 URL Patterns ............................................................................... 63
Figure 2-17 Allowing and Disallowing Sites ...................................................... 64
Figure 2-18 Filter Options for Spider Collections .............................................. 65
Table 2-2  Spider Filter Descriptions ................................................................ 65

Figure 3-1  Collection Status Page .................................................................... 74
Figure 3-2  URL Status Report ......................................................................... 81
Figure 3-3  Page Expert ................................................................................... 85
Figure 3-4  New Page Expert Type Page ........................................................... 87
| Figure 3-5 | Quality Page | 89 |
| Figure 3-6 | Spider Collection Network Page (Top Half) | 92 |
| Figure 3-7 | Form-Based Authentication Specification | 94 |
| Figure 3-8 | Collection Network Page (Bottom Half) | 96 |
| Figure 3-9 | Tuning Parameters | 101 |
| Figure 3-10 | HTML Meta Tag Names | 110 |
| Figure 3-11 | Document Title Replacement | 113 |
| Figure 4-1 | Style Editor | 119 |
| Table 4-1 | Style Sheet Properties | 120 |
| Figure 4-2 | Style Sheet Properties on Initial Search Page | 122 |
| Figure 4-3 | Style Sheet Properties on Search Results Page | 123 |
| Figure 4-4 | Interface | Style Page | 125 |
| Table 4-2 | Style Page Parameters | 125 |
| Figure 4-5 | Interface | Query Page | 127 |
| Table 4-3 | Query Parameters | 128 |
| Figure 5-1 | Reports Page | 135 |
| Figure 5-2 | Top Queries Report | 136 |
| Table 5-1 | Top Query Metrics | 136 |
| Figure 5-3 | No Result Report | 137 |
| Figure 5-4 | No-Clickthrough Report | 138 |
| Figure 5-5 | Usage Report | 139 |
| Figure 5-6 | Top Requested Documents Report | 140 |
| Figure 5-7 | Trend Report | 141 |
| Figure 5-8 | Viewing Topics Sorted by Browse Frequency | 142 |
| Figure 5-9 | Recent Queries | 143 |
| Figure 6-1 | Collection Status | 146 |
| Figure 6-2 | Performance Tab | 148 |
| Figure 6-3 | HTTP Server Threads | 149 |
| Figure 7-1 | Setting Main Server Parameters | 155 |
| Figure 7-2 | Advanced Server Parameters | 159 |
### Figures, Tables, and Listings

| Figure 7-3 | Doc Types Page | 173 |
| Table 7-1 | Supported MIME Types | 174 |
| Figure 7-4 | XML Generic Mappings | 178 |
| Figure 7-5 | XML Attribute Mapping for TEI documents | 180 |
| Figure 7-6 | Sample Quick Link | 181 |

| Figure 8-1 | Spidering Protected Content Using Ultraseek | 185 |
| Table 8-1 | Secure Spidering Terms | 185 |
| Table 8-2 | keyfm Functions | 188 |
| Table 8-3 | keyfm Modifiers | 189 |

| Figure 9-1 | Transactions between Content Assistants and Ultraseek | 208 |
| Figure 9-2 | Content Assistant Service Page | 209 |
| Figure 9-3 | Sample Topic Tree | 212 |

| Figure 11-1 | CCE Topics Page | 248 |
| Figure 11-2 | CCE Topics | 249 |
| Figure 11-3 | Three-Column View | 250 |
| Figure 11-4 | Results Grouped by Topic | 251 |
| Figure 11-5 | Results with Inline Topic Path | 252 |
| Figure 11-6 | CCE Topic Parameters | 253 |

| Figure 12-1 | Creating New Topics | 258 |
| Figure 12-2 | Assigning URLs to a Topic | 259 |
| Figure 12-3 | Creating Cross-References | 260 |
| Figure 12-4 | Changing Names and Blurbs | 262 |
| Figure 12-5 | Editing Topics | 264 |

| Figure 13-1 | Create Rules Region | 268 |
| Figure 13-2 | Test Rules Page | 275 |

| Figure 14-1 | Reports Page | 280 |
| Figure 14-2 | Viewing Documents Sorted by Hierarchy | 281 |
Figures, Tables, and Listings

Figure 14-3  Viewing Topics by Population ............................................................. 282
Figure 14-4  Viewing Queries Sorted by Frequency .............................................. 283
Figure 14-5  Viewing Topics Sorted by Browse Frequency .................................... 284
Figure 14-6  Viewing Query Terms, Sorted by Frequency ................................... 285
Preface

This guide is for Ultraseek system administrators who will maintain one or more Ultraseek servers. It is intended for readers who have installed Verity Ultraseek and assumes that you are familiar with basic Ultraseek concepts.

This preface contains the following sections:

- Using This Book
- Related Documentation
- Release Notes and Document Updates
- Verity Technical Support
Using This Book

This section briefly describes the organization of this book and the stylistic conventions it uses.

Version

The information in this book is current as of Ultraceek version 5.4. The content was last modified June 3, 2005. Corrections or updates to this information may be available through the Verity Customer Support site; see “Release Notes and Document Updates” on page 21.

Organization of This Book

This book includes the following chapters and appendixes:

- **Part I: Managing Ultraceek**
  - Chapter 1, “Getting Started,” describes how to start Ultraceek and launch the Ultraceek administrative interface, and how to access additional Ultraceek utilities and online help.
  - Chapter 2, “Creating Collections,” describes how to create new collections to populate the Ultraceek index.
  - Chapter 3, “Managing Collections,” explains how to administer the Ultraceek server.
  - Chapter 7, “Server Administration,” describes how to monitor server and site activity.
  - Chapter 4, “Customizing Ultraceek Search Forms,” describes how to customize the default Ultraceek search pages.
  - Chapter 5, “Generating Ultraceek Reports,” describes how administrators can monitor the performance of the Ultraceek server and generate reports based on this information.
  - Chapter 6, “Monitoring Activity,” describes how to view server and site search activity.
  - Chapter 8, “Secure Spidering,” explains how to use secure spidering.
Chapter 9, “Using Content Assistants,” describes how to create applications that interact with Ultraseek during indexing to modify documents, assign document categories, or filter documents with inappropriate content.

Chapter 10, “Search Syntax,” describes common and unique search syntax available for obtaining and refining results.

Part II: Content Classification Engine

Chapter 11, “Getting Started with CCE,” provides an overview of how to add content classification to an Ultraseek installation.

Chapter 12, “Creating and Editing Topics,” describes how to create and modify topic hierarchies.

Chapter 13, “Creating Topic Rules,” describes how to set up rules that automatically determine which documents are included in topics and how to categorize documents.

Chapter 14, “Viewing Reports,” describes the kinds of reports you can view using CCE.

Chapter 15, “Mirroring,” describes how to set up mirroring for CCE.

Stylistic Conventions

The following stylistic conventions are used in this book.

<table>
<thead>
<tr>
<th>Convention</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain</td>
<td>Narrative text.</td>
</tr>
<tr>
<td>Bold</td>
<td>User-interface elements in narrative text:</td>
</tr>
<tr>
<td></td>
<td>- Click Cancel to halt the operation.</td>
</tr>
<tr>
<td>Italic</td>
<td>Book titles and new terms:</td>
</tr>
<tr>
<td></td>
<td>- For more information, see the Verity K2 Getting Started Guide.</td>
</tr>
<tr>
<td></td>
<td>- An index is a Verity collection, parametric index, or knowledge tree.</td>
</tr>
</tbody>
</table>
The following command-line syntax conventions are used in this book.

**Convention** | **Usage**
---|---
Monospace | File names, paths, code, and required user input:
- The name .ext file is installed in:
  ```
  C: \ Verity \ Data \
  ```
- In the **User Interface** text box, type **user1**.

**Monospace italic** | Replaceable strings in file paths and code:
- **user username**

**Monospace bold** | Data types:
- **SrvConnect** A connection handle.

**Convention** | **Usage**
---|---
[ optional ] | Brackets describe optional syntax, as in [ -create ] to specify a non-required option.

| | Bars indicate “either | or” choices, as in
| | [ option1 ] | [ option2 ]

In this example, you must choose between **option1** and **option2**.

{ required } | Braces describe required syntax in which you have a choice and that at least one choice is required, as in
{ [ option1 ] [ option2 ] }

In this example, you must choose **option1**, **option2**, or both options.

required | Absence of braces or brackets indicates required syntax in which there is no choice; you must enter the required syntax element.

variable | Italics specify variables to be replaced by actual values, as in
-merge **filename1**

... | Ellipses indicate repetition of the same pattern, as in
-merge **filename1**, **filename2** [, **filename3** ... ]

where the ellipses specify, **filename4**, and so on.

Use of punctuation—such as single and double quotes, commas, periods—indicates actual syntax; it is not part of the syntax definition.
The Ultraseek documentation set provides valuable information. You can download documentation from the Verity Download Center:

http://downloadcenter.verity.com/dlc/documentation.do

- **Verity Ultraseek Release Notes**—Describes new features, important product information, and bug fixes in each release.

- **Verity Ultraseek Customization Guide**—Describes how to customize the search interface to add functionality to your search site and integrate Ultraseek into the look-and-feel of your site.

- **Verity Ultraseek Installation Guide**—Describes how to install and upgrade Ultraseek on Windows, Linux, and Solaris.

- **Verity Ultraseek XPA Programming Guide**—Describes how to use the Java XPA (Extensible Portal Architecture) Application Programmers Interface to extend the Ultraseek feature set.

- **Verity Web Services Guide**—Describes how to implement Verity web services in a client application to access Ultraseek servers.

**Release Notes and Document Updates**

Release notes for the product described in this book are available on the Verity Customer Support site, at

http://customers.verity.com

Updates to documentation are also available at that location. It is recommended that you periodically check the Customer Support site for the existence of updates to this and other Verity product documents.

Access to the contents of the Customer Support site requires a user name and password. To obtain a user name and password, follow the signup instructions on the Customer Support site home page. You will need to supply your Verity entity ID and Verity license key.
Verity Technical Support

Verity Technical Support exists to provide you with prompt and accurate resolutions to difficulties relating to using Verity software products. You can contact Technical Support using any of the following methods:

Telephone: (403) 294-1107
Fax: (403) 750-4100
Email: tech-support@verity.com
Web: www.verity.com
PART I

Managing Ultraseek

- Chapter 1: Getting Started
- Chapter 2: Creating Collections
- Chapter 3: Managing Collections
- Chapter 4: Customizing Ultraseek Search Forms
- Chapter 5: Generating Ultraseek Reports
- Chapter 6: Monitoring Activity
- Chapter 6: Server Administration
- Chapter 8: Secure Spidering
- Chapter 10: Search Syntax
Getting Started

Verity Ultraseek provides intranet and public Web sites with a powerful new approach to information retrieval. Ultraseek delivers excellent accuracy, completeness, and timeliness, combined with remarkable speed and search capabilities.

This chapter contains the following sections:

- Ultraseek Features
- Starting Ultraseek
- Accessing the Administrative Interface
- Accessing Additional Utilities
- Getting Help

Ultraseek Features

Verity Ultraseek is easy to install and administer, enabling you to implement it on your intranet or web site with minimal effort. You can easily customize Ultraseek to meet your specific needs to provide a quick, easy, and robust search experience for users.

The ease of use of Ultraseek allows users to perform accurate searches easily and quickly. The natural language interface, the speed and accuracy of results, and the default interface make Ultraseek an excellent search engine for users of any ability. Administrators can take advantage of a fully customizable, fully tunable spider in Ultraseek’s easy-to-use administrative interface.
Verity Ultraseek helps you make your intranet more accessible and easier to navigate, which improves employee productivity. On your public Web site, Ultraseek gives your visitors an easy and fast way to find what they are looking for, ensuring repeat visits.

### Key Features

This section provides you with a list of the key features of Ultraseek. For a more extensive list of features, see the complete feature list on the Verity web site:


<table>
<thead>
<tr>
<th>Best User Experience</th>
<th>Simple Administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast searches and relevant results</td>
<td>Easy install and startup</td>
</tr>
<tr>
<td>Easy to use natural language queries</td>
<td>Set it, and forget it</td>
</tr>
<tr>
<td>Full set of search operators</td>
<td>Remote management via browser</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Real-Time Index</th>
<th>Adaptive, Flexible Spider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deleted documents are immediately removed; new documents instantly added</td>
<td>Adjustable revisit frequencies, link depth, and directory depth</td>
</tr>
<tr>
<td>Finds new/changed content automatically</td>
<td>Can cross firewalls and index password-protected sites</td>
</tr>
<tr>
<td>Add URL lets users submit their own changes</td>
<td>Adaptable to handle the largest Intranets</td>
</tr>
<tr>
<td></td>
<td>Can be tuned to limit load on servers and networks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Completely Customizable</th>
</tr>
</thead>
<tbody>
<tr>
<td>User interface, results ranking, and more</td>
</tr>
<tr>
<td>Include/Exclude at the site, directory, or document level</td>
</tr>
</tbody>
</table>

### Add-in Modules

Verity provides a number of add-in components and modules that help you optimize your search functionality. Each module is listed below. For detailed information about these modules, see the Verity web site:

http://www.verity.com/products/ultraseek/features.html
Starting Ultraseek

This section assumes that you have already installed Ultraseek, and that you are using the default TCP port of 8765 for Ultraseek's built-in HTTP server. If you have not yet installed Ultraseek, refer to the installation guide on the Verity web site:

http://downloadcenter.verity.com/dlc/documentation.do

For more information on configuring Ultraseek on Solaris and Linux to start automatically when your system reboots, see to the Verity Ultraseek Installation Guide.

To start Ultraseek on Windows NT 4.0 and Windows 2000:

1. Click the Windows Start button.

2. From the Programs menu, select Verity Ultraseek, then select Verity Ultraseek Administration.

To start Ultraseek on Linux/Solaris using the default settings:

1. From a command prompt type:

   `/opt/Ultraseek/ultraseek start`

To start Ultraseek on Linux/Solaris from a specific port:

By default, Ultraseek uses port 8765. This enables you to run the search server on the same host as your normal web server. You can also run the search server on port 80, the standard port for HTTP. However, you should not run any other services on port 80 since this can cause conflicts.

Within the seekd script, specify the following parameters:

- an alternate data directory with `-d <directory>`
- an alternate TCP/IP port with `-p <port>`
For example, if you installed Ultraseek in /opt/Ultraseek, and you decided to use /var/opt/Ultraseek as the data directory with port 80 as its TCP/IP port, run Ultraseek by typing:

```
/opt/Ultraseek/bin/seekd -d/var/opt/Ultraseek -p80
```

If the server is running on a dedicated machine without a web server running on port 80, we recommend you use the `-p80` option, since this is the standard port for HTTP. You need to be super-user to start on any port lower than 1024.

---

**Accessing the Administrative Interface**

The Administrative Interface is a web-based user interface that provides a rich set of features for populating and tuning the search index, monitoring performance and site traffic, and configuring and tuning the Ultraseek server.

You access the Ultraseek administrative interface through your web browser.

**To access the Administrative Interface:**

1. Access one of the following URLs in a web browser:

   **Standard:**
   
   http://host:adminport/admin
   
   **SSL:**
   
   https://host:adminport/admin

   Where `host` is the name of the host where Ultraseek is installed and `adminport` is the number assigned to the Ultraseek search port (the default value for `adminport` is 8765).

   For example: http://search.verity.com:8765/admin

   Use the **https** address to access Ultraseek if you have restricted access to Ultraseek via the add-in SSL module that enables SSL connections; otherwise, use the standard **http** address.

2. If necessary, log on to Ultraseek with the administrator ID and password, or your user account.

   The administrator ID and password are set during the installation process. You can change the ID and password, as well as create and modify user accounts. For more information, see “Managing Administrative Users and Client Access” on page 166.
The Ultraseek administrative interface displays in your web browser, as shown in Figure 1-1.

**Figure 1-1  Administrative Interface**

The following table describes the icons available on each Ultraseek interface page.

<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restart/Shutdown</td>
<td>The shutdown button allows you to restart or shut down the Ultraseek server.</td>
</tr>
<tr>
<td>Search</td>
<td>The search button takes you to the Ultraseek search page.</td>
</tr>
<tr>
<td>Help</td>
<td>The help button takes you to the online help for Ultraseek.</td>
</tr>
</tbody>
</table>

The table below describes the Administrative Interface features for each main button and the corresponding tabs.

<table>
<thead>
<tr>
<th>Main Button</th>
<th>Tabs</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>Status</td>
<td>View the overall status of collections, including indexing status, number of threads, and the number of sites, URLs, and documents indexed by Ultraseek.</td>
</tr>
<tr>
<td></td>
<td>Performance</td>
<td>View the performance of Ultraseek, including the number of queries, the query time, and the query rate served by the Ultraseek server.</td>
</tr>
<tr>
<td></td>
<td>Queries</td>
<td>View a record of recent searches on your collections.</td>
</tr>
</tbody>
</table>
### Threads
View the number of HTTP server threads for each collection, and the status of those threads.

### Server Parameters
Set the following server parameters:
- **Main**: Main server parameters, such as the license key, hostname, binding address, and home URL.
- **Query**: Query parameters, such as the default query mode, default results filter, display number of hits, filter number of hits, and default look.
- **Advanced**: Advanced server parameters, such as the number of server threads, HTTP connection timeout, HTTP keep-alive timeout, and log levels.

### Users
Add and delete administrative users and regulate user access based on IP addresses.

### Doc types
Manage the document types that Ultraseek indexes and specify how document types should be parsed.

### XML Mappings
Create XML mapping fields to Fieldname, Element, Attribute Name, and Attribute Value fields.

### Quick Links
Add Quick Links to your search results pages.

### Ad Server
Modify Ad Server parameters.

### Collections Status
View the status of the indexing task, the sites currently being indexed, and the on disk path to the collection.

### Roots
Modify the root URL of a collection.

### Filters
Filter the sites and documents that are allowed in collections.

### Dupes
Control how Ultraseek defines and handles duplicate URLs.

### Network
Modify network settings for collections, such as authentication and spider throttle settings.

### Tuning
Set collection tuning parameters, such as crawling schedules, META tag definitions, and word spam thresholds.

### URL Status
View the URL status of a document, including whether a URL is allowed/disallowed and if it is in the URL database or search index.

### Add URL
Add a URL to a collection.
Accessing Additional Utilities

If you are using Ultraseek on Windows NT/2000, you can access additional utilities that are shipped as part of the software.

**Obtaining System Information**

The System Information utility provides you with a brief report of your platform, memory, and Ultraseek version statistics.

**To run the System Information Utility:**

1. Click the Windows **Start** button.
2. From the Programs menu, select **Verity Ultraseek, then select System Information**.

**Tuning the System File Cache**

On Windows NT the default File Cache setting is not optimized for the purposes of Ultraseek. The Tune System File Cache utility enables you to create a registry key that sets the File Cache more appropriately. You only need to run this utility once to correctly set the registry key for system file cache tuning.

**To run the Tune System File Cache Utility:**

1. Click the Windows **Start** button.
2. From the Programs menu, select **Verity Ultraseek, then select Tune System File Cache**.
Verifying the Server Installation

The Verify Server Installation utility checks the files in the install directory to determine if any files have changed from the default installation. This utility is useful to determine if any of the HTML pages in the Ultraseek docs directory have been customized. If any of the pages have been customized, Ultraseek displays a dialog that lists the modified pages.

To run the Verify Server Installation Utility:

- Click the Windows Start button.

  From the Programs menu, select Verity Ultraseek, then select Verify Server Installation.

Getting Help

Online help is available for each page in the Administrative Interface. Ultraseek also provides you with an application help page which contains links to Technical Support and product-related documentation, provides useful tips on how to test your search site, and enables user-level access for managing URLs and sites.

Accessing Online Help

To access online help for any page in the application, click the Help button at the top right of the page. These online help files contain reference information for each field in the application, and also provide overview information about the feature, if needed. Our context-sensitive explanations may be just what you need.

Accessing Application Help

To access general help for Ultraseek, click the Help button at the top right of the Administrative Interface. This page provides useful links to the Technical Support site and FAQs, to related product documentation, and to the software support mail list. This page also contains useful tips on how to perform searches to test your site. Non-administrative users can also use this page to manage URLs and re-index individual
sites. See “Managing Administrative Users and Client Access” on page 166 for more information about granting access to user-level administrative rights to features available under the More Services section of the help page.

More Services

The help page contains a section titled More Services that provides user-level access to Ultraseek features, including:

- About Verity Ultraseek
- Add URL—Allows you add a URL.
- URL Status—Shows the status of the URL(s) for the selected collection.
- View Sites—Views the sites currently associated with all of the running collections.
- Revisit Site—Request that all URLs from a specific site be revisited. This allows you to revisit a single site, rather than revisiting an entire collection.
- Server Administration—Takes you to the Server Administration pages.

Functionality available in the More Services section enable users to access Ultraseek features without having administrative privileges. This is useful in intranets or other environments where users can contribute to the search index.
Creating Collections

This chapter takes you through the creation of your first and subsequent collections, fine tuning, allowing and disallowing URLs, and other topics.

This chapter contains the following sections:

- About Indexing Documents
- About Creating Collections
- Spidering the Network
- Scanning the File System
- Following Usenet Groups
- Indexing Microsoft Exchange Public Folders
- Indexing Databases
- Creating Mirrored Collections
- Merging Collections
- Defining Filters
- Importing Collection Configurations
About Indexing Documents

Ultraseek groups information in collections. For example, you may have a collection from your site that includes information for searches on products, a collection used for searches on employment opportunities, and a collection used for searches encompassing your entire site. Users can choose to search in one collection, or in several. Administrators can reserve certain collections to be searched only from special areas of the web site, or can limit their use to certain individuals.

There are several ways to create new collections: by spidering the network, scanning the file system, following Usenet groups, indexing Exchange public folders, mirroring a remote Ultraseek collection, merging a set of collections, or indexing from the XPA API. You can select the collection type only when you define the collection; you cannot later change this information. If you want a different type of collection, you should re-index the data using that collection type.

**Note**  During the indexing process, Ultraseek follows links found in documents and indexes any additional content found there. However, it is recommended that your documents contain no more than 1,000 links. Ultraseek may reach internal limits for documents with large amounts of text or a larger number of links.

**HTTP Spidering**

The Verity Ultraseek spider “crawls” over parts of the World Wide Web. When spidering, Ultraseek works like an end user with a mouse and a browser. It follows HTTP links inside documents like web pages, and follow those links to other data. By default, the spider examines JavaScript in web pages for additional URLs. The spider later visits these URLs and indexes content found there. See the spider filtering options in Table 2-2 on page 65 for more information.

Ultraseek “sees” all of the different types of data, but only certain types can be added to a collection. This is generally any text based information, such as HTML coded web pages, text files, and certain other “text-like” files, such as Adobe PDF and Microsoft Word files. Other file types, such as image files (JPEG, GIF) are not added to collections, even if they appear to contain text.

The Ultraseek spider also accepts automatic updates from weblogs using the standard XML-RPC protocol. When Ultraseek receives a ping from a weblog, it determines whether the weblog’s URL passes the URL filters defined for existing spider collections.
If the weblog’s URL passes any of these filters, the Ultraseek spider crawls the weblog’s URL and updates the content of the corresponding collections. See “Filters” on page 37 for more information about URL filters. See “Adding Blog Entries to Spider Collections” on page 43 for more information about updating collections with weblog content.

MIME Types

Since spidering involves retrieving data using HTTP, the data is “served up” by a web server, such as Apache or Microsoft IIS. When served, it contains a MIME (Multipurpose Internet Mail Extensions) type. It is beyond the scope of this document to provide a detailed description of how MIME works. Besides using HTML, browsers also use MIME types. This enables the browser to display, or output files that are not in HTML format. These MIME types are also used by Ultraseek to decide which parser to use on the data. For more information on MIME types used by Ultraseek, see “Managing Document Types” on page 172.

Filters

From a functional standpoint, filters (when spidering) work as follows.

By its nature, any link on the WWW can lead to any other link on the web, provided a user had the patience to search through all the links. Ultraseek is just such a patient user. Therefore, there needs to be a list of allowed URLs that the spider can crawl to avoid cataloging the entire web. The spider starts at the URL root, and can proceed out from there as long as the links it finds are in its list of allowed sites. This also has the feature of disallowing sites implicitly. If a root is specified as

http://www.verity.com/

and there are links in various web pages within the verity.com domain, Ultraseek will follow them. Suppose that in addition to the allow filter for the verity.com domain, there is an “allow” filter for

http://products.verity.com/
http://services.verity.com/

The spider may find links to documents it can add to the collection at either of these domains, and since they are “allowed”, it will add them as appropriate. Now suppose that the spider finds a link to

http://employees.verity.com/

Since this isn’t specifically “allowed”, the link will not be followed, and no data from the collection will be added to any collection.
Also, please note that if no links exist in any of the documents in the root URL (http://www.verity.com/), to either of the two extra domains, then no documents from those domains will be added. If no links exist in http://www.verity.com/, and documents from http://products.verity.com/ need to be included in a collection, then http://products.verity.com/ would need to be added to the list of root URLs.

**Note** Filters are physically read in the order that they appear within the Ultraseek administrative interface. Each URL pattern is compared to the list of filters, and once one is encountered, it is applied, and the filtering process ends. Therefore, it is generally a good practice to place any “disallow” filters before the “allow” filters.

### File Scanning

File scanning “walks” a directory tree recursively, starting at the document root. It doesn’t follow HTTP links, only the directory structure. This means that anything not listed in the “roots” section for the file scan collection will not be indexed, even if there are HTTP links to data in other directories, or on other file systems. However, on Unix systems, Symbolic (soft) links would be followed, since they are functionally the same as having the data present within the directory structure. WinNT/2000 “shortcuts” would not be followed, since they are actually a file type not indexed by Ultraseek.

### File Types

Since data found using file scanning isn’t served up with a MIME type, Ultraseek has to add a step in order to determine what type of data it is looking at.

Under **Server | Doc Types**, there is a list of file extensions and their associated MIME types. Ultraseek uses this to determine the MIME type of the data being examined, and can then utilize the correct parsing tool for the data, if appropriate. See “Managing Document Types” on page 172 for more information.
Creating Collections

About Creating Collections

This chapter describes the common tasks for creating collections on a step-by-step, collection-by-collection basis. After you have installed Ultraseek, you are requested to create your first collection. Don’t panic! With Ultraseek, it is a simple process that should take a minimum amount of time. To create your first collection, use the instructions in “Spidering the Network” on page 40. This is a simple and straightforward way to create your first collection. Table 2-1 on page 39 shows the different types of collections you can create with Ultraseek.

Table 2-1  Ultraseek Collection Types

<table>
<thead>
<tr>
<th>Collection Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spidering the network</td>
<td>Ultraseek downloads web pages from HTTP servers, just like a web browser does. You supply the root URL of the first page to download, and the spider automatically follows links from the pages it finds. The default for link hops is set to 100. You can set the number of hops on the Filters page (see “Defining Filters” on page 59).</td>
</tr>
<tr>
<td>scanning the file system</td>
<td>Ultraseek includes all allowed files from the directories that you specify. You may limit the amount and type of data retrieved by MIME type, and by including/excluding certain directories. Remember, Ultraseek must have access to the directories in which content is stored, or no documents can be indexed.</td>
</tr>
<tr>
<td>following some Usenet news groups</td>
<td>Ultraseek communicates with your Usenet news server and includes articles from news groups you specify. This option is especially helpful for internal communications for projects, troubleshooting, and customer service.</td>
</tr>
<tr>
<td>indexing Microsoft Exchange public folders</td>
<td>This option is used to communicate with your Microsoft Exchange Server and includes the documents you specify. You can index Microsoft Exchange Public Folders that are published on the web using Microsoft Internet Information Server and Microsoft Outlook Web Access.</td>
</tr>
<tr>
<td>accessing a database</td>
<td>This option is used to index the rows from tables that you have chosen from your database.</td>
</tr>
</tbody>
</table>
Spidering the Network

A spider collection is constructed by downloading web pages from HTTP servers, just as a web browser does. You can also configure spider collections to accept automatic updates from weblogs.

Creating a Spider Collection

To create a new spider collection:

1. Click Collections.
2. Click new.
   
The New Collection window displays.
3. Enter an **Internal Name** for the collection.

   The internal file name should not be longer than 8 characters, and can only contain letters and digits. For example, “Docs”, “west”. Special characters such as *, &, _, #, @, are not supported.

4. Enter a **Display Name** for the collection.

   This is the viewable name Ultraseek uses on the HTML search page and can contain up to 32 characters. For example, “Documentation”, “West Coast Sites”.

5. Select the **Primary Language**.

   The primary language choice determines the default stemmer to use when a document is inserted into the collection and when searches are performed on the collection. If Ultraseek cannot determine the actual language of a document or query, it will use this language. For example, if you specify English as the primary language, the English
stemmer will be used for processing documents and queries where Ultraseek could not determine the language.

6. Choose **spidering the network**.

7. Click **ok**.

The Root Specification page displays, which you use to specify the root location of the content that you want to index.

**Figure 2-2 Root Specification Page for Spidering Networks**


If your site has multiple home pages that may not be linked to each other, you can enter additional root URLs. The indexer starts with these documents and follows any links, indexing those URLs as it goes. A site’s home page makes a good root URL.

9. Click **ok**. The URL Filter Specification page displays. See “Defining Filters” on page 59 for more information on creating filters.

10. Click **ok**.

Ultraseek will now start indexing your site. The amount of content on your web site will determine the length of time Ultraseek needs to complete the indexing. You can click the **Refresh** button in your browser to see the progress of the indexing task.
Adding Blog Entries to Spider Collections

A blog, or weblog, is a journal available on the internet. Most blogs are updated frequently, so it is recommended that you allow weblogs to notify Ultraseek when new content is available, to ensure that new entries are visited promptly. Ultraseek automatically adjusts revisit intervals to visit active blogs more frequently.

**Figure 2-3** Updating collections with blog entries

Figure 2-3 illustrates the process of adding blog content to a spider collection:

1. A blog user posts an entry to the blog server.
2. The blog server uses the XML-RPC protocol to ping Ultraseek about the newly available content.
3. Ultraseek adds the URL to all matching collections
4. Threads allocated to the matching collections visit and index those URLs.

The blog’s URLs must be allowed by Ultraseek’s URL patterns.

To add blog content to spider collections:

1. Edit the URL patterns on the **Collection** | **Filters** page to allow content from the blog server. Do this for each collection to which you want to add blog content.
2. Configure your blog server to ping your Ultraseek server whenever the blog is updated with new entries. Append `/help/ping.xml` to your search application’s URL, as shown in the following example:

   http://www.myultraseekserver.com:8765/help/ping.xml
Scanning the File System

When you create a collection for scanning the file system, Ultraseek includes all allowed files from the directories that you specify. Ultraseek must have access to the directories in which content is stored, or no documents can be indexed.

To scan a file system:

1. Click the Collections | New.
2. Enter an Internal Name for the collection.
3. Enter a Display Name for the collection.
4. Select the Primary Language.
5. Choose scanning the filesystem.
6. Click ok. The Root URL Specification page displays, which you use to specify the root location of the content that you want to index:
7. Enter the **Directory Pathname(s)** or optional **URL(s)** in the fields provided. You can enter more than one pathname and/or URL for various directories in your file system.

---

**Note**  If you specify a directory path, use the standard syntax for pathnames on your platform. All allowed documents in these directories are added to the full text index. If you do not supply a URL, a file URL is automatically derived from the pathname.

---

You must enter at least one directory pathname and the corresponding optional URL to indicate the documents you want included in your collection. A root should be a directory pattern ending with a “/”.

8. Click **ok**. The URL Filter Specification page displays. See “Defining Filters” on page 59 for more information on setting filters.

9. Click **ok**.

   Ultraseek will automatically start indexing the specified paths and URL.
2 Creating Collections
Following Usenet Groups

Following Usenet Groups

When a collection is created and populated by following Usenet news groups, Ultraseek communicates with your Usenet news server and include articles from news groups you specify. This section describes the steps necessary to create a collection from a news group. The procedures are similar to those of “Spidering the Network” on page 40.

To create a collection for a Usenet Group:
1. Click Collections | New.
2. Enter an Internal Name for the collection.
3. Enter a Display Name for the collection.
4. Select a Primary Language.
5. Select following some Usenet news groups.
6. Click ok. The Root Specification page displays.

7. On the Root Specification page, enter the Hostname of your NNTP server that you want to use for this collection. For example, news.mycompany.com.
9. Create any group filter to allow, or disallow groups. Or simply type “*” (without the quotes) in one of the windows, specifying in the Action Syntax pulldown windows “Allow”, and “Wildcard” to index the entire newsgroup.
10. Select the Allowable Languages.
11. Click **ok**.

Once you click **ok**, Ultraseek will begin to automatically index your newsgroups. If there is a large amount of documents, memory usage and time needed to complete the task will increase. For faster indexing, you can reduce the number of indexer threads Ultraseek uses to index the documents.

**Indexing Microsoft Exchange Public Folders**

Ultraseek can index Microsoft Exchange Public Folders that are published on the web using Microsoft Internet Information Server and Microsoft Outlook Web Access. When a collection is populated by indexing Microsoft Exchange Public Folders, Ultraseek communicates with your Microsoft Exchange Server and includes the documents you specify.

To index Microsoft Exchange public folders:

1. Click **Collections** | **New**.
2. Enter an **Internal Name** for the collection.
3. Enter a **Display Name** for the collection.
4. Select a **Primary Language**.
5. Select **indexing Microsoft Exchange public folders**.
6. Click **ok**. The Root Specification page displays:

![Figure 2-6 Root Specification Page for Microsoft Exchange Public Folders](image)
7. Enter the hostname (and optional :port) of the IIS server that is providing web access to the Exchange Public Folders.

For example, if you connect to your Exchange Public Folders on the web using the URL http://server.mycompany.com:8080/exchange/, your root specification would be: server.mycompany.com:8080.

8. Once you click **ok**, Ultraseek will automatically start the indexing process. To view the status, click “Refresh” in your browser.

Once Ultraseek has completed indexing the folders, you can go to the Filters page to view the default filter specification (see “Defining Filters” on page 59). When indexing Exchange V5.5 public folders, the default filter is:

/Public Folders/*

When indexing Exchange 2000 indexes the default filter is:

/public/*(2000)

Both Exchange filters are presented as defaults because Ultraseek does not assume any version preference.

## Indexing Databases

You can create collections from documents stored in supported databases. However, an Ultraseek database collection is not a native database indexing interface. You should not attempt to use a database collection as an ODBC Gateway. Each record in the database represents one document. Larger records require more time to index, so you should attempt to limit the size of database records to around 100 KB.

Ultraseek uses a driver to connect to a database. This driver needs to be installed before the database can be indexed and must be configured by the administrator of the machine where the Ultraseek server is running. The database does not need to be on the same machine as Ultraseek, but it does need to be accessible over the network.

In general, database collections are more difficult to set up than an HTTP spider collection. When possible, we recommend that you use an HTTP spider collection rather than a database collection to serve content to your users. In addition, HTTP spider collections always index content exactly as your users see it. Ultraseek does not format data indexed from a database, so users see raw data unless your application server is configured to format it.
Note Refer to the Windows ODBC documentation or the documentation that came with your third-party database driver for detailed information about installing database drivers.

Supported Databases

Ultraseek supports the following databases:

- Microsoft ODBC for Windows NT/2000/2003
- Oracle 7/8, Microsoft SQL Server, and Sybase Adaptive Server on Linux and Solaris.

The table below describes the databases and platforms that are supported:

<table>
<thead>
<tr>
<th>Database/Platform</th>
<th>Windows NT/2000</th>
<th>Solaris</th>
<th>Linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Access</td>
<td>ODBC</td>
<td>Not supported</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Microsoft SQL Server</td>
<td>ODBC</td>
<td>Native Client</td>
<td>Native Client</td>
</tr>
<tr>
<td>Oracle</td>
<td>ODBC</td>
<td>Native Client</td>
<td>Native Client</td>
</tr>
<tr>
<td>Sybase</td>
<td>ODBC</td>
<td>Native Client</td>
<td>Native Client</td>
</tr>
<tr>
<td>IBM DB2</td>
<td>ODBC</td>
<td>Native Client</td>
<td>Native Client</td>
</tr>
</tbody>
</table>

Configuring Windows NT/2000/2003 Drivers

For Windows NT/2000/2003, either the database developer or a third party must supply an ODBC driver for the database. Ultraseek does not include ODBC drivers. If you are running Ultraseek on Windows, you connect to databases using ODBC. To install database support for NT drivers, you set up ODBC data sources.

To provide ODBC support for Windows NT/2000/2003, use the ODBC Data Source Administrator tool located in the Windows Control Panel to set up the System DSN for your database driver. The following instructions provide you with an example of how to install an ODBC database driver on Windows 2000.

To configure an ODBC driver for Windows 2000:

1. On Windows 2000, open the Control Panel and double-click Administrative Tools (on Windows NT, open the Control Panel and double-click ODBC Data Sources) then Data Sources (ODBC). The ODBC Data Source Administrator window displays.
2. Select the System DSN tab and click Add to open the Create New Data Source window.

3. Choose the driver that you want to install from the list of available drivers. For example, Microsoft Access Driver.

4. Click Finish.

5. Enter the information that is specific to the database that will be indexed. For example, an Access database ODBC driver requires a Data Source Name and Description.

6. Click Select under Database to open the Select Database window.

7. Select the database you want to use and click OK.

   You will be returned to the ODBC Data Source Administrator window where the new ODBC database driver will appear.

8. Click OK.

   Once the ODBC driver is configured, you can build a new collection that will index content from the database.

**UNIX Drivers**

You can only connect to Oracle databases through the native Oracle Client driver. You need to install the Oracle Client software on the same machine as Ultraseek, and configure the tnsnames.ora file in the Oracle installation. This software is supplied by Oracle. The ORACLE_HOME environment variable must be set in database/oracle_config under the Ultraseek program directory. Please read the comments in database/oracle_config for more information. As with ODBC, once the Oracle Client is installed, the Ultraseek administrative interface can be used to index the database.

**Creating Database Collections**

Before you start to index a database, you need to know the following:

- Name and location of the database
- The character set of the database if different from ISO 8859-1

**To index a standard ODBC database:**

1. Create a new collection to access a database, as described on “Spidering the Network” on page 40. After you define the basic collection information, the Root Specification page displays, as shown in Figure 2-7.
2. Specify the database information.

On Windows NT/2000, enter the system data source name as specified in the ODBC Manager as the Database Name. For example, if you add an Access driver, you might name your data source “Access”.

On Linux or Solaris, fill in all the appropriate information for the database you are connecting to, including the Database Name, the Hostname, and the Port. If you are connecting to an Oracle database, enter the Oracle SID as specified in the TNSNAMES.ORA file in your Oracle Client installation.

3. In the Record Display region, choose whether to use the built in page or a custom page to display the database record when it is clicked on in a search result.

If it is set to Use built-in page, Ultraseek delivers the record to the user when they click on a link in the search result, as shown in the following figure.
If it is set to **Use this URL**, the URL entered in the box will be returned in the search results. This option is to be used in conjunction with a custom search page which has been developed for your enterprise. If you do not have a custom page created for such a purpose, use the default option **Use built-in page**.

If you have created a custom search page, the Record Display URL is invoked with the following parameters:

<table>
<thead>
<tr>
<th>col</th>
<th>The name of the collection.</th>
</tr>
</thead>
<tbody>
<tr>
<td>tb</td>
<td>The name of the database table.</td>
</tr>
<tr>
<td>KEYS</td>
<td>For each column selected as a primary key, a parameter of the form KEYNAME=VALUE will be generated.</td>
</tr>
<tr>
<td>charset</td>
<td>The character set the result should be returned in.</td>
</tr>
</tbody>
</table>

For example, if the internal name of your database collection is `dbcol`, indexing a table named `employees`, and one of your database search results is a record whose primary key `EMPNO` has a value of 42, the URL used to display the search hit will be:

```
<contents of Record Display URL field>?col=dbcol&
```

Where the entry for `<contents of Record Display URL field>` may be:

```
db.corp.com/db
```

The entry would then look like this:

```
db.corp.com/db?col=dbcol&
tb=employees&EMPNO=42&charset=iso-8859-1
```

4. Click **ok** to continue. The Authentication page displays.

5. On the Authentication page, click **new** to enter the username and password that Ultraseek should use to access the database.
6. On the New Database Authentication Specification page, enter the username and password and click **ok**.

7. Ultraseek brings you back to the Authentication page. Click **ok** to continue.

8. On the next page, select the allowable languages, then click **new** in the Tables region to add tables to the index.

   The indexer will also check whether a row is allowed based on the language of the row and the allowable languages. If the allowable languages include “Any”, rows in any language are allowed. If a row’s language is not known or is among the allowable languages, the row is allowed. Rows in the primary language defined for the collection are always allowed.

9. On the Collection Default Encoding page, select the default encoding for the collection.

   Since table and column names can be stored in character sets other than ISO-8859-1, Ultraseek needs to know what the default character encoding for the collection should be. Typically, it should be left at the default value of Western (ISO-8859-1).

   If the database contains table names, column names, or column data that is stored in another character set, specify that character set here.

10. Click **ok**.

    The New Database Table page displays, as shown in “New Database Table Page” on page 53.

   ![New Database Table](image)

   **Figure 2-8** New Database Table Page

11. Select the table that you want to index and click **ok**.
You can also specify an owner for the table. Select **Default** to use the default table owner, or select **Custom** and enter a table owner manually in the text field. In many situations, a table owner does not need to be specified. If this is the case, select **None**.

The Edit Table Keys page displays, as shown in “Edit Table Keys Page” on page 54:

**Figure 2-9** Edit Table Keys Page

12. On the Edit Table Keys page, select the primary key(s) and click **ok**.

The Index Columns page displays, as shown in **Figure 2-10**.
13. On the Index Columns page, select the columns that you want to index and click **ok**.

The Table Columns Usage page displays:

**Figure 2-10 Index Columns Page**

**Index Columns**

Select the Table Columns to be added to the Search Index:

- ✔ ProductID
- ✔ ProductName
- ✔ SupplierID
- ✔ CategoryID
- ✔ QuantityPerUnit
- ✔ UnitPrice
- ✔ UnitsInStock
- ✔ UnitsOnOrder
- ✔ ReorderLevel
- ✔ Discontinued

**Figure 2-11 Table Column Usage Page**

**Table Column Usage**

Select the Table Columns to be used for the following purposes:

- Title: <ProductName>
- Description: <ProductName>
- Keywords: <ProductName>
- Data: <ProductName>
- URL: <ProductName>
14. Select the columns that should populate the fields shown on the page. The Date field can only be set to columns that have a date type. The URL field uses the value of the column to replace the URL displayed on the search results page.

15. Click ok to continue.

16. You can now add more tables or click ok again to start the indexing process.

Once the index starts to build, you can go to the search page and search the index. If you click on one of the default links, it will send you to a page that shows you the database record.

Creating Mirrored Collections

When a collection is populated by mirroring a remote collection, Ultraseek copies the search index from another remote mirrored site into the new collection. This allows for faster searching and results generation since the content from both sites is present in one index. After you create a mirrored collection, you can tune the collection by setting parameters specific to mirrored collections, such as setting the Polling Interval, and also modify the mirroring schedule to indicate when the mirror is allowed to update the collection.

**Note**

Ensure that you have enough disk space to store the contents of the mirrored collection.

**To create a mirrored collection:**

1. Creating a new collection to mirror a remote collection follows the procedure outlined in “Spidering the Network” on page 40. After you define the basic collection information, the mirrored collection Status page displays.

2. Specify the information described in the following table.
Remote site

Specify the remote Ultraceek site from which the full text indexes should be copied. This site name should be expressed the same way a site name in a URL is expressed—a host name and an optional colon and port number.

Remote collection internal name

Enter the internal name of the collection on the remote Ultraceek site that should be copied.

Proxy server (optional)

Specify the proxy server to use when communicating with the remote Ultraceek site.

Proxy server username and password (optional)

Specify the proxy username and password to use when communicating through the proxy server.

Load CCE topics from remote

If you are using CCE, and check this option, topic definitions will be mirrored from the remote site along with the collection data. The mirrored topics will be applied to all collections on this server. To edit the topic definitions, use the Ultraceek administrative interface on the remote site.

3. Click ok.

After you create the mirrored collection, the status of the collection changes from idle to polling. For more information on the Collection Status page, see “Basic Collection Management” on page 74.

While Ultraceek is gathering data, you can set Ultraceek’s parameters by clicking the Tuning tab. See “Setting Tuning Parameters” on page 100 for more information.
Merging Collections

Merged collections are built by merging the indexes from multiple collections to create a single set of text indexes. This is useful if you have many collections and you want to produce a single set of indexes for search. For example, you might merge several spidering collections that use cooperative spidering. For more information on cooperative spidering see “Automatic URL Adder (Cooperative Spidering)” on page 99.

To merge collections:

1. Click Collections | New.
2. Enter an Internal Name for the collection.
3. Enter a Display Name for the collection.
4. Select a Primary Language.
5. Select Merging a Set of Collections. The merge collection Status page displays.

Figure 2-12 Merging Collections Status Page

6. Under Collections to Merge, select the collections that you want to merge.
7. Click the merge button.
8. Click ok on the confirmation screen.
Creating Collections Using the XPA API

Indexing using the XPA API is covered in the *Verity Ultraceek XPA Programming Guide*.

Defining Filters

The Filters Specification page specifies how you want Ultraceek to index your site. URL filters tell the spider where to crawl within your web site or intranet. It is here that you specify the allowed and disallowed URLs.

This page generally displays when you create a new collection. You can also access this page by selecting the Collections tab, then selecting the Filters tab.

If you are only going to index the root URL and you do not have a preference on language, you can skip the Filters page. However, you can set specific filters to further customize the content that is included in the collection.

If you do not specify any filters, Ultraceek will completely index the site that you specified on the Root URL page. If Ultraceek is building the collection after you specified custom filtering parameters, then all root URLs must also pass the URL filters for spidering to begin.

There are four steps to defining filters:

1. “Specifying Patterns” on page 59
2. “Allowing or Disallowing Sites” on page 64
3. “Setting Filter Options for Spider Collections” on page 65
4. “Setting Allowable Languages” on page 67

Specifying Patterns

Each pattern and its corresponding action tells the spider what to do when it encounters a new URL. Starting with the top pattern of the URL Filter Specification page, and proceeding toward the bottom, the spider compares the URL with each pattern you specify. When the spider finds a pattern that matches the URL, it performs the
corresponding action you have specified. Only one action per URL is allowed. Working from top to bottom, the first matching pattern indicates the action to take. The possible actions are:

- allow the URL to be indexed
- disallow indexing of the URL

Enter at least one URL pattern or otherwise allowed site (see “Allowing or Disallowing Sites” on page 64) to indicate which documents you want to allow into the collection.

**Note** When creating File System collections the pattern should be a file name pattern. To include content within a directory, enter a wildcard (a URL ending with an asterisk “*”).

Next, select one of the spidering options described in the following table.
Creating Collections

Defining Filters

Each pattern with its corresponding action tells the spider what to do when it encounters a new URL. The spider compares the URL with each pattern. If the URL matches the pattern, the spider performs the action.

<table>
<thead>
<tr>
<th>Allow</th>
<th>The spider indexes the text and follows links in the specified documents.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allow text</td>
<td>The spider indexes the text, but does not follow links to URLs on these pages.</td>
</tr>
<tr>
<td>Allow links</td>
<td>The spider follows links from the document, but ignores the text. This is useful for starting an index at a directory page, without including the directory page itself in the index.</td>
</tr>
<tr>
<td>Disallow</td>
<td>The spider ignores both text and links in the specified document.</td>
</tr>
</tbody>
</table>

Disallowing Directories

If specific directories below the root URLs should not be indexed, choose Disallow from the drop-down list. The spider does not follow links to URLs in the Disallow list, and the pages are not indexed. Wildcards are supported by Disallow so that you can exclude an entire set of URLs or servers with a single entry.

Allowing Directories

Ultraseek does not begin indexing until you enter at least one Allowed URL filter. This is typically the root for your site, followed by *.

http://www.company.com/*

Ultraseek automatically creates an Allow filter from your root URL when you create a new collection.

When you set the root URL, and the URL Allow filter to your site’s home page, Ultraseek immediately begins indexing your site. However, Ultraseek will not follow links to other systems because they will not match the filter. Therefore, in order for Ultraseek to follow links to other servers you must enter a new filter and replace the server name with a wildcard (*). This should look similar to the syntax below:

http://company.com/*
http://*.company.com/*
Sequence of Allows and Disallows

Ultraseek filters URLs sequentially. The spider starts at the top of the list of URL patterns and attempts matches as it proceeds through the list.

The first match encountered indicates the action to take. It is very important that you list filters in order of importance. To exclude a subdirectory, the disallow statement must appear first in the list, followed by the allow for the top-level directory.

For example, it is not valid to set the first item in the list is to be allowed as:

**Figure 2-14 Incorrect Entry for Disallowing Directories**

![Incorrect Entry for Disallowing Directories](image1)

The correct sequence is to first disallow the subdirectory:

**Figure 2-15 Correct Entry for Disallowing Directories**

![Correct Entry for Disallowing Directories](image2)

The reason for this is simple: if the root URL is placed in the Disallow box, Ultraseek will disallow all links from the root URL, including all subdirectories. However, if the subdirectory is placed first, Ultraseek will find the subdirectory and ignore it while indexing the remaining sections of the URL pattern.
Using Wildcard and Regular Expression Patterns

You can identify wildcard patterns or regular expressions by making the appropriate selection from the syntax drop-down list. Wildcard patterns allow you to include entire directories by using * in the pattern. For example, your list of URL patterns might look like Figure 2-16.

The first entry excludes from the index all files with a .pdf extension. The second entry excludes all documents in the archive subdirectory. The third entry allows everything else from the site http://www.mycompany.com/.

By selecting Regex in the Allow drop-down box, you choose to identify specific patterns by using regular expressions. You can make individual selections for each pattern you define. If you are not familiar with regular expressions, use Wildcard, the easier option.
Allowing or Disallowing Sites

The Otherwise Allow Sites and Always Disallow Sites regions enable you to use lookup tables for sites in the event that Ultraseek does not find a match between URLs and patterns that you define.

**Figure 2-17** Allowing and Disallowing Sites

<table>
<thead>
<tr>
<th>Otherwise Allowed HTTP Sites</th>
<th>Always Disallowed HTTP Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Otherwise Allowed HTTPS Sites</th>
<th>Always Disallowed HTTPS Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Otherwise Allowed FTP Sites</th>
<th>Always Disallowed FTP Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Otherwise Allowed Sites**

The **Otherwise Allowed Sites** setting allows you to use an existing list of URLs for top-level sites. If no match between the URL and a pattern is encountered and the protocol is HTTP, HTTPS, or FTP, the site part of the URL is looked up in a table built from lines in the **Otherwise Allowed Sites** box.

If the site part of the URL is in the appropriate **Otherwise Allowed Sites** table, both text and links are allowed. Otherwise, the URL is disallowed. Lookups in the **Otherwise Allowed Sites** table are very fast. There can be hundreds of otherwise allowed sites without any significant spider performance penalty.

Never include subdirectories in the **Otherwise Allowed Sites** setting. Only top-level sites are acceptable.

The following specification is acceptable:

http://www.company.com/

The following specification is not acceptable:

http://www.company.com/support/
Always Disallowed Sites

You can override all other filter specifications and always disallow all documents from a particular site by entering the top-level site name in the appropriate Always Disallowed Sites box. The URL must follow the HTTP, HTTPS, or FTP protocol pattern. If the site part of the URL is in the Always Disallowed Sites table, it is disallowed, no matter what another part of the filter specification says. Lookups in the Always Disallowed Sites table are very fast. There can be hundreds of Always Disallowed Sites without any significant spider performance penalty.

Never include subdirectories in the Always Disallowed Sites setting. Only top-level sites are acceptable.

Acceptable:

http://www.anysite.com/

Not acceptable:

http://www.anysite.com/subsite/

Setting Filter Options for Spider Collections

If you are creating a spider collection, you will see additional fields on the URL Filter Specification page, shown in and described in the Table 2-2.

Table 2-2  Spider Filter Descriptions
### Maximum number of directories in a URL
Indicates the longest URL the spider should allow. You should set it large enough to index interesting documents, but small enough to keep the spider from wasting too much time and space. Files in the root directory of a web site have a directory depth of 0.

### Maximum number of hops from root URL
Indicates how far the spider should follow links. It is also helpful for catching infinite loops in site document hierarchies, and limiting the size of your collections. Be careful if you use this setting to limit the size of your collections; the number of accessible documents can grow explosively with the number of hops.

### Follow links found within JavaScript
Instructs the spider to search for additional URLs by parsing JavaScript found in the following locations:
- Between `<script>` tags
- In event handlers within standard tags
- In strings contained in HTML “`href`” attributes

If discovered URLs pass collection filters that you have defined, the spider crawls them the next time it retrieves content. However, URLs generated by client-side JavaScript are ignored since the spider does not execute JavaScript.

### Disallow all URLs to CGI scripts
By default, the indexer automatically disallows any URLs that include the text `/cgi-bin/` or `.cgi`. Selecting this setting is equivalent to entering these example rules at the top of the filter:

```
http*://*/cgi-bin/* [disallow]
http*://*/*.cgi* [wildcard]
```

Clear this check box to disable this built-in rule and allow all URLs to CGI scripts.

Note: Clearing this setting can put a substantial load on a server with many CGI-generated documents.

### Disallow all URLs with query strings
The indexer normally disallows any URLs that include a question mark. Selecting this setting is equivalent to entering these rules at the top of the filter:

```
http*:///*/?* [disallow]
[ wildcard]
```

Clear this option to disable this built-in rule and allow all URLs with query strings.

### Disallow text from apparent directory listing documents
Select this option to prevent the text of these documents from being indexed. A document is considered to be a directory listing if its URL ends with `"/"` and its title is “Index of /dirpath/”, `"/dirpath/ ="` or “hostname.domain - dirpath/".”
Disallow FrontPage special directories
This option controls the indexing of documents in special directories used by the Microsoft FrontPage Server Extensions. This rule is equivalent to entering these filter patterns in the topmost disallow box in the URL patterns section. For example:

http*://*/_vti_* [disallow]
http*://*/_derived/* [wildcard]
http*://*/_overlay/*

Disallow all links from apparent directory listing
This feature is similar to “Disallow text from apparent directory listing documents”. In this case, you control whether or not links are allowed in these documents, as opposed to whether text is allowed.

Filter Lotus Domino navigation links
Instructs the spider to avoid following an infinite number of dynamically–generated Lotus Domino navigation links, and instead just follow enough to get the Lotus Domino documents. See the online help for a detailed example of how this setting compares to creating equivalent rules at the top of the filter.

Disallow parent directory links from apparent directory listing documents.
Prevents the spider from following parent directory links from documents that are directory listings.

Log disallowed URLs
Records a log entry in the error log for every URL disallowed by the URL filter. Because the log files can get quite large if there are many disallowed URLs, this feature should not be continually running. Instead, run the “Log disallowed URLs” if the spider seems to not be indexing your web site properly, or if this is a new installation, and you are testing the filter properties on your web site, or server.

Setting Allowable Languages

If a URL is allowed, the spider will also check whether each document in that URL is allowed based on the language of the document against the allowable languages you have specified. If the allowable languages include “Any”, documents in any language will be allowed. Documents are also allowed if the language of the document is not known (the document’s language is treated as belonging to the chosen language) or is among the allowable languages. Documents in the primary language are always allowed.

Note
If a document is in a language other than any installed language modules, Ultraseek attempts to parse the language based upon the installed language modules. However, this parsing will not be as accurate as when the correct language module is installed.
If the text of a document is allowed by the URL filter, the spider determines whether the
text may be inserted in the collection based on the language of the document and the
Allowable Languages. Documents that match a disallowed language are not included in
the collection, except as follows:

- Documents in the primary language (specified when you create the collection) are
  always allowed.
- If, when you created the collection, you specified Any allowable language, documents
  in any language are allowed.
- If a document's language is not known (treated as the default, or chosen language, or
  is among the allowable languages), the document is allowed.

Importing Collection Configurations

Importing a collection causes Ultraseek to re-create the original collection on your local
server. An imported collection is not a mirror collection, however; an imported collection
is not updated when the original collection changes.

You can import collection configurations from Ultraseek server and Verity Information
Server.

Importing from Ultraseek

Ultraseek collection configurations are text files named configuration that are located in
the directory you specified during installation to hold search data. This section explains
how to import these files so that Ultraseek can re-create the described collection.

You can import collection configuration from any version of Ultraseek. Importing
Ultraseek collections is useful in the following situations:

- Upgrading Ultraseek
- Migrating collections between Ultraseek servers
- Cloning a collection for testing
- Deploying a collection from a staging server to a production server
When you import a collection configuration, you do not import the collection’s index. Rather, Ultraseek uses the configuration information, contained in a file named configuration, to create a collection on the local server with the same configuration information as the imported collection.

Importing a collection does not set up replication of the original collection; changes made to the original collection are not migrated to the imported collection.

To import a collection configuration from Ultraseek:

1. Click Activity | Status.

2. Click Create a New Collection.

   The New Collection page is displayed.

3. Click Import ...

   The Import Collection page is displayed.

4. Click the “import Verity Ultraseek Server collections” radio button and click OK.

   The Import Collections from a Verity Ultraseek Server wizard is displayed.

5. Click one of the following three radio buttons and fill in the corresponding text box or boxes, then click ok:

   - Import from a running Verity Ultraseek Server
     Enter the fully-qualified domain name of the remote server, its port number, and the administrative username and password.

   - Upload a configuration file
     Enter the path to the Ultraseek collection configuration file.

   - Paste a configuration file or OSStatus report
     Open the collection configuration file in a text editor, copy the entire contents, then paste the contents into the text box. Alternatively, paste an OSStatus report into the text box.

     Ultraseek displays a list of collections described by the configuration file.

6. Verify that the check box next to each collection you want to import is checked and click ok.

   The Activity | Status page is displayed, showing that the imported collection is suspended. The new collection is left in a suspended mode so that you can make changes to the configuration before you index the collection on your local server.
Creating Collections

Importing Collection Configurations

**Note** If you import a database collection, Ultraseek does not verify that the required database drivers are installed on your local machine during the import procedure. Instead, an error message is displayed when you attempt to activate a remote database collection if the drivers are not installed locally.

**Note** A merged collection must be imported with the collections being merged.

**Importing from Verity Information Server**

You can import collection configurations created with Verity Information Server 3.6.1 to Ultraseek. Ultraseek requires access to the following Information Server files to perform the migration:

- `job.db`
- `inetsrch.ini`

It may be possible to import configurations from other versions of Information Server that create `job.db` and `inetsrch.ini` files. However, Ultraseek does not support versions of Information Server other than 3.6.1.

**Note** These files must be located either on the same machine that runs Ultraseek or be accessible through a network file system. After you click OK, the next page in the wizard indicates whether and where it located these two files. If Ultraseek cannot locate them in the directory path that you specified, it asks you again to specify their location.

The following restrictions apply to migrating Information Server collection configurations:

- The Information Server collections must have been created using Information Server’s web administration tool; Ultraseek cannot import collections created with Information Server’s command-line spider.

- Information Server collections created on platforms other than Windows NT 3.5.1 and 4.0 must first be manually copied to one of those platforms before migrating the collection configurations to Ultraseek.
To import a collection configuration from Verity Information Server:

1. Click the Activity | Status.
2. Click Create a New Collection.
   The New Collection page is displayed.
3. Click Import ...
   The Import Collection page is displayed.
4. Click the “import Verity Information Server collection configuration” radio button and click ok.
   The Import Verity Information Server Configuration wizard is displayed.
5. Type the path to the Verity Information Server installation directory and click ok.
   The wizard indicates the locations of the configuration files on the local machine.
6. Click ok.
   The wizard displays a checked list of all Verity Information Server collection configurations found in the installation directory.
7. Uncheck all collection configurations that you do not want to import and click ok.
2 Creating Collections
Importing Collection Configurations
Managing Collections

You can perform the following collection management tasks to control the indexing of each collection:

- Basic Collection Management
- Managing URLs
- Filtering Duplicate URLs
- Improving Document Relevance Scores
- Setting Network Parameters
- Tuning Collections
Basic Collection Management

The Collection Status page provides a complete overview of the current collection, including the number of sites indexed and active, the number of documents in the index, URLs pending and complete, and the structure of in-memory and on-disk databases.

Using this page, you can create, edit and delete collections, and perform additional collection management tasks such as suspending and resuming indexing, clearing and reindexing collections, merging collections, and viewing log files and site reports.

**Note** You can also suspend, resume, clear, and delete collections from the Activity Status page.

**To access the collection status page:**

1. Click Collections.
2. Click the Status tab.
3. Select the collection that you want to manage from the drop-down list.

![Collection Status Page](image.png)

*Figure 3-1 Collection Status Page*
The table below provides a quick reference for the fields on this page; collection management tasks are provided on the pages that follow.

<table>
<thead>
<tr>
<th>suspend/resume</th>
<th>Click suspend/resume to temporarily suspend or continue indexing.</th>
</tr>
</thead>
<tbody>
<tr>
<td>revisit</td>
<td>Revisit all URLs in the collection to check for changes. Ultradeek provides you with three revisit options:</td>
</tr>
<tr>
<td></td>
<td>- Start revisiting every URL, but only reindex each document if it has changed.</td>
</tr>
<tr>
<td></td>
<td>- Start revisiting every URL, and reindex each document whether or not it has changed.</td>
</tr>
<tr>
<td></td>
<td>- Do not start revisiting every URL, but reindex each document whether or not it has changed.</td>
</tr>
<tr>
<td>clear</td>
<td>This option removes the URL databases and search indexes, and rebuilds the contents of the collection from scratch.</td>
</tr>
<tr>
<td>view sites</td>
<td>Displays a report that lists all of the sites included in the collection.</td>
</tr>
<tr>
<td>view log</td>
<td>View log files that contain information on collection spidering and indexing activity.</td>
</tr>
<tr>
<td>merge</td>
<td>Merge indexes.</td>
</tr>
</tbody>
</table>

### Editing Collections

To change the state or parameters of a specific collection, select the collection name from the Collection list. You can then click the appropriate tab in the Ultradeek administrative interface to edit the collection’s content.

### Deleting Collections

To delete a collection, select the name from the Collection list and then click delete. When you delete a collection, Ultradeek removes all information associated with the collection, including root URLs, filters, and the collection name.
Clearing Collections and Reindexing Documents

To delete the contents of a collection’s index and rebuild the entire collection from scratch, click the clear button on the Collection Status page.

When you clear a collection, Ultraseek deletes the contents of the database, search indexes, and site URL queues for the collection. Documents previously added or removed individually are not reflected in the new index. It will also cause Ultraseek to restart.

Once the collection is cleared, Ultraseek will begin indexing again, using the information defined for the specific type of collection.

Suspending and Resuming Indexing

You can choose to suspend or resume the collection. The current status, noted in the upper left corner of the Status card, is either idle, suspended, or recovering. To change the collection’s status, click suspend or resume on the Collection | Status page.

Click suspend to temporarily stop the system from indexing. The index remains searchable during a suspension. Clicking resume causes the indexer to pick up indexing from where it was suspended. Ultraseek captures any changes you make to collection parameters during suspension.

Revisiting Sites and Reindexing Documents

Click revisit on the Collection Status page to cause Ultraseek to begin checking all known URLs in the collection for any changes. This is useful when you make significant changes to content on your site. During the process, the index continues to be complete and searchable, unlike when you choose to clear collections.
Viewing Sites

Click view sites on the Collection Status page to display all sites that are part of the current collection. The first page of the sites report displays the following information about each site in the collection:

- The indexing status of the site, such as idle or suspended
- The number of URLs in the site that are queued for indexing
- The number of URLs that Ultraseek has indexed
- The total number of documents indexed from the site
- The URL for the site

Click a site’s name on the report to view a listing of all known URLs for that site. Very large collections, those with more than 1000 sites, may take several minutes to load. This page displays the HTTP status code for each document that Ultraseek attempts to index at the site. The documents are sorted by revisit queue.

At the bottom of this page, Ultraseek displays the total number of URLs by the following criteria:

- Revisit queue
- HTTP status code
- Document disposition, such as the number of documents disallowed by a URL filter and duplicate documents
- The number of hops, or links, from the root URL
- Document character set

Site and URL listings are also available to non-administrative users through the View Sites link in the help interface. Also included is the total number of indexed documents for each site. See “Accessing Application Help” on page 32 for more information.
Viewing Log Files

Click view log on the Collection Status page to display recent collection activity. Sample log file messages are shown below.

Normal HTTP response:
Info: 200 OK: http://www.verity.com/home.html

Re-directs (page not indexed):
Info: 301 Moved Permanently: http://www.verity.com/gone.html

Duplicate (page unchanged, or duplicate of existing):
Info: 304 Duplicate of http://www.verity.com/other.html

Files Not Retrieved (if in DB, it will be removed):
Info: 400 Bad Request—URL
Info: 403 Not Authorized—URL
Info: 404 Not Found—URL

Server Errors (will retry at pre-determined interval):
Info: 503 Host Not Found

Sample log entries for database and Microsoft Exchange collections are shown below:

Microsoft Exchange
Sep 03 10:44:01 Info: [mse] on-disk C:\Program Files\Verity\SearchData\mse\db\000000
Sep 03 11:24:04 Info: [mse] Polling server:
mail.medicine.uiowa.edu

For a simple database:
Sep 03 10:44:01 Info: [dbase] on-disk C:\Program Files\Verity\SearchData\dbase\db\000000
Sep 03 13:24:01 Info: [dbase] Scanning database
Sep 03 13:24:03 Info: [dbase] Scan complete
**Merging Indexes**

The merge operation improves search performance by reducing the number of indexes that are searched. To merge indexes, click the **merge** button on the Collection Status page.

The time it takes to process a query grows linearly with the number of indexes, so the fewer the indexes, the better the access and return times. However, merging indexes does take some processor time. For example, Ultraseek can currently merge indexes at the rate of roughly 15,000 documents per minute. So very large collections will take a longer time to merge.

Ultraseek automatically merges indexes when the system detects that there are too many search indexes for the number of documents. This helps maintain a balance between the number of indexes and the amount of time it takes to merge them.

Merging also finalizes the deletion of documents. Documents can be deleted from the search indexes manually through **Collections | URLs | Delete URL**. As well as automatically merging indexes, the indexer also automatically deletes old and duplicate documents from the search indexes as it merges. Once deleted, a document won’t show up on a search results page. However, document counts shown at the top of the search results page will still include the document. Only after a merge are document counts updated.
Managing URLs

You can manage URLs to view the status of URLs, including information on the date a document was last indexed and the next scheduled indexed date, add URLs to the index, or delete URLs from the index.

- “Getting URL Status” on page 80
- “Adding URLs” on page 82
- “Deleting URLs” on page 82

You can also use the view sites feature available on the Collection Status tab to obtain a listing of all URLs in the URL database. See “Basic Collection Management” on page 74 for more information.

Features such as URL Status and Add URL are also available through the help interface. Users who do not have full administrative rights can perform minor maintenance functions using the More Services section of the help page. To restrict access and assign partial administrative rights to users, see “Managing Client Access” on page 170. For more information on the More Services section of the help interface, see “More Services” on page 33.

Getting URL Status

You can obtain the status of URLs in a collection, including the following information:

- Date an individual URL or document was last indexed
- Next scheduled revisit
- Last modification date
- Number of hops from the root URL

The URL Status feature automatically determines which collection a URL belongs to. You do not need to know which collection to query for a URL. When more than one collection contains the URL, Ultraceek presents the information from one of the collections and a list of hyperlinks to get the information from the other collections.

To obtain the status of a URL:

1. Click the Collections button.
2. Click the URLs tab.
3. Enter the URL for which you want a status.
4. Click *ok*.

A sample URL Status report is shown in “Adding URLs” on page 81.

**Figure 3-2** URL Status Report

```
more services

URL Status
http://news.ft.com/home/us

Collection
Financial Times

Scheduled to be revisited after Tuesday, 24-Jun-03 16:41:52.

Quality
-1 (15% relevance boost)

From the URL Database
Content-type: text/html
Size: 46820
Document hash: 779da71b25022b15
Referring URL: http://news.ft.com/home/us
Number of hops from root URL: 0
Last response code: 200
Last response: OK
Last modified: Monday, 23-Jun-03 16:42:46
Last put into index: Monday, 23-Jun-03 16:41:48
Last processed by indexer: Monday, 23-Jun-03 16:41:48
Last URL D8 update: Monday, 23-Jun-03 16:41:52
Queue: 1
To be revisited after: Tuesday, 24-Jun-03 16:41:52

Site Status
(0 queued + 4 completed urls) news.ft.com
```
Adding URLs

To easily add a new URL to the queue, or to tell Ultraseek to revisit it, use the Add URL feature. The Add URL feature is useful when only a few pages have changed and you want the index to immediately reflect the changes, instead of waiting for the spider to find the new or changed documents.

You do not need to specify a collection when adding a URL. Ultraseek will add the URL to all collections that allow it, and can also redirect the request to another instance of Ultraseek if cooperative spidering is configured.

To add a URL:

1. Click the Collections button and then click the URLs tab.
2. Click the Add URL tab
3. Enter the URL that you want to add to the collection and then click ok.

Deleting URLs

Documents that have been deleted from a web site are removed from the index the next time the spider checks them.

If the document exists, but you want to prevent the spider from indexing it, disallow it in the filter list, and then remove it. If you do not disallow the URL in the filter list, Ultraseek may find the URL again and index it. Alternatively, after you disallow a document, you can erase the database and re-index the entire site.

To remove documents from the index:

1. Click the Collections button.
2. Click the URLs tab.
3. Click the Delete URL tab.
4. Enter a query to find the documents that you want to delete.
   Ultraseek displays a page that lists all of the documents in the collection that match the results of your query.
5. Click the delete button next to the documents that you want to delete.
   Ultraseek displays a message asking you to confirm the deletion of the documents from the index. You can also choose to delete the URL from the URL database.
6. Click ok.
Filtering Duplicate URLs

The settings on the Collections | Dupes page allow you to indicate which version of a document to enter in the index when more than one copy exists. Ultraseek uses the URL with the highest score based on the wildcard patterns entered on this page.

Duplicate Determination

If Ultraseek finds two documents that have the same title and description, it treats one as a duplicate and does not put it in the index. You can choose, instead, to specify that two documents are duplicates only if the entire content is identical. You can also turn document deduping off.

Duplicate URL Preference

To allow Ultraseek to eliminate duplicate URLs, enter at least one score and wildcard URL pattern in the fields provided. If you index multiple sites and some sites are better choices for the index than others, indicate that preference with the Duplicate URL Preference.

When Ultraseek finds two URLs that refer to duplicate documents, it compares both URLs with each pattern. The first pattern a URL matches indicates its “score”. The URL with the higher score is the one kept in the search index. If both URLs have the same score, then the shorter URL is used.

It is useful to express a preference if you are indexing multiple sites and some sites are better choices to put in the index than others. For example, suppose you have an intranet with a local and a remote site. You could index both sites, but with a preference for URLs on the local site. For documents that are on both local and remote sites, the local URLs would appear on the Search Results pages. For example, say you have two URLs such as:

http://www.verity.com/default.html
http://www.verity.com/index.html

Both of these URLs point to the same documents, however, http://www.verity.com/index.html is the local site and http://www.verity.com/default.html is the remote site. You would give the higher score to http://www.verity.com/index.html in order to index those documents and keep them in the collection.
Duplicate URL Filter Specification

After Ultraseek determines that a document is a duplicate, it decides whether to follow links from the duplicate. Normally, it does not follow links from identical duplicates. It also does not follow links when the only difference in the URL is the hostname portion, which avoids wasting effort spidering mirror sites or hostname aliases.

Use the Duplicate URL Filter Specification to force the spider to allow or disallow links from certain duplicate documents. Enter the appropriate URL patterns and corresponding actions.

Improving Document Relevance Scores

This section describes how to improve the relevance scores of your search results by customizing the indexing process. Most documents contain superfluous content, such as copyrights and legal disclaimers. By default, the Ultraseek spider indexes everything it finds in a document, including low-value content, which can decrease the relevance score that Ultraseek assigns the document in response to user queries. The Page Expert, shown in Figure 3-3, allows you to increase the relevance of documents in spider collections by preventing the indexing of non-relevant content.

The quality factor of a document also affects its relevance score. The controls on the Collections | Indexing | Quality page let you adjust the quality factor of documents based on a variety of factors. Managing the quality factor assigned to documents during indexing is another way to improve the relevance score of documents.
Using Page Types

Use the Page Expert to assign a page type to a URL pattern, so that Ultraseek applies the type to every document in a collection with a URL matching that pattern. A page type consists of one or more page filters.

Each page filter defines the HTML tags corresponding to content that you do not want to index. For example, Ultraseek provides a default page type that contains the following built-in page filters:

- Choose largest blocks of text
- Skip blocks which are mostly links
- Skip copyright notices

You can create your own page types for other content that you do not want indexed into a collection, and you can create custom page filters for each type. If you do not want to use the default page type provided by Ultraseek, you must create your own before you can associate it with a URL pattern.
Creating Page Types

Use the following procedure to create a page type.

To create a page type:

1. Select a collection from the Collection list box.
2. Click the Collections | Indexing | Page Types tab.
3. Click new.
   
   The New Page Expert Type page is displayed, as shown in Figure 3-4.
4. Enter a name for your page type in the Name text box.
5. Optionally, enter a description in the Description text box.
6. Assign a quality factor to the new page type by selecting a value from the Quality Adjustment check box.
7. Check the checkbox next to every built-in page filter that you want to include with the page type.
8. Specify custom page filters to include with the page type.
   
   Check either the “index” or “don’t index” radio button, then specify the HTML or comment tag to include or exclude from the index.
9. Click ok.
Defining a page type is the first step to increasing the relevance scores for documents in a collection. Next, implement your page type by defining a URL pattern to which it applies.
Implementing Page Types

After you create a page type, use the Page Expert to implement it.

To implement a page type:

1. Select a collection from the Collection list box.
2. Click the Collections | Indexing | Page Expert tab.
3. Enter a URL pattern in the URL Pattern text box.
   You can use wildcards or a regular expression.
4. Optionally, use the Content contains text box to specify text that must appear in a document for the page type to be used.
   Leave the text box empty if you do not want to impose any content restrictions on the use of the page type.
5. Select a page type from the Page Type list box.
6. Click ok.

The next time that Ultraseek indexes content for the collection, it checks whether each document URL matches any of the URL patterns defined in the Page Expert. Ultraseek applies the page type associated with the first URL pattern that matches the document URL.

Testing Page Types

You can test your new page type before using it to index content. The Test Page Expert page allows you to see what content your page type includes and excludes for document. Use the following procedure to test a page type.

To test a page type:

1. Select a collection from the Collection list box.
2. Click the Collections | Indexing | Page Expert tab.
3. Click test.
   The Test Page Expert page displays.
4. Enter a URL that matches the URL Pattern for the page type you want to test.
5. Click ok.
   Ultraseek displays the test URL and highlights all content that will be indexed.
Adjusting Quality Factors

Settings on the Collections[Indexing][Quality] page allow you to adjust the quality factor for documents based on URL pattern matches, meta tags, and other criteria. The quality factor is added to scores computed in search results. Quality factor scores range from a minimum value of -16 to a maximum value of +15. To increase a URL pattern’s importance, give it a positive quality score. To decrease its importance, give it a negative quality factor. The indexer assigns a quality factor to each document it indexes based on this baseline quality value, with some automatic adjustments based on built-in heuristics.

Figure 3-5 Quality Page
To adjust a document’s quality factor based on its URL:

1. Select a collection from the Collection list box.
2. Click the **Collections > Indexing > Quality** tab.
3. Enter a URL pattern in the URL Pattern text box.
   You can use wildcards or a regular expression.
4. If you used a regular expression to specify the URL pattern, select regex from the Syntax list box.
   The default value is wildcard.
5. Select a quality factor from the adjustment list box.
6. Click **ok**.

To adjust a document’s quality factor based on its meta tags:

1. Click the **Collections | Indexing | Quality** tab.
2. Enter the name of a meta tag in the Name text box.
3. Optionally, enter a value for the meta tag in the value text box.
4. Select a quality adjustment from the Adjustment list box.
5. Click **ok**.

The following table describes the other controls on the **Quality** page that allow you to adjust a document’s quality score.

<table>
<thead>
<tr>
<th><strong>Quality Adjustment</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Quality Value</td>
<td>The initial quality factor assigned to documents in this collection. All other quality factors specified on this page are added or subtracted from this starting value. The default value is 0.</td>
</tr>
<tr>
<td>Penalize for direct depth of document</td>
<td>Decreases the quality factor of a document based on how many directories from root it is located. The further a document is removed from root, the higher the penalty.</td>
</tr>
<tr>
<td>Boost for inbound links to document</td>
<td>Increases the quality factor of a document depending on the number of documents from other sites with links to it. The more inbound links, the higher the boost to the quality factor.</td>
</tr>
</tbody>
</table>
### Quality Adjustment

<table>
<thead>
<tr>
<th>Quality Adjustment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document URL ends in a slash</td>
<td>Adjusts the quality of documents that appear to be home pages. Since home pages typically contain high-value content, the default value is 2.0.</td>
</tr>
<tr>
<td>Document URL contains a query string</td>
<td>Adjusts the quality of documents that appear to be dynamically generated. URLs for these pages contain a question mark and a list of form parameters. Since the content in dynamically generated pages is usually less relevant than other documents, the default value is -1.0.</td>
</tr>
<tr>
<td>Document is from a /~user directory</td>
<td>Adjusts the quality of documents in user directories. Since content in user directories is typically unmanaged, the default value is -2.0.</td>
</tr>
</tbody>
</table>
Setting Network Parameters

To enable spider access throughout complex networks, including those with many security features, specify the Collections | Network settings discussed in the following sections.

**Figure 3-6** Spider Collection Network Page (Top Half)
User Agent Specification

Some sites allow only specific types of users, such as spiders, to access the server. The User Agent name is what the system presents by HTTP when it requests pages for indexing.

One way you can use the user agent is to allow certain sites to be indexed only by your search software, but not by other spiders. By giving your collection a unique user agent name, and allowing that user agent to index the site, only spiders presenting that user agent are enabled. You can have different user agent names for different URL patterns.

When the spider retrieves a document, it compares the URL of the document with each user agent URL pattern. The first pattern the URL matches indicates which user agent value to use. The spider presents this value to the HTTP server in the “User-Agent” HTTP header line. If no match is found, no “User-Agent” HTTP header line is presented.

“Ultraseek” is the default user agent that Ultraseek uses while spidering.

---

**Note** Some sites will only work with specific browser versions and will return a “Upgrade Your Browser” page or message. In order to index these sites, set the “User-Agent” to version of web browser accepted by the site.

Since web servers record the “User-Agent” in their access logs, and this information is used to generate reports about the source traffic, if you are running more than one spider you can use different “User-Agent” strings to distinguish between the different agents.

Authentication

The Authentication pane enables you to index content at password-protected sites.

Click **new** to enter the username and password for any password-protected sites you want to index. Users can view search results from these password-protected sites, but are required to enter their passwords to access the document. Password-protected sites without an appropriate entry in this field are not indexed and are noted in the log.

When the spider retrieves a document, it compares the URL of the document with each password-protected URL pattern. The first pattern the URL matches indicates which username and password values to use.

When it fetches the document, the spider presents these values to the HTTP server in the “Authorization” MIME header line. If no match is found, no “Authorization” MIME header line is presented.
Form-Based Authentication

Web sites occasionally require a session cookie before they can be viewed in a browser or spidered by a crawler. This type of authorization is usually acquired by receiving a cookie from a form on the particular site. Usually this is done by entering a username and password for that site. Once the information is entered into the form, the end-user is given a cookie and redirected back to the original content. Ultraseek allows the crawler to simulate this behavior, enabling the indexing of content behind a site with form-based authentication.

Note You will only be allowed to spider a site if you have access rights to that site. You will not be allowed to access a secure site without the proper credentials.

To create a form-based authentication entry:

1. Click the Collections tab
2. Click the Network tab.
3. On the Network page, scroll down to the Form-Based Authentication region.
4. Click new. The New Form-Based Authentication Specification page displays.

Figure 3-7 Form-Based Authentication Specification

5. Enter the URL pattern to those documents that are protected with the Form-Based Authentication.

6. Select either wildcard or regex.

7. Enter the URL of the authorization form.
8. Click **ok**.

   After clicking **ok**, you will be directed to a second page. This page will normally contain only two fields, username and password. However, the number of fields will be determined by the web site’s authentication fields.

9. Enter the information for the site and click **ok**.

---

**Note**  
If the authentication form cannot be retrieved, an error page is returned displaying an error message as to what went wrong. This page contains the form(s) that were parsed from the page located at the Form URL (Step 6).
Additional HTTP Header Entries

Identify additional HTTP header information with the HTTP Header Entries.

When the spider fetches a document, it compares the URL of the document with each additional HTTP header entry’s URL pattern. For each pattern the URL matches, the spider presents the corresponding HTTP header entry name and value to the remote HTTP server when downloading documents.

A typical example of this is when providing a cookie. This setting is implemented primarily by advanced users, however, it is rarely needed.

**Figure 3-8** Collection Network Page (Bottom Half)
Spider Throttle

On rare occasions, you might need to adjust your spider throttle settings. For example, a public site administrator might ask you to increase the time your spider waits before revisiting that site. Or, you might routinely index a large number of documents from a single host and want to use multiple spider threads to index the site more quickly. The spider throttle pane allows you to specify the maximum number of spider threads that may simultaneously index a single web site, and specify the amount of time those threads wait between HTTP requests to the same server.

Ultraseek uses site patterns to associate throttle settings with a web site. After the spider retrieves a document, it compares the site where the document was found with the site patterns specified in the spider throttle pane. The spider waits the number of seconds specified by the first matching site pattern. If no match is found, the spider does not wait.

To increase time between the spider’s HTTP requests to a web site:

1. Click the Collections > Network tab.
2. Select a collection from the Collection list box.
3. Enter a site pattern that corresponds to the site URL.
   Optionally, choose “regex” from the Syntax list box if you prefer to use a regular expression to identify the site.
4. Enter the number of seconds that you want the spider to wait in the Seconds text box and then click ok.

Delay values greater than a few seconds may make it difficult for the spider to finish spidering the site. For example, a 60 second delay for a 1000-document site would take over 16 hours to complete. It is recommended that you throttle the spider only when necessary and keep the default delay value.

To change the maximum number of spider threads per site:

1. Select the Collections | Network tab.
2. Select a collection from the Collection list box.
3. Enter the maximum number of threads in the “Maximum number of spider threads per site” text box and then click ok.

   You can assign as many spider threads as are available to the collection. This value is set on the Collections | Tuning page.

Note  Verify that your HTTP server is configured to handle multiple threads before increasing the number of spider threads.
Proxy Server Specification

You can associate different URL patterns with different proxy servers, allowing the spider to index content on both sides of firewalls. Click **new** to create add proxy server information. You can designate a site pattern, a proxy server, and username and password information if applicable.

When the spider retrieves a document, it compares the site name in the URL of the document with each proxy-server site pattern. The first pattern the site name matches indicates which proxy server to use.

When the spider retrieves a document using the proxy server, if no match is found, no proxy server is used. You can specify a port in the proxy server value with a colon. For example, if the host “venus” has a proxy server running on port 8080, use **venus:8080**.

---

**Note** In mirror collections, you designate proxy server information using the **Collections | Status** page.

Direct File Access

The files you seek are often in a system that is accessible to the server where Ultraseek resides. You can point Ultraseek to these files by specifying direct file access information.

If you enter the URL prefix and its corresponding pathname prefix, Ultraseek opens the file rather than the URL.

Direct file access is useful to eliminate spider traffic to a local HTTP server. When the spider retrieves a document, it compares the beginning of the URL with each direct file access URL prefix. If a prefix is the same as the beginning of the URL, the spider fetches the document by directly accessing a local file. The local file pathname is constructed by replacing the URL prefix with the direct file access pathname prefix.
Automatic URL Adder (Cooperative Spidering)

The Automatic URL Adder section allows you to enable cooperation between multiple spidering instances of Ultraseek. Click **New** to create a new cooperative spidering instance. You can designate a URL pattern, an Ultraseek destination site, proxy information, and proxy authentication information.

You can specify that disallowed URLs be automatically fed to other instances of Ultraseek. Using this feature, spidering instances of Ultraseek can cooperate by configuring them to feed hyperlink URLs to each other.

This is useful if you want to build collections of cross-referenced documents. For each disallowed hyperlink URL the spider finds, an add-URL request will automatically be generated for the corresponding destination Ultraseek site. An optional proxy server can be specified for the add-URL request to go through. Both the destination Ultraseek site and proxy server parameters are specified as in the site part of a URL - a hostname and optional colon and port number.

Enter a URL pattern and corresponding remote Ultraseek site to automatically feed disallowed URLs to another running instance of Ultraseek.

Tuning Collections

To tune collections, choose the **Tuning** tab for a selected collection name under the **Collections** menu. Depending on the type of collection that you are tuning, you will see different regions and parameters on the tuning screens. Some options are not available for all collection types; for example, you cannot set indexer weights or word spam detection thresholds for mirrored collections.

<table>
<thead>
<tr>
<th>Tuning Parameters</th>
<th>“Setting Tuning Parameters” on page 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activities</td>
<td>“Scheduling Activities” on page 106</td>
</tr>
<tr>
<td>Activity Curfews</td>
<td>“Setting Activity Curfews” on page 107</td>
</tr>
<tr>
<td>Indexer Weights</td>
<td>“Setting Indexer Weights” on page 108</td>
</tr>
<tr>
<td>HTML META Tag Names</td>
<td>“Specifying HTML META Tag Names” on page 109</td>
</tr>
<tr>
<td>Document Titles</td>
<td>“Replacing Document Titles Automatically” on page 113</td>
</tr>
</tbody>
</table>
3 Managing Collections

Tuning Collections

Setting Tuning Parameters

Tuning parameters enable you to optimize the performance of your collections and search site. Depending on the type of collection that you are working with, you will see different parameters on this page. This section first describes the common tuning parameters for most collections. Additional parameters for specific collection types (such as additional parameters for mirrored collections) are described at the bottom of the tuning parameters reference table.

Some of the tuning parameter names vary slightly, depending on collection type. For example, the Usenet Tuning Parameters card has a field to set the number of indexer threads, instead of spider threads for spider collections.

To set tuning parameters:

1. Click Collections.
2. Select the Tuning tab.
3. Choose the collection that you want to modify.

The Tuning Parameters page for the collection type displays. Figure 3-9 shows an example Tuning page; each parameter is described in detail in the table that follows.
### Spider Tuning Parameters

- **Number of spider threads**
- **HTTP connection timeout (seconds)**
- **Maximum parse cpu time (seconds)**
- **Number of revisit queues**
- **Minimum document revisit interval (days)**
- **Maximum document revisit interval (days)**
- **Default document age estimate (days)**
- **Automatic-merge minimum on-disk index ratio**
- **Automatic-merge minimum on-disk index size (documents)**
- **In-memory index size (documents)**
- **Maximum document download (megabytes)**
- **Data directory**
- **Primary Language**
- **Default Encoding**
  - Use HTTP keep-alives
  - Use primary host names
  - Discover sites
  - Follow redirects immediately
  - Accept cookies
  - Ignore case differences in URLs
  - Remove session IDs from URLs
  - Show this collection by default
  - Search this collection by default
  - Use sitelist.txt files

### Scheduled Activity

Click to schedule a new task to be started at regular intervals.
### 3 Managing Collections

**Tuning Collections**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of spider threads</strong></td>
<td>Specifies how many threads to reserve for indexing the collection. The default is 5. If you have a large number of collections, you can reduce this value so that the total number of threads is not too large. Each indexer thread requires one megabyte of memory for its stack. Generally, query performance suffers if you devote more than 15 threads for indexing all of your collections. Regardless of how many threads have been specified here, <em>only one thread at a time indexes each site</em>. To be network-friendly, Ultraseek allows only one outstanding request to a site at any time.</td>
</tr>
<tr>
<td><strong>HTTP connection timeout</strong></td>
<td>Specifies the maximum number of seconds to wait for a response from a remote HTTP server before giving up. The default is 60 seconds.</td>
</tr>
<tr>
<td><strong>(seconds)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Maximum parse cpu time</strong></td>
<td>Specifies the maximum number of CPU seconds to allow for the parsing of a document.</td>
</tr>
<tr>
<td><strong>(seconds)</strong></td>
<td>Most HTML files parse in well under one second, but some large files in other document formats can take very long to parse. The default of 60 seconds should cover most circumstances.</td>
</tr>
<tr>
<td><strong>Number of revisit queues</strong></td>
<td>Specifies the number of different revisit queues to set up between the minimum and maximum intervals.</td>
</tr>
<tr>
<td></td>
<td>Each queue has its own revisit interval, which is logarithmically interpolated between minimum and maximum, depending on how many queues you define.</td>
</tr>
<tr>
<td></td>
<td>The indexer automatically puts every URL it visits in one of the queues, depending on how often it seems to change.</td>
</tr>
<tr>
<td></td>
<td>The recommended value is 6. With a minimum revisit interval of one day, a maximum interval of 32 days, and 6 queues, the queues have revisit intervals of 1, 2, 4, 8, 16, and 32 days.</td>
</tr>
<tr>
<td><strong>Minimum document revisit</strong></td>
<td>Specifies the minimum time the indexer should wait before revisiting a URL that changes frequently. The recommended value is one day. If you are indexing sites that are rapidly changing, specify a shorter period.</td>
</tr>
<tr>
<td><strong>interval (days)</strong></td>
<td>Decimals are supported; specifying 0.5 days is the equivalent of specifying 12 hours. Choose a frequency that allows the entire collection to be indexed without significant revisit traffic interfering with the completion of the queue.</td>
</tr>
<tr>
<td><strong>Maximum document revisit</strong></td>
<td>Specifies the maximum time the indexer should wait before revisiting a URL that does not appear to change.</td>
</tr>
<tr>
<td><strong>interval (days)</strong></td>
<td>The default is 32 days. If you are indexing sites that do not change very often, you may want to specify a longer period. Decimals are also supported.</td>
</tr>
</tbody>
</table>
### Default document age estimate (days)

The age of a document is used to direct its URL to a revisit queue. Normally, the indexer uses the “Last-Modified” HTTP header line to determine the age of a document, but many times this information is not present.

If the indexer cannot determine the age of a document, it uses the default age estimate. The recommended value is 16 days which is the default. The estimate is only used the first time the indexer visits a document. On subsequent visits, the age is determined from whether the document has changed. Decimals are also supported.

### Automatic-merge minimum on-disk index ratio

This setting specifies when Ultraseek should automatically merge on-disk indexes. After an in-memory index is saved to disk, the size of each on-disk index is compared with the sum of the sizes of all smaller on-disk indexes. If the ratio of the size to the sum is smaller than the minimum, then the index and all the smaller indexes are merged.

After each in-memory index is saved to disk, Ultraseek automatically merges indexes when the number of search indexes created exceeds the number of documents. This helps balance load on your server by reducing the number of indexes so that Ultraseek does not spend a lot of time merging indexes.

### Automatic-merge minimum on-disk index size (documents)

Specifies when Ultraseek should automatically merge on-disk indexes. After an in-memory index is saved to disk, the smallest on-disk indexes are merged so that none is smaller than the minimum.

### In-memory index size (documents)

The in-memory index size is an estimate of the maximum size for the in-memory index. When the indexer finds a new document, it puts it into an in-memory index. When the in-memory index gets full, the indexer saves it, resulting in an on-disk index, which eventually is merged with other on-disk indexes.

The in-memory index should not be too large, or the indexer uses too much memory. It should not be too small, or the indexer too frequently performs time-consuming saves and merges.

When you set the in-memory index size, consider the overall RAM resources of your server. The default designated for Ultraseek is derived from the amount of memory in your system.

### Maximum document download (megabytes)

Specifies the maximum size of download allowed per document, in megabytes, to keep Ultraseek from downloading excessively large documents. The default value is 15.0 MB. When Ultraseek downloads a document, it will stop receiving data after this limit has been reached, and will proceed to parse and index just what it has downloaded.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data directory</td>
<td>Specifies the pathname of the directory where Ultraseek should keep all its data for this particular collection. If you specify a relative path name, the path is interpreted as relative to Ultraseek’s data directory. There is usually no need to change this unless you want to put collection data on a different disk volume. It is not recommended to place data directories on different machines across a network. For example, file servers. This will slow down Ultraseek since this data will also be accessed for searches. Verity does not support this type of installation, and should not be performed.</td>
</tr>
<tr>
<td>Default Encoding</td>
<td>Select the character set encoding to use to decode a document if encoding has not been specified for the document, and the encoding identifier is unable to identify the character set encoding.</td>
</tr>
<tr>
<td>Primary Language</td>
<td>Determines which stemmer Ultraseek should use when inserting a document into the collection. For example, if you cannot determine the language of the document from the HTTP header, META tags, or language identifier. If this is a mirrored collection, this specification should match the primary language setting for the collection on other systems. To have access to stemmers for languages other than English, you must install the appropriate language support software for your platform. Please contact a Sales Representative for further information. For information on filtering documents by language, refer to “Setting Allowable Languages” on page 67.</td>
</tr>
<tr>
<td>Use HTTP keep-alives</td>
<td>Specifies that Ultraseek should use HTTP keep-alives when indexing. Using HTTP keep-alives results in more efficient indexing, since fewer HTTP connections need to be opened. However, the indexer sometimes has problems talking to some HTTP servers with keep-alives enabled.</td>
</tr>
<tr>
<td>Use primary host names</td>
<td>Specifies that the indexer should automatically identify all hostnames in URLs by their primary name. To find a host’s primary name, the indexer first looks up the name it has to get an address. It then looks up the address to get a name. That name is considered the primary name. Some sites are known by multiple names, and can therefore end up being indexed more than once. Ultraseek eliminates duplicates automatically, and by using a primary hostname you can reduce the overhead of redundant indexing.</td>
</tr>
<tr>
<td>Discover sites</td>
<td>Specifies that whenever the indexer finds a link to anywhere on a newly-discovered site, it also starts indexing at the root URL of the new site.</td>
</tr>
</tbody>
</table>

---
## Tuning Collections

### Parameter Description

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Follow redirects immediately</strong></td>
<td>Specifies that when Ultraseek receives a 302 code from a web server, it should immediately start indexing from the new target location. This can be useful when indexing web servers that use single sign-on methods for authorization.</td>
</tr>
</tbody>
</table>
| **Accept cookies**                 | Specifies that Ultraseek should:  
- Accept HTTP session cookies from web sites it spiders.  
- Return the cookies in subsequent HTTP requests. |
| **Ignore case differences in URLs** | Specifies whether the spider should ignore differences in case in URLs. For example, if Ultraseek encounters two URLs that are identical, except for differences in case, the spider treats both URLs as the same URL. |
| **Remove session IDs from URLs**   | Controls whether the spider removes commonly used session IDs from newly discovered URLs before adding them to the URL database.            |
| **Show This Collection by Default** | When checked, this parameter specifies that this collection should be shown when collections are listed on query forms. If this parameter is not checked, then this collection will not be shown. This option can be overridden on a per-query basis through the use of the `qc` form variable. See the *Verity Ultraseek Customization Guide* for more information. |
| **Search This Collection by Default** | When checked, this parameter specifies that this collection should be searched by default. If the collection is listed on the query form, this option controls the default value of the checkbox next to the listing. If it is not listed, then the collection will be unconditionally searched (the user does not have the option of checking or un-checking the checkbox). This option can be overridden on a per-query basis through the use of the `col` form variable. See the *Verity Ultraseek Customization Guide* for more information.  
Tip: If you have several collections and you want them all to be unconditionally searched, then uncheck all the “show this collection by default” checkboxes, but leave checked all the “search this collection by default”. The user will see no collection checkboxes, and will always search all collections. |
| **Use sitelist.txt files**         | Specifies that the indexer should look for *sitelist.txt* files. These files allow the indexer to rapidly learn what documents at a site have changed.  
This function is off by default because this is an earlier implementation. The revisit queues feature should be used instead. |

### Additional Parameters for Other Collection Types

---

Verity Ultraseek Administrator Guide 105
Scheduling Activities

You can schedule spidering activities to specify that certain tasks start at regular intervals. A few examples of how you can use scheduled activities are described below:

- For spidered collections, you can specify when to revisit URLs and re-index changed documents.
- For mirrored collections, you can specify when to poll the remote server for updates.
- For database collections, you can specify when to merge on-disk indexes.

To schedule an activity:

1. Click Collections.
2. Select the Tuning tab.
3. Choose the collection that you want to modify.
4. On the Tuning page, scroll down to the **Scheduled Activity** region.
5. Click **new**. The New Scheduled Activity Specification page displays.
6. Specify the frequency and start time for the activity.
7. Select the type of action that you want to schedule:

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rescan</td>
<td>Revisit URLs and reindex changed documents.</td>
</tr>
<tr>
<td>reindex</td>
<td>Revisit URLs and reindex all documents.</td>
</tr>
<tr>
<td>save</td>
<td>Save in-memory indexes.</td>
</tr>
<tr>
<td>merge</td>
<td>Merge on-disk indexes</td>
</tr>
<tr>
<td>clear</td>
<td>Clear and rebuild the collection</td>
</tr>
<tr>
<td>suspend</td>
<td>Suspend the collection.</td>
</tr>
<tr>
<td>resume</td>
<td>Resume the collection.</td>
</tr>
</tbody>
</table>

8. Click **ok**.

### Setting Activity Curfews

You can set activity curfews that determine when collections are allowed to be updated. For example, you may not want to spider your web site during peak site visiting hours, and can instruct the spider to run only at the time that you specify.

**To set an activity curfew:**

1. Click **Collections** and then select the **Tuning** tab.
2. Choose the collection that you want to modify.
3. On the Tuning page, scroll down to the Activity Curfew region.
   
   **Note** If there are existing curfew entries, select the **Allow operation at all times** option to override the listed schedule entries and allow updates at any time. When a curfew is in effect, the collection status (shown on the **Status** tab) is **curfew in effect**.

4. Click **new**.
   
   The New Allowed Period of Operation page displays.
5. Set the frequency, start, and run time options for the activity.
Setting Indexer Weights

Indexer weights define the importance of various portions of a document’s text, relative to the body text of the document. The greater a weight for a section, the more a word in that section counts against a query. In a section with a weight of zero, a word is not counted. Do not set the weights so that the combined value is more than 32. If this is done, Ultraseek will set the weights back to 0.

To set indexer weights:

1. Click Collections.
2. Select the Tuning tab.
3. Choose the collection that you want to modify.
4. On the Tuning page, scroll down to the Indexer Weights region.
5. Enter values for each indexer weight field, following the guidelines below.

Many of the indexer weights are based on various forms of META tags. Acceptable characters in META tags are letters, numbers, dots, dashes, and underscores.

<table>
<thead>
<tr>
<th>Title</th>
<th>You can provide document titles with additional weight relative to the body text. The title is extracted from the HTML Title tag. The Title field has a value from 0 to 10; the default setting is 8.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>If a document has an HTML META tag with the name designation “description”, or if there is an override setting designated, that content displays in the summaries on the Search Results page. This field has a value from 0 to 10; the default setting is 4.</td>
</tr>
<tr>
<td>Keyword</td>
<td>If a document has an HTML META tag with the name designation “keywords”, Ultraseek weighs those keywords into the document’s relevance ranking. This field has a value from 0 to 10; the default setting is 4.</td>
</tr>
</tbody>
</table>
The default values are designed for sites that have fairly consistent use of META tags. See “Searching by Metadata” on page 240 for more information.

### Specifying HTML META Tag Names

You can specify which META tags to use as overrides for title, summary, date, keywords, URL, and publisher information. Some examples of use of the HTML META tag names feature include using "dc.title" for the title override META tag, or removing the "description" META tag name to stop the description META tag from overriding the automatically generated document description.

You can specify more than one META tag name by entering names as a comma-separated list. If you specify a META tag name in lowercase, it will match any capitalization of META tag name in a document, but if you have any capitalized letters in the META tag name, it will match exact case only.

**To specify META tag names:**

1. Click **Collections**.
2. Click the **Tuning** tab.
3. Select the collection that you want to modify.
4. On the Tuning pane, scroll down to the HTML Meta Tag Names region.
5. Enter the values for applicable override fields, using the guidelines in the table below.
### HTML Meta Tag Names

<table>
<thead>
<tr>
<th>Meta Tag Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Used when displaying a document’s title in the search results.</td>
</tr>
</tbody>
</table>
| Description   | Used when displaying a document’s summary in the search results. By default, the summary is taken from the “description” META tag. If there is no “description” META tag available, Ultraseek will create a smart summary. A smart summary is created by grabbing the first bit of body text found in a document, ignoring text within heading tags `<h1>`.
| Date          | Used when displaying the document date. Ultraseek can parse the “Full Date” format defined in the HTTP 1.1 protocol specification or web profile of ISO 8601. The following are examples of date stamps which can be parsed:
|               | 1994-11-06T08:49:37Z (ISO 8601 format) |
|               | Sun, 06 Nov 1994 08:49:37 GMT |
|               | Sunday, 06-Nov-94 08:49:37 GMT |
|               | For more information on date stamping and the currently accepted protocol practices, see: |
|               | http://www.w3.org/Protocols/rfc2068/rfc2068 |
|               | The web profile of ISO8601, common for the `dc.date` META tag, is available at: |
The default values are designed for sites that have fairly consistent use of META tags. See “Searching by Metadata” on page 240 for more information.

### Getting External Metadata from a Database

A spider collection can be configured to automatically query a database for additional meta data for a document. To use this feature, you must have at least one database record allowed by your Ultraseek license.

The following databases are supported:

- Oracle 8i, 9i, 10g (with or without client)
- DB2 Universal Server (UDB) 6.1, v7.x and v8.x
- Sybase Adaptive Server (ASE) 11.5, 11.9, 12.0, 12.5, 12.5.1
- Microsoft SQL Server 7, SQL Server 2000

**Note** For Windows support, you must obtain the ODBC drivers from the database vendor.

### Configuring External Metadata from a Database

To set up retrieving external metadata:

**Note** For a given collection, you can configure only one database from which to retrieve metadata.

1. On Collections > Tuning, in the External Metadata from a Database section, click new.

<table>
<thead>
<tr>
<th>Override</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keywords</strong></td>
<td>Used in calculating relevance.</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td>Specify a URL override to modify the URL displayed for the document in a search results page. You can use the URL override to index one document, but when a user clicks the search results it leads to another document. This would be useful if you wanted to provide search of an image, audio, or video document database. This is providing that the HTML documents describe each of the media types.</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>Used to display the document’s publisher.</td>
</tr>
</tbody>
</table>
2. Select a database type and specify the appropriate information for the other fields. What you see depends on your operating system.

3. As prompted, provide appropriate credentials.

4. Select the table that contains the extra metadata, specify a table owner if there is one, and then click ok.

5. Select a column name and then the item to which you want to match the column’s data.
   For example, if you have a column with the whole URL in it, select URL and the whole URL.

6. Optionally add the columns to field and/or normal queries, and then click ok.

7. Optionally map Ultraseek field names to table column names, and then click ok.

8. Optionally click ok to revisit all URLs in the collection (recommended). Otherwise, click no revisit.

**Editing the Configuration**

Click edit next to the section that applies. Only the relevant section appears, as it did when you set up the configuration.

**Deleting the Configuration**

To delete the authentication information, click delete that appears on the same row as the Authentication section title.

To delete the table information, click delete that appears on the same row as the Table section title.

To delete the entire configuration, click delete that appears on the same row as the External Metadata from a Database section title.
Replacing Document Titles Automatically

Many documents have junk or missing titles, which degrades the ability of your users to search for documents by title. Ultraseek can automatically replace these missing and junk titles for you.

**To automatically replace document titles:**

1. Click the Collections button.
2. Choose a collection from the Collection list box.
3. Click the Tuning tab.
4. Scroll down to the Document Title Replacement frame on the Tuning page.
5. Check the Enable replacement of missing and junk titles checkbox.
6. (optional) If you do not want filenames used as titles, check the Titles which appear to be filenames are junk checkbox.
7. (optional) Enter any titles that you want the indexer to consider as junk in the text fields below the checkboxes, as shown in Figure 3-11.
8. Click OK.

**Figure 3-11** Document Title Replacement

Several junk titles, such as Untitled Document, are specified by default.
Setting Word Spam Detection Thresholds

Ultraseek includes an automatic spam detection algorithm. The indexer decides that a document is spamming a certain word if that word occurs more often than a certain number of times within a range of 100 words. If it decides that a document is spamming the word, it punishes the document by giving it a very low relevance for the word.

To set word spam detection thresholds:

1. Click Collections.
2. Select the Tuning tab.
3. Choose the collection that you want to modify.
4. On the Tuning page, scroll down to the Word Spam Detection region.
5. Specify the threshold of how many word occurrences there needs to be for Ultraseek to decide that a document is spamming a word. For each threshold, the minimum value is 1, the maximum is 100. To disable spam detection, enter a value of 100 for each item. You can set separate word occurrence values for several sections in HTML documents.

<table>
<thead>
<tr>
<th>Section</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>4</td>
</tr>
<tr>
<td>Description (summary)</td>
<td>6</td>
</tr>
<tr>
<td>Keywords</td>
<td>6</td>
</tr>
<tr>
<td>Image alt text</td>
<td>2</td>
</tr>
<tr>
<td>Body</td>
<td>10</td>
</tr>
</tbody>
</table>

Customizing Title Records

Title records contain the information that is displayed on search results pages. You can specify how document metadata is assembled into title records as documents get indexed.

To enter title records:

1. Click Collections.
2. Select the Tuning tab.
3. Choose the collection that you want to modify.
4. On the Tuning page, scroll down to the Title Records region.
5. Enable passage-based summaries if you want to include summaries that are created from matching passages in the document text. This text is saved in the description component of the title record.

6. Enter the Title Record Size, which is the amount of disk space to reserve for each title record.

7. The size can be any value from 512 bytes to 8192 bytes. Choose the smallest title record size that can hold the information you need. Ultraseek performs best when the record’s size is an integer multiple of 512 bytes.

8. Enter the Maximum Component Lengths that should be used to specify the number of characters that can be used for a document’s title, description, URL, or publisher.

9. No component may contain more characters than the title record size minus 129 bytes. For all but the URL component, Ultraseek automatically truncates the component’s text, if necessary. For the URL component, Ultraseek disallows longer URLs.

10. Enter additional component names and field lengths in the available fields, if needed.

Checking the Internal and Display Names

The display name of your collection is used when Ultraseek needs to provide a name on an HTML page. For example, the display name is the name of your collection to the right of a collection checkbox on your search page. It can contain up to 32 characters.

The internal name, designated when you first set up a collection, is shown above the display name. The internal name is used when Ultraseek needs to refer to a collection, such as when setting a value for a form variable.

Note Once the internal name has been set, you cannot change it. However, you can change the external display name at any time.
Customizing Ultraseek Search Forms

This chapter explains how to use the Administrative Interface and Ultraseek styles to customize your search interface and search results forms. An Ultraseek style consists of a set of server configuration parameters and a cascading style sheet (.css).

This chapter contains the following sections:

- Customizing Styles
- Editing Style Sheets
- Setting Server Parameters
Customizing Styles

Ultraseek provides four styles as starting points for customizations:

- Simple
- Standard (default)
- Ruby
- Metal

You use the Style Editor, shown in Figure 4-1, to edit Ultraseek styles and the Interface | Style and Interface | Query pages to edit the server parameters associated with a particular style. You can also create new styles. Once you designate a default style, Ultraseek applies it to all your search interface and search results forms.

You can still customize your forms using form variables, as described in the Verity Ultraseek Customization Guide. However, these customizations cannot use features new to Ultraseek versions 5.1 and above, such as Group by Location and In-document Highlighting. In addition, you will likely need to re-implement your form variable customizations each time you upgrade Ultraseek.

IMPORTANT The Ultraseek server process must have write permission for the install_dir/docs directory. The location of this directory can be viewed and changed in the Document directory text field on the Server | Parameters | Main page in the Ultraseek administrative interface.
Editing Style Sheets

Use the Style Editor to graphically edit the style sheets associated with your Ultraseek styles. The control pane on the left lets you select the field that you want to edit, and the preview pane on the right automatically shows your changes. You can also submit queries in the preview pane, but you cannot browse the returned search hits. Click the reset button to clear all unsaved changes.

**Note**  
Your web browser must support frames to use the Style Editor.

**Figure 4-1** Style Editor

To edit a style sheet:

1. Click **Interface** and then select the Ultraseek style that you want to edit in the Style list box on the Style pane.

2. Click **launch style editor**.

3. Click the property link in the control pane for the style sheet property that you want to edit. describes each property. **Figure 4-2** and **Figure 4-3** show where many of the properties appear in the search pages.

4. Use the displayed controls to edit the style sheet and then click **save**.
Table 4-1  Style Sheet Properties

<table>
<thead>
<tr>
<th>Style Sheet Property Link</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML›Link</td>
<td>Specifies how HTML links are displayed.</td>
</tr>
<tr>
<td>HTML›Active link</td>
<td>Specifies how active HTML links are displayed.</td>
</tr>
<tr>
<td>HTML›Visited link</td>
<td>Specifies how visited HTML links are displayed.</td>
</tr>
<tr>
<td>HTML›Entire page</td>
<td>Specifies how the entire HTML page is displayed.</td>
</tr>
<tr>
<td>Search box›Collection Names</td>
<td>Specifies whether to display or hide the names of collections managed by Ultraseek.</td>
</tr>
<tr>
<td>Search box›Various labels</td>
<td>Specifies the size and look of labels in the search box, such as Tip: and Example:.</td>
</tr>
<tr>
<td>Search box›Languages</td>
<td>Specifies the language to use for prompts. Ultraseek stems query terms in the selected language.</td>
</tr>
<tr>
<td>Search box›Search input box</td>
<td>Specifies whether to display the search input box.</td>
</tr>
<tr>
<td>Search box›Search button</td>
<td>Specifies whether to display the search button in the search input box.</td>
</tr>
<tr>
<td>Search box›Tips</td>
<td>Specifies whether to display tips to the right of the search box.</td>
</tr>
<tr>
<td>Search box›Thesaurus</td>
<td>Specifies whether to display the Thesaurus.</td>
</tr>
<tr>
<td>Search box›Custom logo</td>
<td>Specifies whether and how to display a custom logo above the search box.</td>
</tr>
<tr>
<td>Results›Score bar</td>
<td>Specifies whether to display a score bar on the search results page. A score bar is two-color horizontal bar that graphically illustrates the relevance score of a query.</td>
</tr>
<tr>
<td>Results›Collection in hit</td>
<td>Specifies whether and how to display the names of collections containing the search hits displayed on the search results page.</td>
</tr>
<tr>
<td>Results›Date in hit</td>
<td>Specifies whether and how to display a document’s date on the search results page.</td>
</tr>
<tr>
<td>Results›Description in hit</td>
<td>Specifies whether and how to display a description of a document’s contents on the search results page.</td>
</tr>
<tr>
<td>Results›Extra info in hit</td>
<td>Specifies whether and how to display extra information about a hit on the search results page.</td>
</tr>
<tr>
<td>Results›Find similar link</td>
<td>Specifies whether and how to display a “Find Similar” link to the right of each hit on the search results page. Clicking this link causes Ultraseek to return documents similar to the search hit.</td>
</tr>
<tr>
<td>Results›Highlighted term</td>
<td>Specifies how to highlight query terms in search hits on the search results page.</td>
</tr>
</tbody>
</table>
Table 4-1 Style Sheet Properties (continued)

<table>
<thead>
<tr>
<th>Style Sheet Property Link</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results›Highlighted link</td>
<td>Specifies whether and how to display the “Highlight” link to the right of each hit on the search results page. Clicking this link causes the query terms to be highlighted in the document.</td>
</tr>
<tr>
<td>Results›Left score bar</td>
<td>Specifies how to display the left side of the score bar.</td>
</tr>
<tr>
<td>Results›Publisher in hit</td>
<td>Specifies whether and how to display the publisher of a document in the search hit.</td>
</tr>
<tr>
<td>Results›Results</td>
<td>Specifies how search hits are displayed in the results page.</td>
</tr>
<tr>
<td>Results›Right score bar</td>
<td>Specifies how to display the right side of the score bar.</td>
</tr>
<tr>
<td>Results›Score in hit</td>
<td>Specifies whether and how to display the relevance score of search hits.</td>
</tr>
<tr>
<td>Results›Size in hit</td>
<td>Specifies whether and how to display the size of documents in search hits.</td>
</tr>
<tr>
<td>Results›Title in hit</td>
<td>Specifies whether and how to display the title of documents in search hits.</td>
</tr>
<tr>
<td>Results›URL in hit</td>
<td>Specifies whether and how to display a document’s URL in search hits.</td>
</tr>
<tr>
<td>Results›Word scores</td>
<td>Specifies whether and how to display the relevance score of individual query terms within a query.</td>
</tr>
<tr>
<td>Miscellaneous›Advertisement</td>
<td>Specifies whether to display an advertisement.</td>
</tr>
<tr>
<td>Miscellaneous›Navigation</td>
<td>Specifies whether to display the navigation bar above search hits.</td>
</tr>
<tr>
<td>Miscellaneous›Quick Links</td>
<td>Specifies whether to display Quick Links at the top of the results page after a user searches for predefined keywords and phrases. See “Creating Quick Links” on page 181 for more information.</td>
</tr>
<tr>
<td>Miscellaneous›Result count</td>
<td>Specifies whether and how the result count is displayed in the navigation bar.</td>
</tr>
<tr>
<td>Miscellaneous›“Skip to content” link</td>
<td>Specifies whether and how the Skip to content link is displayed in the navigation bar.</td>
</tr>
<tr>
<td>Miscellaneous›Spelling suggestion</td>
<td>Specifies whether and how to display spelling suggestions above the search results and navigation bar.</td>
</tr>
<tr>
<td>Miscellaneous›Spelling suggestion emphasis</td>
<td>Specifies how to apply emphasis to words in a spelling suggestion.</td>
</tr>
</tbody>
</table>
Table 4-1  Style Sheet Properties (continued)

<table>
<thead>
<tr>
<th>Style Sheet Property Link</th>
<th>Description</th>
</tr>
</thead>
</table>
| Miscellaneous›Various search options | Specifies whether and how to display the following links within the navigation bar:  
  - Score using date  
  - Group by location  
  - Hide summaries |
| Topics›Group topics | Specifies whether and how to display the group topics. |
| Topics›Subtopic names | Specifies how to display subtopic names. |
| Topics›Related topics | Specifies whether and how to display related topics. |
| Topics›Result subtopics | Specifies whether to display result subtopics. |
| Topics›Topic breadcrumb | Specifies whether and how to display the topic navigation path. |
| Topics›Topic in hit | Specifies whether and how to display topics that appear in hits. |
| Topics›Topic names | Specifies how to display topic names. |
| Topics›Topic section | Specifies whether and how to display topic sections. |
| Topics›Topic text | Specifies whether and how to display topic text. |

**Figure 4-2** Style Sheet Properties on Initial Search Page

![Style Sheet Properties on Initial Search Page](image-url)
Various search options

Result count

Navigation

Collection in hit

Score in hit

Description in hit

Highlighted term

Title in hit

URL in hit

Date in hit

Size in hit

Highlight link

Score bar link

Find similar

Tips for Using the Euro

September 2001, banks in member nations have been stockpiling up to 14 billion euro banknotes and 50 billion euro coins meant to meet the demand. The colorful euro bills, which include holograms and...


Tips for Using the Euro

September 2001, banks in member nations have been stockpiling up to 14 billion euro banknotes and 50 billion euro coins meant to meet the demand. The colorful euro bills, which include holograms and...

http://www.sperren.com/gentechnik_euro_28_6.htm - 25 KB

Tips for Using the Euro

September 2001, banks in member nations have been stockpiling up to 14 billion euro banknotes and 50 billion euro coins meant to meet the demand. The colorful euro bills, which include holograms and...

http://www.sperren.com/gentechnik_euro_28_6.htm - 25 KB

Tips for Using the Euro

September 2001, banks in member nations have been stockpiling up to 14 billion euro banknotes and 50 billion euro coins meant to meet the demand. The colorful euro bills, which include holograms and...

http://www.sperren.com/gentechnik_euro_28_6.htm - 25 KB

Tips for Using the Euro

September 2001, banks in member nations have been stockpiling up to 14 billion euro banknotes and 50 billion euro coins meant to meet the demand. The colorful euro bills, which include holograms and...

http://www.sperren.com/gentechnik_euro_28_6.htm - 25 KB
Setting Server Parameters

After you’ve edited a style sheet, use the Interface | Style and Interface | Query pages to set the server parameters associated with the style sheet:

- Setting Style Parameters
- Setting Query Parameters

Setting Style Parameters

Style parameters specify the default style, language, header, and footer for your search forms.

To set style parameters:

1. Click Interface in the Administrative Interface.
2. Click the Style tab.
   The Style page is displayed, as shown in Figure 4-4.
3. Use the controls to edit the style parameters described in Table 4-2 on page 125.
4. Click OK.
The following table describes the controls and style parameters shown in Figure 4-4.

**Table 4-2  Style Page Parameters**

<table>
<thead>
<tr>
<th><strong>Style Parameter</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Style</td>
<td>Lists all available Ultraseek styles.</td>
</tr>
<tr>
<td>New</td>
<td>Displays the New Interface Style page. Use this page to name a new style and designate its base style. Style names may contain only letters and numbers.</td>
</tr>
<tr>
<td>Default style</td>
<td>Specifies that the current style be the default for all search and results forms.</td>
</tr>
<tr>
<td>Enter any HTML that should always appear at the top of the search page</td>
<td>Specifies a header to appear in all forms associated with the selected style.</td>
</tr>
</tbody>
</table>
Enter any HTML that should always appear at the bottom of the search page | Specifies a footer to appear in all forms associated with the selected style.

Default Language | Specifies the default language used with the selected style. Ultraseek dynamically negotiates the language of displayed content by examining the Language and Accept-Language fields in the HTTP request header, so the default language may not always be used. For example, if the request header fields indicate that the user prefers German, then a German-supporting instance of Ultraseek will display content in that language rather than the default language associated with the style.

Launch Style Editor | Displays the Style Editor window. Use this editor to edit the style sheet associated with your Ultraseek style, as described in “Editing Style Sheets” on page 119.

**Setting Query Parameters**

Query parameters enable you to customize the appearance, content, and functionality of your search and search results pages.

**To set query parameters:**

1. Click **Interface** in the Administrative Interface.

2. Select the Ultraseek style that you want to configure in the Style list box on the Style page.

3. Click the **Query** tab.
   * The Query page is displayed, as shown in Figure 4-5.

4. Use the controls to set the query parameters described in Table 4-3.

5. Click **OK**.
Figure 4-5  Interface | Query Page

4  Customizing Ultraseek Search Forms
Setting Server Parameters
### Table 4-3  Query Parameters

<table>
<thead>
<tr>
<th>Query Parameter</th>
<th>Description</th>
</tr>
</thead>
</table>
| Default query mode       | Indicates which query mode to present by default on the initial search results page. Users can override this setting by selecting a different option when refining search results or starting a new search. Ultraceek provides the following query modes on the Search Results page:  
  - *start new search*—Each successive query is treated as a new search, independent of all previous queries.  
  - *search these results*—Each successive query is treated as a refinement of the previous query, narrowing the results down each time.  
  - *search entire Web*—Each query is redirected to search a web-wide search engine.  
You can also manually set this parameter with the rq form variable. See the *Verity Ultraceek Customization Guide* for more information. You can also manually set this parameter with the rq form variable. For more information, see the *Verity Ultraceek Customization Guide*. |
| Default results filter   | Specifies the default filter to use for sorting results on the search results page:  
  - *sort by relevance*—The results display in the original relevance-ranked order. This setting does not affect the results of a query.  
  - *score using date, sort by relevance*—The results display in the original relevance-ranked order, with more recent dated documents given a boosted score.  
  - *sort by title*—The results are sorted alphabetically by document title.  
Users can override this setting by selecting a different option when refining search results or starting a new search.  
You can also manually set this parameter with the rf form variable. For more information, see the *Verity Ultraceek Customization Guide*. |
### Table 4-3 Query Parameters (continued)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Display number of hits</td>
<td>Specifies the number of search hits to display on the search results page. The default is 10. You can change this setting to display 25, 100, or 500 hits per page. Note that displaying more than 10 hits slows queries and increases the size of results pages. Consider selecting the Hide Summaries option in the Default look field if you choose to display more than 10 hits per page. You can also manually set this parameter with the nh form variable. For more information, see the Verity Ultraseek Customization Guide.</td>
</tr>
<tr>
<td>Default look</td>
<td>Indicates whether Ultraseek should hide or display summary information on the search results page. Users can override this setting by choosing to hide or display results summaries on the search results page.</td>
</tr>
<tr>
<td></td>
<td>- Show summaries—Ultraseek displays detailed information for each search result, including score, title, description, date, URL, and collection name.</td>
</tr>
<tr>
<td></td>
<td>- Hide summaries—Ultraseek displays only the score and title for each search result.</td>
</tr>
<tr>
<td>Default query form complexity</td>
<td>Indicates whether to display a simple or advanced query form in the search interface. Users can override this setting.</td>
</tr>
<tr>
<td></td>
<td>- Simple—Display a simple search box, with rotating tips.</td>
</tr>
<tr>
<td></td>
<td>- Advanced—Display an advanced search form that contains options for users to include or exclude terms, search by date ranges, and specify how to filter and display search results. The advanced form helps users create more specific queries, ensuring more specific search results.</td>
</tr>
<tr>
<td>Results page width</td>
<td>Specifies the width of the search results page, as an integer or a percentage:</td>
</tr>
<tr>
<td></td>
<td>- Integer—The width of the browser, in number of pixels.</td>
</tr>
<tr>
<td></td>
<td>- Percentage—The percentage of the browser window that should be used to display results, from 0 to 100%.</td>
</tr>
</tbody>
</table>
### Table 4-3  Query Parameters (continued)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Show query form above or below hits</strong></td>
<td>You can control the placement of the search box relative to the hits on search results pages. By selecting the corresponding check box for this parameter, you control whether the search box displays above the returned results, below the returned results hits, or both, on search results pages. When you select both options, the query form displays twice in the user interface, once above the results and once below the results. When you de-select both options the query form does not display on the results page.</td>
</tr>
<tr>
<td><strong>Advanced query fields</strong></td>
<td>Enables you to customize the field search option on the advanced query form. The first part of the entry (up to the comma) reflects the field name in the resulting query. The rest of the entry reflects the label that displays in the drop-down list on the advanced search form. The default field search values are below: , in the body, in the title, in the URL, in the site name, in a link, in an image link, in image alt text, in the description, in the keywords, in remote anchor text. You can add additional field search menu entries by adding the field search syntax to the Advanced query fields text box. For more information about field searches, see “Field Searches” on page 236.</td>
</tr>
<tr>
<td><strong>Group search results by location by default</strong></td>
<td>Ultraseek dynamically defines a location as a URL, a URL + directory, or a URL + two directories. You can also set this parameter with the <code>sc</code> form variable. For more information, see the Verity Ultraseek Customization Guide.</td>
</tr>
<tr>
<td><strong>Show individual word scores</strong></td>
<td>Displays individual word scores on search results pages for each word in multi-word queries. Individual word scores are helpful for queries that contain many unrelated words. You can also set this parameter with the <code>ws</code> form variable. For more information, see the Verity Ultraseek Customization Guide.</td>
</tr>
<tr>
<td>Table 4-3</td>
<td>Query Parameters (continued)</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------</td>
</tr>
<tr>
<td><strong>Provide option to search the Internet</strong></td>
<td>Displays an option for users to search the Internet on the initial search pages and on results pages. You can also set this parameter with the <code>si</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>Require login for result filtering</strong></td>
<td>Requires all users to enter a username and password before making queries. You can also set this parameter with the <code>rl</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>Show spelling suggestions</strong></td>
<td>Enables Ultraseek to display spelling suggestions for search queries. You can also set this parameter with the <code>enable_spell</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>Show passage-based summaries</strong></td>
<td>Enables Ultraseek to display passage-based summaries of documents returned with search results. You can also set this parameter with the <code>enable_passages</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>Lowercase query terms</strong></td>
<td>Formats all search queries in lowercase. You can also set this parameter with the <code>lowercase_query</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>Highlight matching search terms in results</strong></td>
<td>Applies formatting for highlights to search terms where they appear in results. For more information on configuring highlight formatting, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>When using this style, the user is restricted to these collections:</strong></td>
<td>Sets the collections to search and display on the search page when using the selected style. You can also set these parameters with the <code>col</code> and <code>qc</code> form variables. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
<tr>
<td><strong>When using this style, the user is restricted by this query prefix</strong></td>
<td>Query prefixes prepend a field to all queries, which allows you to create multiple search presences, or virtual collections, from the same collection. The following query prefix specifies that all queries are made against indexed content from the BBC news site: <code>site:www.bbc.com.uk</code> You can also set this parameter with the <code>qp</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
</tbody>
</table>
### Table 4-3  Query Parameters (continued)

<table>
<thead>
<tr>
<th>When using this style, the user is restricted by this query suffix</th>
<th>Query suffixes append a field to all queries, which allows you to create multiple search presences, or virtual collections, from the same collection. The following query suffix specifies that search queries do not return documents that were marked as private when they were indexed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-status:private</td>
<td>You can also set this parameter with the <code>default_qs</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date Range Default</th>
<th>Restricts the documents returned to users to those indexed within a specified date range. Check the “Anytime” radio button if you do not want to restrict documents based on the date they were indexed.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>You can also set this parameter with the <code>dt:</code> form variable. For more information, see the <em>Verity Ultraseek Customization Guide</em>.</td>
</tr>
</tbody>
</table>
Generating Ultraseek Reports

This chapter describes how administrators can monitor the performance of the Ultraseek server and generate reports based on this information. Ultraseek creates daily and biweekly log digests that contain query and trend data compiled from the server’s log files. These digests are created at midnight and reflect data through the previous day.

Ultraseek reports complement third-party logging tools, but are not intended to replace them. You can generate the following types of reports:

- Generating Reports
- Generating Top Queries Reports
- Generating No-Results Reports
- Generating No-Clickthrough Reports
- Generating Usage Summaries
- Generating Most Requested Documents Reports
- Generating Trend Reports
- Generating Reports Per Topic
- Monitoring Recent Queries
To generate Ultraseek reports:

1. Click Activity | Reports.
2. Select the type of report that you want to generate in the Report column.
3. Specify the date or date range for the report in the Duration column.
4. Click make report.

**Note** Optionally specify that the report is downloaded as tab separated values by selecting the **download as tab separated values** checkbox.
### Figure 5-1 Reports Page

#### Query Reports

<table>
<thead>
<tr>
<th>Report</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Queries</td>
<td></td>
</tr>
<tr>
<td>Top Queries with No Results</td>
<td></td>
</tr>
<tr>
<td>Top Queries with No Clickthrough</td>
<td></td>
</tr>
<tr>
<td>Top Requested Documents</td>
<td></td>
</tr>
<tr>
<td>Usage Summary</td>
<td></td>
</tr>
</tbody>
</table>

- **By Date:** May 23 2005
- **Range:** From: Apr 23 2005 To: May 23 2005

- Download as tab separated values

#### Query Trend Report

<table>
<thead>
<tr>
<th>Old Month</th>
<th>New Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 2005</td>
<td>May 2005</td>
</tr>
</tbody>
</table>

- Download as tab separated values

#### Topic Report

<table>
<thead>
<tr>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents per topic, sorted by hierarchy</td>
</tr>
<tr>
<td>Documents per topic, sorted by population</td>
</tr>
<tr>
<td>Topics, sorted by browse frequency</td>
</tr>
</tbody>
</table>

- Download as tab separated values

#### Recent Queries

- List of recent queries

---

*Verity Ultraseek Administrator Guide*
Administrators can generate a report listing the most frequent queries, as shown in Figure 5-2. Click on a query to perform that search.

**Figure 5-2** Top Queries Report

<table>
<thead>
<tr>
<th>Rank</th>
<th>Query</th>
<th>Times</th>
<th>Avg. Clicks</th>
<th>Null Results</th>
<th>Avg. Page Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>url:http</td>
<td>17</td>
<td>0.41</td>
<td>0</td>
<td>1.06</td>
</tr>
<tr>
<td>2</td>
<td>url:biy烽silicon technology</td>
<td>9</td>
<td>0.11</td>
<td>0</td>
<td>2.67</td>
</tr>
<tr>
<td>3</td>
<td>lninfo:nACOORDINATORTYPE</td>
<td>4</td>
<td>0.25</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>cas</td>
<td>4</td>
<td>0.00</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>Iraq</td>
<td>4</td>
<td>0.00</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>size:large</td>
<td>4</td>
<td>0.00</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>-topic:833647206, Clean Energy</td>
<td>4</td>
<td>0.25</td>
<td>0</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5-1 describes the metrics used in a top queries report:

**Table 5-1** Top Query Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>Integer between 1 and 100, where 1 is the highest rank.</td>
</tr>
<tr>
<td>Query</td>
<td>The ranked query. Click the query to perform the search.</td>
</tr>
<tr>
<td>Times</td>
<td>The number of times users queried the server with the search term.</td>
</tr>
<tr>
<td>Avg. Clicks</td>
<td>The average number of clicks users performed on the results for this query. A value of 1.00 indicates that there was one click on each set of results pages, on average.</td>
</tr>
<tr>
<td>Null Results</td>
<td>The number of times that no documents were retrieved for the search term.</td>
</tr>
<tr>
<td>Avg. Page Views</td>
<td>The average number of results pages viewed for this query.</td>
</tr>
</tbody>
</table>
Generating No-Results Reports

Administrators can generate reports listing queries that return no search hits and the number of times these queries were submitted to the server. Examine this report to see if it includes queries that you think should return results or to identify new content to add to your collections. Click a query to perform that search.

Use any of the following methods to address no-result queries:

- Add the query to the Thesaurus
- Add a Quick Link
- Add new content which matches the query
- Enable spelling suggestions

Figure 5-3 shows a no-results report for a single day.

**Figure 5-3** No Result Report

<table>
<thead>
<tr>
<th>Rank</th>
<th>Query</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bas</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>sizelarge</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>size:small</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td><a href="http://www.boomer-bikes.com">www.boomer-bikes.com</a></td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>38500</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>+topic:833647206, Clean Energy II amorphos</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>i.shore</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>&quot;learn chinese&quot;</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>NKWERA</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>size:medium</td>
<td>2</td>
</tr>
</tbody>
</table>
Generating No-Clickthrough Reports

Administrators can generate reports listing queries that produced no clickthroughs. A clickthrough occurs when a user clicks a link to a document on a search results page. There are two reasons why users did not click search hits on the results pages: they found the information they wanted in the summary, or the query did not return relevant results. If the results are not relevant, use one of the methods described in “Generating No-Results Reports” on page 137 in to improve the results.

Figure 5-4 shows a no-clickthrough report.

**Figure 5-4** No-Clickthrough Report

![No-Clickthrough Report](image)
Generating Usage Summaries

Usage summaries display information about queries, unique queries, results served, pages viewed, and clicks for a day or a period of time. Figure 5-5 shows a daily usage summary.

Figure 5-5 Usage Report
Administrators can generate a report of the most requested documents for all collections managed by a single Ultraseek server, as shown in Figure 5-6. Each row in the report displays the URL associated with the document, the queries that retrieved the document, and the total number of times the document was retrieved. Click a document’s URL to open that document.

**Figure 5-6** Top Requested Documents Report

<table>
<thead>
<tr>
<th>Rank</th>
<th>URL</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://intranet.verity.com/development/docs/Archives/k2v500/html/VerityIntelligentClassificationGuide/app_b_VQL8.html">http://intranet.verity.com/development/docs/Archives/k2v500/html/VerityIntelligentClassificationGuide/app_b_VQL8.html</a></td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://conn00.verity.com:8765/db/dbrec.html?col=kaliba&amp;db=TSI_ACTIONS_TEXT_UTASEQ_TS1=449310&amp;charset=utf-8&amp;Top_Queries=swimming(12)">http://conn00.verity.com:8765/db/dbrec.html?col=kaliba&amp;db=TSI_ACTIONS_TEXT_UTASEQ_TS1=449310&amp;charset=utf-8&amp;Top_Queries=swimming(12)</a>;</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://conn00.verity.com:8765/db/dbrec.html?col=kaliba&amp;db=TSI_ACTIONS_TEXT_UTASEQ_TS1=365950&amp;charset=utf-8&amp;Top_Queries=marianne_kilkenny%22+%22charu%22(6);%22+%22marianne_kilkenny%22+%22charu%22(1)">http://conn00.verity.com:8765/db/dbrec.html?col=kaliba&amp;db=TSI_ACTIONS_TEXT_UTASEQ_TS1=365950&amp;charset=utf-8&amp;Top_Queries=marianne_kilkenny&quot;+&quot;charu&quot;(6);&quot;+&quot;marianne_kilkenny&quot;+&quot;charu&quot;(1)</a>;</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td><a href="http://conn00.verity.com:8765/db/dbrec.html?col=biqdb&amp;tb=SHARE_VIEW&amp;n">http://conn00.verity.com:8765/db/dbrec.html?col=biqdb&amp;tb=SHARE_VIEW&amp;n</a> thí deliberately_s=0959-1&amp;Top_Queries=FILE_NAME&quot;resume.g025.txt&quot;(2);urlhttp(1);</td>
<td>3</td>
</tr>
</tbody>
</table>
Generating Trend Reports

Trend reports rank queries over time, as shown in Figure 5-7. Each row in a trend report shows a search term's ranking at the beginning and end of any period of time that you choose. Queries occurring the same number of times have the same rank. Clicking a query performs that search.

Figure 5-7  Trend Report
Generating Reports Per Topic

You can generate reports on topics by several criteria:

- Documents per topic, sorted by hierarchy
- Documents per topic, sorted by population
- Topics, sorted by browse frequency

A report on topics by frequency is shown in Figure 5-8.

**Figure 5-8** Viewing Topics Sorted by Browse Frequency

<table>
<thead>
<tr>
<th>topics</th>
</tr>
</thead>
</table>

**Topic Visits by Frequency**

Results from past 2 days

22 Total visits

11 Total topic browses

4 Arts and Culture > Books
3 Arts and Culture > Music
2 Arts and Culture
1 Arts and Culture > Dance
1 Science and History
Monitoring Recent Queries

Administrators can monitor queries recently submitted to the Ultraseek server by viewing the recent queries text box on the Reports page. The text box displays the last 50 user queries still in memory. When you restart the server, you lose all of the previous queries. Monitoring recent queries is useful for determining the type of information that your users want. You may find that some queries are very long. These queries occur when users click the Find Similar link for a search hit on the results page.

To monitor recent queries:

1. Click Activity | Reports.

   The most recent queries are displayed in the recent queries text box.

2. Refresh your browser to view queries submitted to the server since you clicked the Reports tab.

Figure 5-9 shows the most recent queries for an Ultraseek server.

Figure 5-9  Recent Queries
5 Generating Ultraceek Reports
Monitoring Recent Queries
Monitoring Activity

To view server and site search activity, click the Activity button. The Activity window displays system performance and activity information, such as recent queries and the status of your collections. You can use this window to quickly and easily obtain information about the status of Ultraseek.

To refresh the Activity window, click the Ultraseek banner at the top of the page.

The Activity window includes the following information:

- Collection Status
- Performance Information
- HTTP Server Threads
Collection Status

The Status tab displays the status of each collection indexed by Ultraseek, including the number of sites being indexed and the total number of documents in each collection. You can use this report to determine the current size of each index in terms of site, URL, and document indexing statistics.

Figure 6-1 Collection Status
The Status page for each collection in the Ultraseek administrative interface will automatically refresh every 15 seconds if the status is not *idle*, *suspended*, or *curfew in effect*. To change the refresh rate, add the following patch to the bottom of patches.py with the new value:

```python
import config
cfg.refresh_rate = 15
```
A value of 0 will disable the automatic refresh.

### Status

Indicates one of the following collection statuses:

- **Idle**—The indexer has completed crawling all links in the collection, and is now waiting for some event to trigger a revisit request.
- **Active**—The indexer is currently evaluating one or more URLs for updates to the index. This may indicate that your site or network has not yet been completely indexed, or that some event has triggered the spider to revisit documents currently in the index.
- **Suspended**—The indexer has been stopped, while still allowing search of the existing index.
- **Error**—The collection has encountered a problem with one of the sites or databases. If viewing the log generated from the indexing task does not enable you to solve the problem, check the FAQs on the Verity support site.

### Threads

Indicates the number of server threads that are currently running.

### Sites

Shows the number of different web sites that have been contacted in the process of following all links allowed by collection settings.

### URLs

Indicates how many URLs have been found by the indexer. A URL that is discovered by the indexer is not necessarily included in the index.

### Docs

The document count shows how many unique documents are in the index. If essentially identical documents are found on different Web servers on your network, only one copy is kept in the index. You can tell Ultraseek which one to keep (see “Filtering Duplicate URLs” on page 83).
Performance Information

The Performance tab includes query statistics for the last 10 minutes, 1 hour, and 24 hour time periods. Using this report, you can analyze the response times for queries against the server, view the statistics for query rates, and view the total number of queries requested on your site.

Figure 6-2 Performance Tab

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Query Time</th>
<th>Query Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 min</td>
<td>0 ms</td>
<td>0 q/s</td>
</tr>
<tr>
<td></td>
<td>0 ms</td>
<td>0.000 q/s</td>
</tr>
<tr>
<td></td>
<td>0 ms</td>
<td>0 q/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 hour</td>
<td>0 ms</td>
<td>0 q/s</td>
</tr>
<tr>
<td></td>
<td>0 ms</td>
<td>0.000 q/s</td>
</tr>
<tr>
<td></td>
<td>0 ms</td>
<td>0 q/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 hours</td>
<td>0 ms</td>
<td>0 q/s</td>
</tr>
<tr>
<td></td>
<td>0 ms</td>
<td>0.000 q/s</td>
</tr>
<tr>
<td></td>
<td>0 ms</td>
<td>0 q/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

total queries = 0
HTTP Server Threads

The Threads tab displays the current status of HTTP server threads, the client IP address that a thread is currently contacting, and the request line the server is sending to the client. HTTP threads are sent to Ultraseek when an end-user or another system needs to communicate with the server. For example, HTTP threads can be initiated by an Ultraseek administrator, an end-user using the search site, or by XPA calls to the server.

HTTP server threads can have the following statuses.

- **waiting**: The server is idle, and is waiting for an HTTP request.
- **working**: The server is performing a transaction in response to an HTTP request.
- **accepting**: The server is in the process of establishing a connection in response to an HTTP request.
- **in keep-alive status**: The server is in keep-alive status, which allows clients to keep the HTTP connection open after making a request to the server. Persistent keep-alive HTTP sessions allow clients to send multiple requests over the same TCP connection. To set the keep-alive status of the server, see “Setting Advanced Server Parameters” on page 158.

![Figure 6-3 HTTP Server Threads](image_url)
6 Monitoring Activity
HTTP Server Threads
You manage the overall system, server, and user parameters using the Server features of the Ultraseek administrative interface. Click the Server button on the Navigation bar to access the server management features.

This chapter explains how to perform the following server administration tasks:

- Managing the Server
- Setting Server Parameters
- Managing Administrative Users and Client Access
- Managing Document Types
- Creating XML Mappings
- Creating Quick Links
Managing the Server

This section describes how to obtain information on the operating system status and how to restart and shut down the Ultraseek server.

Viewing Ultraseek Status

To view Ultraseek system information:

1. Click Server.
2. Select the Parameters tab.
3. Select the Main tab.
4. Click OS Status.

Ultraseek displays a report that contains information about Ultraseek system warnings, collection settings, server parameter settings, Ultraseek program files, and additional information about your system. The information in this report can help you determine the nature of system problems without exiting Ultraseek.

Restarting the Server

You need to restart the Ultraseek server if you have modified the patches.py file. You may also need to restart the server if Ultraseek is utilizing too much CPU load, memory, or other system resources.

To restart the server, click the Restart/Shutdown icon in the Administrative Interface.
Shutting Down and Restarting the Server

This section describes how to shut down and restart the Ultraseek server.

**Shutting Down the Server**

To shut down the server, click the Restart/Shutdown icon in the Administrative Interface.

We recommend that you restart Ultraseek for most purposes instead of shutting it down. In addition, do not shut down the server if you intend to restart the service immediately. Restart the server instead.

When you shut down Ultraseek, you cannot access the Administrative Interface from a browser (Solaris only) until you restart the server from the command line or reboot your system.

In Solaris, use **Shutdown** only if you can restart the server in one of the following ways:

- Restart Ultraseek from the machine it runs on
- Restart Ultraseek from the command line
- Rebooting the machine Ultraseek resides on

**Restarting the Server**

See “Starting Ultraseek” on page 27 for more information on restarting Ultraseek after a complete server shutdown.
Setting Server Parameters

To change server parameters, choose the Server button in the Administrative Interface. You can set the following server parameters:

- Setting Main Server Parameters
- Setting Advanced Server Parameters

Changing some of the parameters will result in a prompt asking to restart the server. You can postpone restarting the server until you have completed your changes.

Setting Main Server Parameters

Server parameters on the Main tab enable you to specify information about your server, such as the hostname, server port, and encoding properties.

To set main server parameters:

1. Click the Server button.
2. Click the Parameters tab.
3. Click the Main tab.

The Main Server Parameters page displays, as shown in Figure 7-1; each parameter is described in the table that follows the figure.
Advanced server parameters enable you to optimize the performance of the Ultraseek server. Using server parameters on the Advanced Server Parameters page, you can set HTTP connection information, specify log file message levels, and set additional server management settings.

**Figure 7-1 Setting Main Server Parameters**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Users</th>
<th>Doctype</th>
<th>XML Mappings</th>
<th>Quick Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **view logs**
- **08 status**

License key
License hostname
Licensed languages
Automatically check for updates
Hostname
Binding address
Email addresses for important automatic notifications (comma-separated)
Email addresses for automatic bug reports (comma-separated)
SNMP mail forwarding host
Document directory
HTTP server port
HTTPS server port
Encoding
Language

Ultraseek generates the following types of log files:

- **Access**—Displays information that logs access to the Ultraseek server.
- **Error**—Displays a detailed list of Ultraseek activity, including information about any errors encountered during spidering and indexing. Each error log entry shows the date and time that an error occurred, followed by a message describing the event. If the event was a HTTP command, the message also includes the HTTP response code.
- **Query**—Displays a list of recent queries on your search site, including the date and time of the queries, the number of results returned, and the internal name of collections searched by users.

Note: The level of information in the error log file is specified by the log level parameters on the Advanced Server Parameters page. See “Setting Advanced Server Parameters” on page 158 for more information.
License key
Displays the current license key for Ultraseek. Click **new** to enter a new license key if your license key is invalid or has expired.

Each instance of Ultraseek requires a license key. The license key determines the maximum number of documents that can be indexed, the number of web sites that can be spidered, and the number of collections that you can create using Ultraseek. The license key also determines access to add-in modules, such as CCE and support for stemming in multiple languages.

As of version 5.2, valid license keys are issued only for one of the following:

- Domain-qualified hostname
- IP address
- Unqualified hostname, excluding localhost (time-limited key)

Existing license keys become invalid in the following scenarios:

- If you are upgrading to a new major version. For example, you need a new key if you are upgrading from version 4.x to 5.x, but not if you are upgrading from 5.0.x to 5.1.x.
- If you are upgrading to Ultraseek 5.2 and your existing license key was issued to an unqualified hostname, such as localhost.
- If you are upgrading to Ultraseek 5.2 and want to index database documents, but your existing Ultraseek license does not distinguish between database and non-database documents.

Contact software-sales@verity.com for more information about obtaining new license keys.

Automatically check for updates
Specifies that Ultraseek automatically connect to the Verity support site to check for newer versions of the software. A notification appears at the bottom of the administration interface when a newer version is available.

Hostname
Specifies the name of the server where you are running Ultraseek. For example, searchhost.verity.com. You can also use symbolic aliases for the hostname syntax.

Ultraseek uses the hostname when generating titles on results page and when generating redirect responses.

Binding address
Specifies the name of the Internet address to bind to on the host. It is necessary to bind Ultraseek, the web server, or other applications to separate addresses. This is to alleviate any conflicts which may arise due to different applications attempting to use the same IP address. This value can either be an IP address or a host name that resolves to an IP address.

Leave this setting blank if you want to bind Ultraseek to all addresses on the host.
### Email addresses for important automatic notifications

This field is for operational notifications from Ultraseek. Notifications such as low memory, collections which cannot start up, or spidering problems will be sent through this setting. Separate multiple email addresses with commas. Click **test** to ensure that Ultraseek can send email to the intended recipients.

### Email addresses for automatic bug reports

This field is for development notifications from Ultraseek. Bug reports are sent when Ultraseek detects internal error conditions, and include Python stack traces and other information about the errors. Separate multiple email addresses with commas. Click **test** to ensure that Ultraseek can send email to the intended recipients.

### SMTP mail forwarding host

Specify the SMTP mail forwarding host for your site. Ultraseek uses the SMTP host first when sending email for important notifications and bug reports. If Ultraseek cannot reach email recipients using the SMTP host, it then attempts to send email to the site host name specified in the email address.

### Document directory

Ultraseek checks the document directory for scripted HTML documents for its built-in HTTP server. You can modify most pages generated by Ultraseek, including the search and results pages. If you choose to modify these pages, do not overwrite the default Ultraseek pages. Instead, store your modified pages in a different directory and identify the new directory in the document directory field. For example, if you created a new directory named `company` in the Ultraseek program directory, specify:

```
C:\Program Files\Verity\Ultraseek\company
```

as the path in the document directory field. The above example is based upon an absolute pathname, however, the pathname can also be relative to the Ultraseek program directory.

For performance reasons, Ultraseek keeps in memory the HTML template files that generate each screen. Click **reload** to flush the page cache of all current templates.

If you change any documents in the documents directory, you must click **reload** so that Ultraseek can cache the new documents.

### HTTP server port

By default, Ultraseek communicates on port 8765. To change the TCP port that Ultraseek uses for its built-in HTTP server, enter it in this field. The standard port number for HTTP servers is 80. If you change the port address, remember to also change any URLs that point to your search engine.

If you are running any other web server or application on port 80, you will not be able to run Ultraseek on the same port. Make sure that only one Web or Ultraseek server is running on port 80 using the same IP address at any given time to alleviate any possible conflicts. See “Binding address” on page 156 for more information on running separate applications on the same port using different IP addresses.
Setting Advanced Server Parameters

This sections describes how to set advanced server parameters for Ultraseek.

To set advanced server parameters:

1. Click the Server button.
2. Click the Parameters tab.
3. Click the Advanced tab.

The Advanced Parameters page displays, as shown in Figure 7-2. The parameters are described in the table following the figure.

<table>
<thead>
<tr>
<th><strong>HTTPS server port</strong></th>
<th>If you are using the SSL security module, specify the TCP port to use for secure access to the Ultraseek HTTPS server. The standard port number for HTTPS servers is 443.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Encoding</strong></td>
<td>Specifies the default character encoding for pages returned by Ultraseek's built-in web server.</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>Specifies the default language to use for Ultraseek’s search and help pages. Users can override this setting on the search and help pages by choosing a different language if you have multiple language support enabled.</td>
</tr>
</tbody>
</table>
**Figure 7-2** Advanced Server Parameters

![Advanced Server Parameters](image-url)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of server threads</td>
<td>10</td>
</tr>
<tr>
<td>HTTP connection timeout (seconds)</td>
<td>15</td>
</tr>
<tr>
<td>HTTP keep-alive timeout (seconds)</td>
<td>5</td>
</tr>
<tr>
<td>Operating system log level</td>
<td>Warning</td>
</tr>
<tr>
<td>File log level</td>
<td>Info</td>
</tr>
<tr>
<td>Number of old access, click, and query log files to keep</td>
<td>400</td>
</tr>
<tr>
<td>Number of old error log files to keep</td>
<td>7</td>
</tr>
<tr>
<td>Term list cache size (megabytes)</td>
<td>64</td>
</tr>
<tr>
<td>Maximum term list length (documents)</td>
<td>131071</td>
</tr>
<tr>
<td>Title cache size (megabytes)</td>
<td>0</td>
</tr>
<tr>
<td>Maximum storage heap size (megabytes)</td>
<td>256</td>
</tr>
<tr>
<td>Filter number of hits</td>
<td>100</td>
</tr>
<tr>
<td>Average group size</td>
<td>5</td>
</tr>
<tr>
<td>Maximum group subhits</td>
<td>3</td>
</tr>
<tr>
<td>Group score threshold</td>
<td>10</td>
</tr>
</tbody>
</table>

**In-document highlighting filter**
- wildcard
  - allow

**In-document highlighting proxy server**
- new

**Recreate log digests**
- recreate

- Collect clickthrough information
- Allow HTTP keep-alives
- Resolve client addresses for log
- Log requests for images
- Use Stemming
- Debug mode
- Parse binary data in database tables
- Remove from results duplicate hits with the same URL
- Allow Wildcard queries
- Interleave Sites
- Filter results based on user credentials

[Ok] [Cancel] [Help]
### Number of server threads
Specifies the number of server threads that Ultraseek should devote to its HTTP service. The number of threads determines how many simultaneous HTTP connections Ultraseek's built-in HTTP server can handle. The default is 5. Each server thread requires one megabyte of main memory for its stack.

Note: The NNTP spider uses multiple threads to access the same news server. This is different from the HTTP spider, which only uses one thread per host. If the news server is heavily loaded or on a slow machine, reduce the number of spider threads.

### HTTP connection timeout
Specifies the maximum number of seconds to wait for a request from a remote HTTP client before giving up. The default is 15 seconds.

### HTTP keep-alive timeout
Specifies the maximum number of seconds to wait for another request from a remote HTTP client after servicing an initial request on a connection. The default is 5 seconds.

### Unix username (Solaris/Linux Only)
By default, the Unix username is the same username that was entered when Ultraseek was started as “root” the first time. Ultraseek uses the username “daemon” to preserve security. Specify a username to change the name associated with the Ultraseek process at run time.

Note: Ultraseek generally does not require special account privileges to run.

### Syslog facility (Solaris/Linux Only)
This setting indicates how Ultraseek should identify itself when creating syslog messages.

The default is “daemon”. See the man page for `syslog.conf` for more information.

### Operating system log level
Specifies the minimum log message severity required to send log messages to the operating system. You can view messages using the operating system log facilities.

The operating system log levels are as follows:

- **error**—The operating system logs only Ultraseek errors. This log level allows you to maintain the cleanest log.

- **warning**—The operating system logs both warning and error messages. This log level allows you to maintain a more detailed log. This is the default setting.

- **notice**—The operating system logs all notification, warning, and error messages.

- **info**—The operating system logs all messages generated by Ultraseek, including informational, notice, warning, and error messages. Information includes collection pathnames, indexing status, language modules installed, and restart warnings.
### File log level

Specifies the minimum log severity required to record log messages to Ultraceek’s error.log file.

You can set the following file log levels:

- **error**—Log only Ultraceek errors. This log level allows you to maintain the cleanest log.
- **warning**—Log both warning and error messages. This log level allows you to maintain a more detailed log.
- **notice**—Log all notification, warning, and error messages.
- **info**—Log all messages generated by Ultraceek, including informational, notice, warning, and error messages. This is the default setting.

### Number of old access, click, and query logs to keep

Controls when Ultraceek deletes access, click, and query log files. By default, Ultraceek keeps 400 sets of these files. Since Ultraceek relies on these log files to produce reports, it is recommended that you retain a large number of them.

### Number of old error log files to keep

Controls when Ultraceek deletes old error log files. By default, Ultraceek keeps seven sets of old log files.

The default setting is generated by Ultraceek, based on the amount of physical disk space available on your system. This value is likely to provide the best performance.

### Termlist cache size

Specifies how much main memory to devote to Ultraceek’s search engine termlist cache, in megabytes. The minimum query cache size is dictated by the maximum termlist length. If you have a large collection and are processing many queries, you might want to choose a larger cache size than the default value.
### Maximum termlist length

Specifies the maximum length of any termlist in an index file, in terms of number of documents. In order to keep its index file size and query cache size limited, Ultraseek keeps track of a limited number of documents per term in each of its index files. This parameter specifies the maximum number of documents per term to keep track of. The maximum allowed value for this parameter is 2097151 (2 megabyte documents). The default setting is relative to the number of documents in your license. Also, the maximum termlist length will not be greater than 131071 (128k documents) unless you change the setting in this field.

If a termlist becomes longer than the value set for this parameter as the result of an index merge operation, Ultraseek truncates the termlist to this length, and references to the least relevant documents are dropped.

Smaller maximum termlist lengths allow the use of smaller query caches and result in smaller index files. Larger maximum termlist length values ensure the correct operation of the include (+) and exclude (-) query operators on popular terms. The maximum termlist length value does not need to be larger than the total number of documents indexed, since no termlist would ever contain more documents than all indexed documents.

### Title cache size

Specifies how much main memory to devote to Ultraseek's search engine title cache. This cache holds the per-hit title information displayed on search results pages, including each URL, title, summary, publisher, and date.

This cache is useful for improving the performance of queries against large collections. Each document title record takes approximately 448 bytes. A size of 0 disables the title cache.
### Maximum storage heap size

Indicates the amount of memory to allocate in megabytes, to prevent Ultraseek's virtual image size from growing too large.

If `malloc` asks for more than the designated amount of memory, Ultraseek restarts itself. Ultraseek's total memory usage depends on the heap size specified by this parameter, the query cache size (see “Termlist cache size” on page 161), each collection's in-memory index (see “In-memory index size (documents)” on page 103), and the total number of server and indexer threads (one megabyte per thread, see “Number of server threads” on page 160).

If you experience frequent restarts of Ultraseek, consider adjusting this parameter. Generally, you can set the maximum storage heap size in increments of 50% of the initial setting. For example, if you have heap size set to 512 megabytes, the next increase would be 768 megabytes and so on.

The default setting is generated by Ultraseek, based on the amount of physical memory in your system. This value is likely to provide the best performance.

### Filter number of hits

Specifies the number of most relevant query hits to feed into the results filter for processing. The default is 100.

### Average group size

Specifies the average number of search hits to cluster in a group before returning search results to the user. If the actual group size is higher than the specified number, Ultraseek decreases group size until the number is reached or it cannot further subdivide the group.

This setting is used only when the Group by Location feature is enabled.

### Maximum group subhits

Specifies the maximum number of search hits shown per group. This setting is used only when the Group by Location feature is enabled.

### Group score threshold

Filters search hits displayed in a group. Ultraseek determines a high and low boundaries for each group: the top boundary is the document score of the top search hit in the group, and the lower boundary is the top boundary minus the Group score threshold. Documents are included in the group only if they are between these two numbers.

This setting is used only when the Group by Location feature is enabled.

### In-document highlighting filter

Specifies which URLs can be highlighted. By default, documents from all URLs can be highlighted. Ultraseek examines the list of URL patterns in the order they are specified. The first matched URL pattern determines whether in-documenting highlighting is allowed for documents originating from the specified URL. If there is no match, Ultraseek disallows highlighting on documents from the URL.
## In-document highlighting proxy server
Specifies which proxy server to use for document highlighting. This can be the same proxy server used for spidering purposes and specified on the Collections | Network page.

## Recreate log digests
Click the **recreate** button to rescan the query logs and rebuild the reporting digests.

*Note:* This operation takes a while to complete, so it should only be performed during periods of low search activity.

## Collect clickthrough information
Select this setting if you want to collect clickthrough information for the reporting feature. Clickthroughs count the number of links that users click while browsing results in the search interface.

## Allow HTTP keep-alives
Select this setting if you want Ultraseek’s built-in HTTP server to allow clients to keep the HTTP connection open after making a request to the server. This reduces the TCP connection setup time for client queries, but at the cost of tying up a server thread while the connection is kept open. We recommend you select this setting, since most page leads require several HTTP requests to load the HTML and any associated images.

If you clear this setting, you can decrease the number of server threads. Connections are closed if they remain idle for more than the set time. The default is 60 seconds. For more information on keep-alives see “HTTP keep-alive timeout” on page 160.

## Resolve client addresses for log
When this parameter is selected, Ultraseek attempts to find a host name for each client IP address using DNS lookups and includes the host name in the access log file. If this parameter is not selected, Ultraseek includes the client’s IP address in the access log file and does not attempt to resolve the client IP address to the host name.

Resolving client addresses can decrease performance, since many IP addresses do not resolve to host names. Instead, deselect this setting and use a log post-processor tool. For example, if you are running an Apache HTTP server to analyze log files, you should use their logresolve tool available from http://www.apache.org

## Log requests for images
Enables the logging of successful HTTP requests for images in the access log file. Typically, requests for images are not logged in order to keep the access log file smaller.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
</table>
| **Use Stemming**                             | Causes stem forms of words to be used for indexing and query terms. When stemming is used:  
  - A query for “goose” finds either “goose” or “geese”.  
  - A query for “computers” finds “computer” or “computers”.  
  - A query for “run” finds “run”, “ran”, or “running”.  
  This parameter is checked by default. When **Use Stemming** is changed, all indexes must be cleared and rebuilt. Stemming for Japanese, Korean, and Chinese is always enabled. We recommend keeping this parameter on to improve relevance rankings. |
| **Debug mode**                               | Enables more verbose error reporting.                                                                                                                                                                     |
| **Parse binary data in database tables**     | Causes a database collection to attempt to parse binary data (BLOBs) discovered in a table.                                                                                                                                                                      |
| **Remove from results duplicate hits with the same URL** | Enables the automatic removal of duplicate hits with the same URL from the results list during query time. Normally enabled, this feature is useful when users search multiple collections and there are duplicate URLs between the collections. |
| **Allow Wildcard queries**                   | Allows users to use wildcard characters in their queries. Ultraseek uses wildcards to expand the possible matches for a query.                                                                                                                                 |
| **Interleave Sites**                         | Specifies how results from the same site should be listed on the search results page. When selected, search results that have the same score are sorted so that hits from the same site do not appear in groups. Instead, results with the same score, but from different sites, are interleaved. |
| **Filter results based on user credentials** | Authenticates all search results before displaying them to users. Checking this parameter causes Ultraseek to display only search results to which users have access. In addition, Ultraseek places a lock icon next to each search hit that requires the user to log in before the document may be viewed. Enabling this feature may decrease query performance.  
This feature supports HTTP and HTTPS URLs. Use XPA to provide this level of authentication for other URLs. See the *Verity Ultraseek XPA Programming Guide* for more information.  
You can customize your authentication by modifying the *new_check_auth* wrapper function in *patches.py*. See the *Verity Ultraseek Customization Guide* for more information.  
You can also require users to log in before performing searches by checking the “Require login for result filtering” checkbox on the **Interface** | **Query** page. |
Managing Administrative Users and Client Access

You can provide password-based access control to administrative functions and manage client access to Ultraseek. All of the access rights should be granted by the administrator (superuser) of Ultraseek.

Creating Administrative User Accounts

You can grant individual users administrative access to specific features in the Administrative Interface.

To create administrative user accounts:
1. Click Server.
2. Click the Users tab.
3. Click new.
   The New Administrative User page displays.
4. Enter the username and password the user will use to access the Administrative Interface.
5. Select the features that the user can access with administrative rights. You can select individual collections, all collections, server access, and topic access.

Deleting Administrative User Accounts

To delete administrative user accounts:
1. Click Server.
2. Click the Users tab.
3. Click delete next to the username that you want to remove.
Changing Passwords

Administrative users can change their passwords using the password feature on the Users tab. If a user has forgotten their password, you can also delete the old user ID and create a new user.

To change your password:

1. Click Server.
2. Click the Users tab.
3. Click the password link next to the username. The Change Password page displays.
4. Enter the old password in the Old Password field.
5. Enter the new password in the New Password field.
6. Re–enter the new password in the Confirm Password field.
7. Click ok.

Additional Password Security

You can manage passwords by defining the password length, password longevity, password uniqueness, and number of failed administrator logins by modifying the patches.py file in the Ultras eek program directory. See the Verity Ultraseek Customization Guide for more information about modifying the patches.py file.

Note If you modify password security values in the patches.py file, the new configuration values will override the internal Ultraseek default values. To return to the default values, either use the code snippets below to explicitly set the parameters back to their default values or manually remove the entries from the configuration file when Ultraseek is not running. You cannot modify these values using the Administrative Interface.
Password Length

The minimum password length is managed by the configuration variable `password_min_length`. The default length is 5 characters. You can disable this feature by setting the `password_min_length` variable to 0.

If you create a new user on the New Administrative User page with a password shorter than the length specified in patches.py, Ultraseek displays a message indicating the minimum password length.

To set the minimum password length to 7 characters, for example, add this code block to the end of patches.py:

```python
#--------------
import config
import log

config.password_min_length = 7
if config.password_min_length:
    log.log(log.info,'password minimum length set to %d characters' %
            config.password_min_length)
else:
    log.log(log.info,'password minimum length not enforced')
#--------------
```

Password Longevity

You can set passwords to expire after a specific length of time by setting the `password_expire_interval` variable. The default interval is 30 days. You can disable this feature by setting the `password_expire_interval` variable to 0 days.

To set the password lifetime interval to 90 days, for example, add this code block to the end of patches.py:

```python
#--------------
import config
import log

config.password_expire_interval = 90
if config.password_expire_interval:
    log.log(log.info,'password lifetime set to %d days' %
            config.password_expire_interval)
else:
    log.log(log.info,'passwords never expire')
#--------------
```
Password Uniqueness

Ultraceek requires that users enter a unique password when changing existing login information. You can override this setting to allow users to reuse existing passwords by setting the `unique_pwd` variable in the configuration file.

If an administrative user tries to keep their same password by submitting the same value for both the old password and new password fields, Ultraceek displays a message indicating that a unique password needs to be entered.

The default value is enabled with a setting of `true (1)`. You can disable this feature by setting the `password_require_change` variable to `false (0)`.

To allow users to reuse existing passwords, add this code block to the end of `patches.py`:

```python
#--------------
import config
import log

config.password_require_change = 0
if config.password_require_change:
    log.log(log.info, 'password change enforcement enabled')
else:
    log.log(log.info, 'passwords change enforcement disabled')
#--------------
```

Failed Administrator Logins

After a sequence of failed login attempts to the Administrative Interface, the IP address of the user is forbidden from accessing Ultraceek and the user will receive a “403 Forbidden” error.

The user will be locked out for the designated lockout period or until the next server restart, whichever occurs first. Ultraceek records failed login attempts in the error logs and also sends an error message to the administrator email address.

This feature is managed by two configuration variables:

- `password_max_error_retries (# of attempts)—the number of failed attempts before IP lockout. The default value is 5.
- `password_IP_lockout_interval (# of hours)—the duration of IP lockout after hitting the threshold. The default value is 24 hours.

To disable this feature, set the `password_max_error_retries` value to zero (0).
To set a more strict threshold of failed login attempts of 20 attempts and a longer lockout interval of 72 hours, for example, add this code block to the end of patches.py:

```python
#--------------
import config
import log

cfg.password_max_error_retries = 20
cfg.password_IP_lockout_interval = 72
if cfg.password_max_error_retries:
    log.log(log.info,
            'after %d failed login attempts, clients will be locked out for %d hours' %
            (cfg.password_max_error_retries, cfg.password_IP_lockout_interval))
else:
    log.log(log.info,'no limit to failed login attempts')
#--------------
```

Managing Client Access

In addition to password protection, you can restrict access to the Administrative Interface by specifying numeric Internet Address Patterns (IP) using the Address-Based Access Level Specification feature on the Server|Users tab.

By default, certain administrative functions can be performed through the help interface (see “Accessing Application Help” on page 32). To see these functions, look under the More Services heading in the navigational column that runs down the right side of the help interface. For example, users can use the Add URL, Delete URL, and Revisit site features without having administrative privileges. This is useful in intranets or other environments where users can contribute to the search index. However, it is sometimes necessary to hide these functions from users. To do this, you set values in the Address-Based Access Level Specification region.

The address-based access level specification form provides levels of access to clients based on the numeric Internet address of the client machines. The web server automatically hides links to functionality for which access is not allowed.

Each pattern and its corresponding action tells Ultraseek’s built-in web server what kind of access level to provide to a particular client. The web server compares the numeric Internet address of the client with each pattern. If the address matches the pattern, the web server provides the level of access. It either allows administrative access with a
password, allows mirror access, allows add url access, allows search and help access, or disallows any access at all. If no match is found, administrative access is allowed with a password.

**To specify client access:**

1. Click **Server**.
2. Click the **User** tab.
3. Specify values in the Address-Based Access Level Specification region.

   The following values describe the options.

   **Syntax**

   Choose wildcard patterns or specific IP addresses. “Wildcard” is the default. Choose “regex” (regular expression), or you can use specific IP addresses. The table below shows some examples of wildcards.

<table>
<thead>
<tr>
<th><strong>Client Numeric IP Address Pattern</strong></th>
<th><strong>Access Level</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter an IP address or wildcard pattern.</td>
<td>Choose the client search access:</td>
</tr>
<tr>
<td></td>
<td>■ allow administrative access with a password</td>
</tr>
<tr>
<td></td>
<td>■ allow mirror access</td>
</tr>
<tr>
<td></td>
<td>■ allow addurl access</td>
</tr>
<tr>
<td></td>
<td>■ allow search access</td>
</tr>
<tr>
<td></td>
<td>■ disallow any access.</td>
</tr>
</tbody>
</table>

   For example, suppose your company administrators share similar IP prefixes that start with 123.456.78.XX. The form entries would look something like this:

<table>
<thead>
<tr>
<th>Syntax</th>
<th>Client Numeric IP Address Pattern</th>
<th>Access Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>wildcard</td>
<td>123.456.78.*</td>
<td>allow admin</td>
</tr>
<tr>
<td>wildcard</td>
<td>*</td>
<td>allow search</td>
</tr>
</tbody>
</table>

   This will give every computer with an IP address that matches 123.456.78.* the capability for full administrative access, but all other computers will be limited to searches only.

   The IP addresses need to be listed from most restrictive to least restrictive. In other words, exact IP addresses come first, then regex IP addresses, and then wildcard IP addresses. The type of access does not matter as much as the order of the IPs.

   If on a different subnet, change the admin pattern to match your subnet.
Because the default access for an unmatched host is “allow admin”, the final “allow search” pattern must be specified, or access will not be limited.

In the help interface, links under the More Services heading will only be visible to administrators who have a qualifying IP.

Allowing HTTPS Administrative Access

If you have a license key that enables the SSL security module and have installed the SSL package, you can provide secure HTTPS access to Ultraseek. Contact software-sales@verity.com for more information about obtaining this module.

To access the server only through HTTPS:

1. If necessary, specify an HTTPS port number in the HTTPS server port field on the Server | Parameters | Main page.
2. Click OK.
   
   Ultraseek will restart.

Managing Document Types

You can manage document types to define mappings between file extensions and content types. You can also specify how Ultraseek parses documents, by adding entries to the document parsing table.

Using the document type management features, you can:

- View the supported and unsupported MIME document types (see “Document Type Specification” on page 173).
- Change the list of MIME types that the indexer associates with each file extension to reflect your network conventions (see “Document Type Specification” on page 173).

For more information on how to parse documents based on content type see “Document Type Parsing” on page 177.

To manage document types:

1. Click Server.
2. Click the Doc types tab.
Document Type Specification

Ultraseek determines each document’s content type based on the file extension for each file. You use the Document Type Specification section to define how various file extensions are mapped to available content types. If a mapping does not exist for a particular file extension, Ultraseek uses the content type `x-unknown-content-type`.

- The web server uses the mapping to determine the content of files stored in the `docs` directory.
- The file system scanning collection uses the mapping to determine the content of the files it scans.
- The spider uses the mapping to determine the appropriate protocol, such as FTP or HTTP, when retrieving files and to identify files that cannot be indexed. The spider also uses the mapping, combined with the document type parsing table, to automatically disallow HTTP URLs that appear as if they cannot be indexed. For example, a URL ending in “.tar” is automatically disallowed by the spider because the file extension “.tar” is matched to “application/x-tar”, which does not have an assignment in the document type parsing table.

Figure 7-3 Doc Types Page
Supported MIME Types

MIME types indicate how Ultraseek classifies files. Documents with supported MIME types are indexed by Ultraseek. Documents with unsupported MIME types are not indexed, preventing unnecessary load on the server. Table 7-1 lists common MIME types that Ultraseek supports.

The addition of MIME types to the configuration of your web server will affect how you compose searches. For example, if your web server lists Excel files under a MIME type application/x-msexcel, you can search these files by typing doctype:msexcel or doctype:x-msexcel.

XML documents must be well-formed (lexically correct) but they do not need to be valid (consistent with their Document Type Declaration).

Table 7-1  Supported MIME Types

<table>
<thead>
<tr>
<th>Application</th>
<th>Mime Type (Document Type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe Acrobat</td>
<td>application/pdf</td>
</tr>
<tr>
<td>Autocad Interchange</td>
<td>application/x-dxf</td>
</tr>
<tr>
<td>Corel WordPerfect</td>
<td>application/wordperfect5.1</td>
</tr>
<tr>
<td>HTML</td>
<td>text/html</td>
</tr>
<tr>
<td>Ichitaro</td>
<td>application/x-js-taro</td>
</tr>
<tr>
<td>Lotus 1-2-3</td>
<td>application/vnd.lotus-1-2-3</td>
</tr>
<tr>
<td>Lotus Freelance</td>
<td>application/vnd.lotus-freelance</td>
</tr>
<tr>
<td>Lotus WordPro (Windows NT only)</td>
<td>application/vnd.lotus-wordpro</td>
</tr>
<tr>
<td>Microsoft Excel</td>
<td>application/vnd.ms-excel</td>
</tr>
<tr>
<td></td>
<td>application/x-msexcel</td>
</tr>
<tr>
<td>Microsoft Powerpoint</td>
<td>application/vnd.ms-powerpoint</td>
</tr>
<tr>
<td></td>
<td>application/x-mspowerpoint</td>
</tr>
<tr>
<td>Microsoft Project</td>
<td>application/vnd.ms-project</td>
</tr>
<tr>
<td>Microsoft Word</td>
<td>application/msword</td>
</tr>
</tbody>
</table>
## Application Mime Type (Document Type)

<table>
<thead>
<tr>
<th>Application</th>
<th>Mime Type (Document Type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multipart</td>
<td>multipart/alternative</td>
</tr>
<tr>
<td></td>
<td>multipart/digest</td>
</tr>
<tr>
<td></td>
<td>multipart/mixed</td>
</tr>
<tr>
<td></td>
<td>multipart/parallel</td>
</tr>
<tr>
<td></td>
<td>multipart/related</td>
</tr>
<tr>
<td></td>
<td>multipart/signed</td>
</tr>
<tr>
<td>OpenOffice 1.0, 1.1</td>
<td>application/vnd.sun.xml.calc</td>
</tr>
<tr>
<td>StarOffice 6.0, 7.0</td>
<td>application/vnd.sun.xml.calc.template</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.draw</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.draw.template</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.impress</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.impress.template</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.math</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.writer</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.writer.global</td>
</tr>
<tr>
<td></td>
<td>application/vnd.sun.xml.writer.template</td>
</tr>
<tr>
<td>Plain Text</td>
<td>text/plain</td>
</tr>
<tr>
<td>RFC822</td>
<td>message/rfc822</td>
</tr>
<tr>
<td>Rich Text</td>
<td>application/rtf</td>
</tr>
<tr>
<td>Visio</td>
<td>application/vnd.visio</td>
</tr>
<tr>
<td>XML</td>
<td>application/vnd.wap.xml</td>
</tr>
<tr>
<td></td>
<td>application/xml</td>
</tr>
<tr>
<td></td>
<td>text/vnd.wap.wml</td>
</tr>
<tr>
<td></td>
<td>text/xml</td>
</tr>
<tr>
<td>Zip</td>
<td>application/zip</td>
</tr>
</tbody>
</table>
Supported Highlight Types

Ultraseek can highlight query terms in any document with a source URL beginning with “http://” or “https://”, except in database, mirror, and merged collections. If a document returned from a search is highlightable, Ultraseek inserts a “Highlight” link directly beneath the “Find Similar” link on the search results page.

For certain document types, a “View as HTML” link appears when results return from searches where there are no terms to highlight. For example, fielded searches or “Find Similar” searches. When you click “View as HTML”, Ultraseek converts the result document to HTML and displays it in your browser.

**Note** Highlighting in Chinese language documents is not supported.

Ultraseek supports highlighting search terms in the following document types:

- Microsoft Word
- Microsoft Excel
- Microsoft PowerPoint
- Adobe Acrobat
- Corel WordPerfect
- Lotus 1-2-3
- Lotus WordPro (Windows NT, Windows 2000, and Windows 2003 only)
- Lotus Freelance
- Rich Text Format
- XML
- HTML
- .txt

**Note** PDF results can be converted to HTML before highlighting. If you want to highlight PDF results as HTML, for the application/pdf document type, set Parse as to Adobe Acrobat (Key View) on the Server > Doc types pane. If you want to continue to highlight PDF results within the Acrobat Reader, set Parse as to Adobe Acrobat.
Document Type Parsing

The Document Type Parsing region, shown in Figure 7-3, specifies how the indexer should parse documents, based on content type. You can use this region to configure Ultraseek to recognize new or additional document types.

To add a document type to the parsing table:

1. Enter the MIME type of the document in the Document Type field.
2. Select an entry in the Parse as field to specify which built-in parsing procedure the indexer should use for the document type.

Creating XML Mappings

Element names can be mapped to field names which can be searched across. This allows searches to be focused on a specific element within an XML document. You can set mappings for particular Document Type Definitions (DTDs), which specify the syntax of a web page in SGML, the Standard Generalized Markup Language on which HTML is based.

DTDs are identified by the root element. There is a generic mapping for documents which do not match a specific mapping. Whenever you apply a new mapping, you will need to re-index the document.

XML documents using the namespace mechanism can be parsed, and elements with namespace qualifiers can be mapped for fielded search. Ultraseek can associate the text inside XML elements with searchable fields, so a search for title:Hamlet will find an XML document containing the play *The Tragedy of Hamlet, Prince of Denmark* if the document’s title: field has been populated.

You can search values of XML attributes. To map an attribute to a field name for search, enter an ampersand (&) in the last field of the mapping table. To map the title attribute in this element to the "title:" field:

```
<article title="Content in Attributes">
<x xmlns:title="The Tragedy of Hamlet, Prince of Denmark"/>
```

Use this mapping in the XML mappings page of the Server section of the Administrative Interface:

```
FieldnameElementAttr. NameAttr.Value
titlearticletitle&
```
The string “Content in Attributes” will be mapped to the title of the document.

To edit existing XML mappings or create new mappings:

1. Click the Server tab.
2. Click the XML Mappings tab.
3. Click edit next to the predefined mapping that you want to edit, or click new to create new element-to-fieldname mappings.

**XML Generic Mappings**

**Figure 7-4  XML Generic Mappings**

The Generic Mappings are a set of commonly-used XML elements along with their most common meanings. If an XML document doesn’t match a specific mapping, the Generic Mapping is used.

The most important fields to map are title:, which is used for the title in search results pages, and description:, which is used for the summary in those pages. Providing mappings to these two tags will greatly improve the search results.
Document-Specific Mappings

Mappings are selected according to the root element of the XML file. The root element is the element which is not inside any other element. In the example below, `<PLAY>` is the root element.

```xml
<?xml version="1.0"?>
<PLAY>
<TITLE>The Tragedy of Hamlet, Prince of Denmark</TITLE>
<fm>
<p>Text placed in the public domain by Moby Lexical Tools, 1992.</p>
<p>This work may be freely copied and distributed worldwide.</p>
</fm>
<Personae>
<TITLE>Dramatis Personae</TITLE>
<Persona>CLAUDIUS, king of Denmark.</Persona>
<Persona>HAMLET, son to the late, and nephew to the present king.</Persona>
<Persona>POLONIUS, lord chamberlain.</Persona>
<Persona>HORATIO, friend to Hamlet.</Persona>
</Personae>
<!-- rest of characters and play deleted -->
</PLAY>
```

For this example, we map the title, the description, and the list of characters (contained in `<PERSONAE>` elements). This document does not have a summary, but the `<fm>` element describes the rights and provenance, which is useful information. We do not map the `<p>` element to anything, since it is not an interesting aspect of the text. Unlike in HTML, case is important in XML elements, so we need to make `<PLAY>` and `<fm>` upper case and lower case, respectively. If you have used the wrong case, your syntax errors show up in the log files.

This improves the information on the results pages, and will allow searches that are limited to information about characters in the play. With all of Shakespeare's plays in this format, a search for `personae:nephew` would find all plays with characters described as someone's nephew.
Specifying Attributes

If an element’s role is determined by an attribute, set the attribute name and value fields. For example, in documents using the *Text Encoding Initiative* (TEI) standard, many kinds of text divisions use the `<div1>` element, but abstracts (descriptions) should have a type attribute with the value abstract:

```
<div1 type="abstract">A summary.</div1>
```

*Figure 7-5  XML Attribute Mapping for TEI documents*

Element-to-fieldname mapping happens when a document is retrieved and indexed. When mappings are changed, the documents need to be re-indexed. To re-index documents, use the **Delete URL** tab on the collection status page, but make sure the **Delete from URL database?** button is not checked.

Each XML document is indexed with a `rootelement:field` and with a MIME type of `text/xml` or `application/xml`. A search for `doctype:xml` will match all XML documents. A search for `rootelement:PLAY` will match just XML documents with a root element of `<PLAY>`. That search can be used with the Delete URL command to re-index all documents affected by a mapping change.

Conflicting Root Elements

You may have more than one kind of XML document that uses the same root element. To handle this, work out the mappings for each DTD, then put them together under the same root element. As each one of the documents are indexed, the appropriate mappings are available.
Creating Quick Links

A Quick Link is a URL that appears at the top of the results page after a user searches for certain predefined keywords or phrases. Each Quick Link requires the keywords or phrases that need to be matched and the URL that will be returned by the search. If you define multiple keywords for a particular URL, only one keyword needs to match the search criteria to serve the URL in the results. For example, if you associate three queries "stock market", "funds", and "stock symbols" with a URL, a search on "funds" only will return the URL at the top of the results page.

You can specify Custom Titles or Summaries for URLs, and also choose which collections to include when searching for Quick Link keywords.

An example of the search results for a Quick Link defined for the keyword “java” and associated with the URL http://developer.java.sun.com/developer/infodocs/ and the hyperlink text “Java Documentation” is shown below.

Figure 7-6 Sample Quick Link
To create Quick Links:
1. Click the Server button.
2. Click the Quick Links tab.
3. If you have multi-language support, select the language that you want to edit.
4. Click new.

The Quick Links page displays.
5. In the Keyword field, enter the keywords or phrases that you want to associate with a particular URL. Separate multiple entries with commas. For example: stock market, funds, stock symbols.
6. Enter the URL that should display at the top of the result page when a keyword is matched in the URL field. This should be a fully formed URL, for example:

   http://www.verity.com/index.html

7. Select a Title option from the following table.

<table>
<thead>
<tr>
<th>Automatic Title</th>
<th>Ultraseek tries to look for the title of the URL in the search index and to apply that title to the Quick Link. If the title tag is not populated in the URL, then no title will appear.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom Title</td>
<td>Allows you to specify the title of the Quick Link. If you choose this option, enter the title in the text field.</td>
</tr>
</tbody>
</table>

8. Select a Summary option:

<table>
<thead>
<tr>
<th>Automatic Summary</th>
<th>Ultraseek tries to look up the summary of the URL in the search index and to apply that summary to the Quick Link. If the summary tag is not populated, then no summary will appear.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom Summary</td>
<td>Allows you to specify the summary of the Quick Link. If you choose this option, enter the summary in the text field.</td>
</tr>
</tbody>
</table>

9. Select the collections that will enable this Quick Link. The URL will only appear if a user searches a collection that is checked here.

10. After completing the fields, click ok. The new Quick Link with the corresponding URL of the site will appear on the Quick Links page.
Secure Spidering

This chapter describes the types of authentication that Ultraseek uses, how to manage your digital certificates and keys.

- Authentication
- Managing Digital Certificates
- Managing Keys
Authentication

Ultraseek supports the following types of authentication when spidering sites:

- Basic + SSL (with or without X.509 client certificates)
- NTLM + SSL (with or without X.509 client certificates)
- Form-based

SSL support is provided in a separate module and is not packaged with Ultraseek package. Please contact your sales representative for more information about obtaining the SSL module.

**Note** SSL spidering does not apply to database, scanner, mirror, and merged collections. For more information about basic authentication with SSL, see “Authentication” on page 93.

Figure 8-1 illustrates how Ultraseek spiders protected content:

1. The spider tries to access the protected resource, such as a web server.
2. The spider is challenged by the web server; the spider then passes credentials to the web server.
3. The web agent passes credentials to the policy server.
4. The spider is authenticated and authorized to access the protected resource.

After the spider is authenticated, the documents are crawled and indexed.
Table 8-1  Secure Spidering Terms

**Certification Authority**

A certification authority (CA) is an authority trusted by one or more users to create and assign certificates. Optionally, the certification authority may create the user's keys. Certificate users depend on the validity of information provided by a certificate. A certification authority should be someone that certificate users trust, and usually holds an official position created and granted power by a government, a corporation, or some other organization.

**Key Pair**

A key pair is a set of related keys—a public key and a private key—that are used for asymmetric cryptography and are generated in a way that makes it infeasible to derive the private key from knowledge of the public key. A key pair's owner discloses the public key to other system entities so they can use the key to encrypt data, verify a digital signature, compute a protected checksum, or generate a key in a key agreement algorithm. The matching private key is kept secret by the owner, who uses it to decrypt data, generate a digital signature, verify a protected checksum, or generate a key in a key agreement algorithm.
<table>
<thead>
<tr>
<th><strong>Table 8-1</strong> Secure Spidering Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distinguished Name</strong></td>
</tr>
</tbody>
</table>
| **Common Name**  | A key pair’s or certificate’s *Common Name* is a character string that meets the following criteria:  
  - May be a part of the X.500 Distinguished Name of a Directory object (“commonName” attribute).  
  - Is a (possibly ambiguous) name by which the object is commonly known in some limited scope (such as an organization).  
  - Conforms to the naming conventions of the country or culture with which it is associated.  
Examples of Common Names are “Dr. E. F. Moore”, “The United Nations”, and “Human Resources Department”. |
| **Mandatory Access Control**  | A *Mandatory Access Control (MAC)* is an access control service that compares the following:  
  - **Security labels** (which indicate how sensitive or critical system resources are)  
  with  
  - **Security clearances** (which indicate the system entities that are eligible to access certain resources)  
This kind of access control is called “mandatory” because an entity that has clearance to access a resource may not, just by its own volition, enable another entity to access that resource. |
Managing Digital Certificates

This section describes how to manage digital certificates using the keyfm and mykeys.kdb files.

- “Using keyfm” on page 188
- “Adding a Trusted Root to the mykeys.kdb File” on page 190
- “Adding a Client Certificate” on page 192
- “Changing the Password” on page 193
- “Removing a Key or Certificate” on page 194

The mykeys.kdb file stores all key, certificate, and certification authority information for SSL communications. On Windows NT and Windows 2000, the file is located in the following directory: C:\Program Files\Verity\Ultraseek\lib. On Linux and Solaris, the file is located in /var/opt/Ultraseek/lib.

By default, the mykeys.kdb file is password protected by the password “password.” Instructions for changing this password are described below.

The Ultraseek web spider uses the list of Trusted Root Certificate Authorities found in mykeys.kdb when connecting to SSL-protected (HTTPS) web sites. Protected web sites must present a SSL Certificate issued by a trusted certification authority. Instructions for adding trusted roots to mykeys.kdb are described below.

When connecting to a remote web site using SSL (HTTPS), the spider chooses a key-client certificate pair from the mykeys.kdb file. If the mykeys.kdb file has only one key, the spider will automatically choose that key. If mykeys.kdb has more then one key, the spider does the following to choose a key:

1. The spider examines every key in mykeys.kdb to determine which are certified.
2. For every certified key, the spider examines the list of key/value pairs in the certificate's Distinguished Name.
3. If the value part of the key/value pair matches the authentication userid for any key/value pair in the certificate, that key/certificate will be used for client authentication.
4. If no match is found, the spider will randomly choose a key from the mykeys.kdb file.

Use the Collections | Network page to edit the userid in the Authentication region. Instructions for adding and removing keys and certificates from the mykeys.kdb file are provided in the sections that follow.
Using keyfm

`keyfm` demonstrates basic key-management tasks.

```
keyfm -[kszcDEItuTp][b][A][d][f][r][X][l][v][23] [-i infile]
[-o outfile] [-K keyfile] [-P password] [common_name | -N dn]
```

Several `keyfm` commands operate on a key pair or trusted root key in the keyfile. You can specify the choice of key at the command line by specifying either the Common Name or the full Distinguished Name of the desired key. If neither of these is included, `keyfm` lists the Distinguished Name of all relevant keys and prompts you to choose one.

**IMPORTANT** The common name for a key should be the hostname of the computer that search is going to be accessed on (the one entered in the URL). Otherwise you will get an SSL error from the browser saying the key does not match.

### Options (Functions)

The following mutually exclusive options specify the function to be performed.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>Generate a new RSA key pair and, optionally, a certificate request for it.</td>
</tr>
<tr>
<td>s</td>
<td>Create a certificate signing request for an existing key pair.</td>
</tr>
<tr>
<td>z</td>
<td>Delete a key pair.</td>
</tr>
<tr>
<td>c</td>
<td>Receive a certificate for a key in the keyfile.</td>
</tr>
<tr>
<td>D</td>
<td>Display the Distinguished Names and, optionally, the certificates, for key pairs in the keyfile.</td>
</tr>
<tr>
<td>E</td>
<td>Export (to the output stream) a key pair and its certificate chain (if present) in PKCS#12 format.</td>
</tr>
<tr>
<td>I</td>
<td>Import key pairs or certificates from a PKCS#12 file.</td>
</tr>
<tr>
<td>t</td>
<td>Add a trusted root by reading its X.509 certificate.</td>
</tr>
<tr>
<td>u</td>
<td>Delete a trusted root from the keyfile.</td>
</tr>
<tr>
<td>T</td>
<td>Display trusted roots and their certificates, when available.</td>
</tr>
<tr>
<td>p</td>
<td>Change the password protecting a keyfile.</td>
</tr>
</tbody>
</table>
Options (Modifiers)

The following options modify the behavior of the selected function. Some options may not be applicable to every function.

**Table 8-3  keyfm Modifiers**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Generate a DSA key—the default is RSA. DSA keys are relevant only for creating key pairs (controlled by the <code>-k</code> flag).</td>
</tr>
<tr>
<td>d</td>
<td>Generate a 1024-bit key—the default is 512 bits; 1024-bit keys are relevant only for creating key pairs (controlled by the <code>-k</code> flag).</td>
</tr>
<tr>
<td>f</td>
<td>Use fake random numbers, which are faster for testing; fake random numbers are relevant only for creating key pairs (controlled by the <code>-k</code> flag).</td>
</tr>
<tr>
<td>r</td>
<td>Create a certificate signing request and generate a new key pair; certificate signing requests are relevant only for creating key pairs (controlled by the <code>-k</code> flag).</td>
</tr>
<tr>
<td>b</td>
<td>Use Base64 encoding for outputting <code>csr</code> data; Base64 encoding is relevant for <code>-k</code> and <code>-s</code> flags.</td>
</tr>
<tr>
<td>X</td>
<td>Receive an X.509 certificate—the default is PKCS#7; X.509 certificates are relevant only for adding certificates (controlled by the <code>-c</code> flag).</td>
</tr>
<tr>
<td>l</td>
<td>Display the long form of certificates; the long form is relevant for <code>-T</code> and <code>-D</code> flags.</td>
</tr>
<tr>
<td>v</td>
<td>Produce verbose output on <code>stderr</code>.</td>
</tr>
<tr>
<td>2</td>
<td>Use a 40-bit RC2 (CA_RC2_CBC) cipher to encrypt the <code>keyfile</code> when it is created—the default is 56-bit DES, or CA_DES_CBC; 40-bit RC2 ciphers are useful for creating export <code>keyfiles</code>.</td>
</tr>
<tr>
<td>3</td>
<td>Use a Triple-DES (CA_DES_ECE3_CBC) cipher to encrypt the <code>keyfile</code> when it is created—the default is 56-bit DES, or CA_DES_CBC; Triple-DES ciphers are useful for creating domestic-only <code>keyfiles</code>.</td>
</tr>
<tr>
<td>i</td>
<td>Read input data from <code>infile</code>, rather than <code>stdin</code>.</td>
</tr>
<tr>
<td>o</td>
<td>Write output data to <code>outfile</code>, rather than <code>stdout</code>.</td>
</tr>
<tr>
<td>K</td>
<td>Use <code>keyfile</code> as the <code>keyfile</code>—the default is <code>mykeys.kdb</code>.</td>
</tr>
</tbody>
</table>
Adding a Trusted Root to the mykeys.kdb File

When you use the keyfm utility, it writes all of the root information to the mykeys.kdb file in the bin directory. The bin directory is not accessed by the Ultraseek server for authentication or for checking trusted roots. Ultraseek checks the mykeys.kdb in the lib and SearchData directories. The mykeys.kdb file is present in the bin directory to allow keyfm to write to a separate file so that in the event the mykeys.kdb file is damaged it does not affect Ultraseek. Therefore, in order for any changes to take affect to the mykeys.kdb file which Ultraseek uses, the mykeys.kdb file must be overwritten in the lib and SearchData directories.

WARNING! Make a backup copy of the original mykeys.kdb file and to save it in another location, so if it is necessary to start with a fresh mykeys.kdb file, you will have a clean one to utilize.

How to add trusted roots on Windows NT/2000:

1. Stop the Ultraseek server. On Windows NT/2000 machines Ultraseek is run as a service, so it needs to be stopped in the services window. Use the keyfm tool’s -t option to install a certificate file for the certification authority you want to add as a trusted root.

   The certificate file should be in the X.509 file format. For the default installation you will need to cd into the following directory:

   Program Files\Verity\Ultraseek\bin
2. Once you are in the bin directory, run:

   keyfm -K mykeys.kdb -t -i MyCa.cer

   Where MyCa.cer is the name of the certificate file.

3. Use the keyfm tool's -T option to ensure that the certificate was properly loaded as a trusted root.

   keyfm -K mykeys.kdb -T

4. If this is the last change you will be making to the mykeys.kdb file, copy it into the Program Files\Verity\Ultraseek\lib directory. When you are asked if you want to overwrite the existing file, click OK. If this is not the last change you will make, follow the steps in the following sections, then, when you have completed your tasks, copy the file over.

5. Restart the Ultraseek server.

How to add trusted roots on Linux/Solaris systems:

1. Stop the Ultraseek server (seekd).

2. Use the keyfm tool's -t option to install a certificate file for the certification authority you want to add as a trusted root.

   The certificate file should be in the X.509 file format. For the default installation you will need to cd into the following directory:

   /var/opt/Ultraseek/bin

3. Once you are in the bin directory, run:

   keyfm -K mykeys.kdb -t -i MyCa.cer

   Where MyCa.cer is the name of the certificate file.

4. Use the keyfm tool's -T option to ensure that the certificate was properly loaded as a trusted root.

   keyfm -K mykeys.kdb -T

5. If this is the last change you will be making to the mykeys.kdb file, copy it into the / var/opt/Ultraseek/lib directory. If this is not the last change you will make, follow the steps in the following sections, then, when you have completed your tasks, copy the file over.

6. Restart seekd.
Adding a Client Certificate

This section describes how to add a client certificate to your Ultraseek installation.

**How to add client certificates to Windows NT/2000:**

1. Stop the Ultraseek server.

2. Use the `keyfm` tool's `-I` option to import the PKCS#12 key with certificate. If you are not in the `bin` directory, cd into the following directory:
   
   `Program Files\Verity\Ultraseek\bin`

3. Once you are in the `bin` directory, run:
   
   `keyfm -K mykeys.kdb -I -i MyCert.pfx`

4. Enter a password to decrypt the PKCS#12 file.

   The password is the password that was entered when the certificate was imported into the Internet Explorer browser. Enter this password to decrypt the `.pfx` certificate. When the PKCS#12 Certificate file is decrypted, you can delete or move it.

5. Use the `keyfm` tool’s `-D` option to ensure the key was properly loaded.

   `keyfm -K mykeys.kdb -D`

6. If this is the last change you will be making to the `mykeys.kdb` file, copy it into the `Program Files\Verity\Ultraseek\lib` directory. When you are asked if you want to overwrite the existing file, click **OK**. If this is not the last change you will make, follow the steps in the following sections, then, when you have completed your tasks, copy the file over.

7. Restart the Ultraseek server.

**How to add client certificates to Linux/Solaris:**

1. Stop the Ultraseek server (`seekd`).

2. Use the `keyfm` tool’s `-I` option to import the PKCS#12 key with certificate. If you are not in the `bin` directory, cd into the following directory:

   `/var/opt/Ultraseek/bin`

3. Once you are in the `bin` directory, run:

   `keyfm -K mykeys.kdb -I -i MyCert.pfx`

4. Enter a password to decrypt the PKCS#12 file.

   The default password is “password.” When the PKCS#12 Certificate file is decrypted, you can delete or move it.
5. Use the `keyfm` tool's `-D` option to ensure the key was properly loaded.

   ```
   keyfm -K mykeys.kdb -D
   ```

6. Restart `seekd`.

### Changing the Password

This section describes how to change the password to your Ultraseek installation.

#### Changing the password in Windows NT/2000:

1. Stop the Ultraseek server.

2. Use the `keyfm` tool's `-p` option to change the password for the `mykeys.kdb` file.

   ```
   If you are not in the bin directory, cd into the following directory:
   Program Files\Verity\Ultraseek\bin
   ```

3. Once you are in the bin directory, run:

   ```
   keyfm -K mykeys.kdb -p <newpassword>
   ```

4. Edit the configuration file `\Program Files\Verity\SearchData\configuration` and change the value for `ssl_password` to the new `mykeys.kdb` password.

5. If this is the last change you will be making to the `mykeys.kdb` file, copy it into the `Program Files\Verity\Ultraseek\lib` directory. When you are asked if you want to overwrite the existing file, click OK. If this is not the last change you will make, follow the steps in the following sections, then, when you have completed your tasks, copy the file over.

6. Restart the Ultraseek server.

   When the Ultraseek server restarts, it automatically encrypts the SSL password in the `\Program Files\Verity\SearchData\configuration` file so that the password is not stored in plain text. When you use the `keyfm` tool after you have changed the password, you will need to use the `-p` option to specify the new `mykeys.kdb` password.

#### Changing the password in Linux/Solaris:

1. Stop the Ultraseek server (`seekd`).

2. Use the `keyfm` tool's `-p` option to change the password for the `mykeys.kdb` file.

   ```
   keyfm -K /var/opt/Ultraseek/mykeys.kdb -p <newpassword>
   ```

3. Edit the configuration file `/var/opt/Ultraseek/configuration` and change the value for `ssl_password` to the new `mykeys.kdb` password.
4. Restart seekd.

When seekd restarts, it automatically encrypts the SSL password in the /var/opt/Ultraceek/configuration file so that the password is not stored in plain text. When you use the keyfm tool after you have changed the password, you will need to use the -p option to specify the new mykeys.kdb password.

Removing a Key or Certificate

This section describes how to remove a key or client certificate from your Ultraceek installation.

Removing a key or certificate in Windows NT/2000:

1. Stop the Ultraceek server.
2. Use the keyfm tool’s -z option to remove a key. If you are not in the bin directory, cd into the following directory:
   
   Program Files\Verity\Ultraceek\bin

3. Once you are in the bin directory, run:
   
   keyfm -K mykeys.kdb -z

   A list of key pairs will be available depending upon how many key pairs have been generated.

4. Enter the number of the key you want to remove.

5. Use the keyfm tool’s -D option to make sure the key was properly removed by entering:
   
   keyfm -K mykeys.kdb -D

   A list of the remaining keys will be shown. If there was only one key then you will see a message stating that the mykeys.kdb file contains no key pairs.

6. If this is the last change you will be making to the mykeys.kdb file, copy it into the Program Files\Verity\Ultraceek\lib directory. When you are asked if you want to overwrite the existing file, click OK. If this is not the last change you will make, follow the steps in the following sections, then, when you have completed your tasks, copy the file over.

7. Restart the Ultraceek Server.

Removing a key or certificate in Linux/Solaris:

1. Stop the Ultraceek server (seekd).
Managing Keys

This section describes how to use the keyfm program included with Ultraseek to perform common key management tasks such as creating, deleting, and displaying public and private key pairs, as well as importing and exporting public and private key pairs and certificates.

Creating Public or Private Key Pairs and Installing Certificates

This example describes the steps required to generate and install a public or private key pair, create its certificate signing request, and install the corresponding certificate.

To create a key pair and install a certificate:

1. Create an RSA key pair using the keyfm demonstration application.

   **Command line:** keyfm -kvf "Common name"

   **Example:** keyfm -kvf "William Shakespeare"

   The example command line generates a new RSA public/private key pair and stores it in the keyfile. The k flag generates a key pair, the v flag specifies verbose output, and the f flag specifies use of a deliberately insecure but fast random number seeding function. The key pair in this example is an RSA key pair. (Use the -A flag to create a DSA key pair.)
To identify this key pair, the spider assigns it a *label*, which consists of the key’s Distinguished Name represented as an RFC 1485 string. The name is formed by adding the supplied Common Name (“William Shakespeare”) to a preset string containing the remaining fields of a sample Distinguished Name. The resulting label for the new key is:

```
CN=William Shakespeare,OU=Engineering,O=“Terisa Systems, Inc.”,
ST=California,C=US
```

Note that the new key pair is not assigned the Distinguished Name until a certificate signing request is generated for it. The `keyfm` code snippet for this command begins at “case MAKE_KEYPAIR:”.

2. Display the `keyfile` bye using the command `keyfm -D`, to confirm that the key pair has been entered in the `keyfile`.

Make sure the key pair’s Distinguished Name is followed by “(no csr)”. This tag indicates that a certificate signing request (csr) has not been generated for this key. The key is, therefore, also un-certified.

Generate a `csr` for the new key.

**Command line:** `keyfm -svb -o outfile "Common name"

**Example:** `keyfm -svb -o csrfile "William Shakespeare"

The `s` flag generates the certificate signing request, the `v` flag specifies verbose output, the `b` flag specifies Base64-encoded output, and the `Common name` selects the key pair for which the `csr` is generated. When this step has been successfully completed, the Distinguished Name contained in the label for the key pair is assigned as the key pair’s Distinguished Name, and the Base64-encoded PKCS #10 `csr` is written to `csrfile`. The `keyfm` code snippet for this command begins at “case CREATE_CSR:”.

Because the output is in Base64, the contents of `outfile` looks like the example below:

```
---BEGIN CERTIFICATE REQUEST---
MIIBlzCB2gIBADB1MQswCQYDVQQGEwJVUzETMBEGA1UECBMKQ2FsaWZvcm5p
YTEdMBsGA1UEChMUVGlyanNhIFN5c3RlbXMsIEluYy4xFDASBgNVBAsTC0Vu
Z2luZ3VyaW5nMRwwGgYDVQQDEwNXaWxsanFtIFNoYWtlc3BlYXJlMFwwDQYJ
KoZIhvcNQAEBQAdSwAwSAJBAkTrKv1GwABHo3F83NPNxCIqEdsM+R+uEwMd
y0JAg020cyD+3K13V3zuZueZF8vi/gZZHtB1NEx7ej/DR7UieRE8CAwEAaAA
MA0GCSqGSIb3DQEBAUAAOEAaTn8dvDty8U2EVFAgtBtX83wGlzOGLzlxD
Mh6jnXRXUtgadPUusQY6fMvnR46CR1eN1/CSpDCMcXIKra6cw==
---END CERTIFICATE REQUEST---
```

3. Send the `csr` to a certification authority by e-mail, online Web submission, or other exchange method.
Obtain the corresponding certificate and save it to a file. A test certification authority (ca) provided by Verisign can be found at https://digitalid.verisign.com/ TestGetCSR.html.

The certificate may look like the example shown below:

```
---BEGIN CERTIFICATE---
MIICGDCCAcIICF8yYQIhBz77AwDQYJKoZIi8EAkQRADAQ==
BgNVBAsTDV2lcm1TaWduLCBjbmMxRzBFBgoNVBAoTb3ZjYjOjckmibj0v
cMVw3NpdG9yeS9jZXN0Q1BTIE1uY29ycC4gQnkgUmVmLiBMaWFILiBMVEQuMUYw
RAYDVQQLzEBZV5CB3JiGmYyVAyXyZ2gXYV0aG9yaXp1ZCB0ZXJ0aWNsZS5j
IcMFzV5fYw5jZXMgKEMpVlMxOTk3MB4XDTk3MTEwNDIzNTk1MDowTDEakGAMhC
MVVMxZzARBgNVBAoFMjAwMDAwMFoXDTk0MDcyNTk3MDoxMTAwNDAwNjQw
BAoUUFFRlcmlzYSBTTEx0MDk5NjAwMDAwMFowcTAklmdmFsdWVzdXJl
MBoGA1UEAxQTV2lsbGlhbSBTaGFrZXNwZWFyeSByb290aW9ucyBh
MEQgCQQChUK5RsAAR6NxfNzZcQ1KhHBdpfrhMDHctCQIKINjNMG/ttytd1d8?mbn
mRfL4v4GWR7NTRMe3o/w0e1IKRPaGMBAAEwDQYJKoZIhvcNAQEEBQADQGQA9A6J
BerAsi4pNBt+iAPzLrr5Qd1xpB1FE+N1txhmICWF70ftwYq02a32v4RjxP9UOuCWF
eQA5Axwmx+9ECcg+Y
---END CERTIFICATE---
```

Make sure the CA’s key is included as a trusted root.

In this example, the certificate for the Verisign Test ca can be obtained by storing the X.509 certificate returned by https://digitalid.verisign.com/cgi-bin/getcacert in a file and then using keyfm to install it as a trusted root. (See <CrossRef>Managing Trusted Roots, on page 201.)

Load the certificate into the keyfile.

Command line: `keyfm -cXv -i infile`

Example: `keyfm -cXv -i x509certfile`

The example command line reads a certificate from the file `x509certfile` and installs the certificate into the `keyfile`. The `c` flag installs the certificate, and the `X` flag specifies that `x509certfile` contains an X.509 certificate. Without the `X` flag, the input stream is assumed to be a PKCS#7 certificate chain. Note that `keyfm` automatically analyzes the input data and performs any necessary removal of BEGIN/END lines and Base64 decoding. The `keyfm` code snippet for this command begins at "case RECV_CERT:"

If the CA has modified the Distinguished Name so that the subject Distinguished Name in the certificate does not match the Distinguished Name in the `csr`, `keyfm` will prompt the user to verify whether or not the certificate should be accepted anyway. To do this, the spider uses the `change_verify` callback function given to `TSW_install_certificates()`. The Verisign Test ca has been known to change the ASN.1 encoding of the subject Distinguished Name.

4. Display the `keyfile` to confirm that it contains the certificate.
It should now appear as a “certified” key. (See “Displaying Distinguished Names for Public or Private Key Pairs” on page 198.)

**Note**  Steps 1 and 3 above can be combined by including the \( r \) flag when you are generating the key. Thus, the following example both creates a new key pair and generates its csr.

**Example:** `keyfm -kvfrb -o csrfile "William Shakespeare"

Also note that the full Distinguished Name can be specified using the \( N \) flag at the command line. If no Common Name or full Distinguished Name is included in the command line, `keyfm` provides you with a list of keys (represented by their Distinguished Name), and prompts you to select one.

### Displaying Distinguished Names for Public or Private Key Pairs

This example displays the Distinguished Names for all key pairs in the keyfile.

**Command line:** `keyfm -D

**Example:** `keyfm -D

The \( D \) flag specifies that key pairs should be listed. The output will be similar to the following:

```
CN=demo key,O=Demonstration Issuer,C=US (certified)
CN=William Shakespeare,OU=Engineering,O="Terisa Systems,Inc.",ST=California,C=US (uncertified)
```

Each output line displays the Distinguished Name of a key pair in the keyfile. The terms “certified” and “uncertified” in the output denote whether the key pair’s associated certificate is installed in the keyfile. If a certificate signing request has not been generated for a key pair, the Distinguished Name displayed will be obtained from the key pair’s label, and it will be followed by “no csr” instead of “certified” or “uncertified.”

To display more information about the certificate, include the \( l \) flag for “long form” output.

**Command line:** `keyfm -Dl

**Example:** `keyfm -Dl

The D flag specifies that key pairs should be listed.

The output will be similar to the following:
Subject Distinguished Name:
CN=demo key,O=Demonstration Issuer,C=US
Issuer Distinguished Name:
CN=Terisa CA,O=Demonstration Issuer,C=US
Certificate Serial=342ab1f0
RSA public key length: 512 bits
Valid End: Sat Sep 25 11:48:15 1999
Status: REVOCATION UNKNOWN
No Certificate Extensions Exist
The keyfm code snippet for this command begins at “case DISPLAY_CERTS:”.

Using PKCS#12 to Export Public or Private Key Pairs and Certificates

This example exports a key pair and any associated certificates into a PKCS#12 file. The export data is encrypted and MAC-protected with a password that was supplied interactively.

Command line: keyfm -Ev -o pkcs12_filename "Common name"

Example: keyfm -Ev -o outfile.p12 "William Shakespeare"

Enter password for encryption and MAC check:

The E flag exports the key pair and certificates to the PKCS#12 file, the v flag specifies verbose output, and the -o outfile.p12 argument writes the PKCS#12 data to the file outfile.p12. The key pair to be exported contains “William Shakespeare” in the Common Name field of its Distinguished Name. If a key pair is not specified by its Distinguished Name on the command line (for example, if “William Shakespeare” were omitted above), keyfm displays a list of available key pairs to export.

keyfm prompts the user for a password for encrypting data and generating the MAC for the PKCS#12 file. For security, the user will not see the password characters as they are typed.

Once this command executes, the resulting PKCS#12 file can be imported into web browsers, e-mail programs, and other applications supporting PKCS#12.
Using PKCS#12 to Import Public or Private Key Pairs and Certificates

This example imports a key pair or certificate from a PKCS#12 file. The imported data is decrypted and MAC-verified with a password that was supplied interactively.

Command line: keyfm -Iv -i pkcs12_file

Example: keyfm -Iv -i infile.p12

Enter password for decrypting and MAC check:

The I flag imports the key pair or certificate from the PKCS#12 file, the v flag specifies verbose output, and the -i infile.p12 argument specifies the file containing the PKCS#12 data. If the key pair or certificates were successfully imported, keyfm displays information about the certificates or key pairs it has imported.

Note that PKCS#12 files can contain only certificates or only key pairs. Also, the trusted root for an imported certificate must be present for the associated key pair to become certified. keyfm currently does not employ an UnknownCAHook when it imports PKCS#12 data. Thus, if the keyfm output displays No certificate found for key pair for a given key pair, the trusted root for that key pair's certificate chain probably is not installed in the keyfile.

Deleting Public or Private Key Pairs and Certificates

This example describes the steps required to delete a public or private key pair and any associated certificates from the keyfile.

To delete a key pair and associated certificate:

1. Use keyfm to delete the key pair and any certificates.

   Command line: keyfm -zv "Common name"

   Example: keyfm -zv "William Shakespeare"

   The z flag deletes the key pair and certificates, and the v flag specifies verbose output. The keyfm code snippet for this command begins at "case DELETE_KEYPAIR:"

2. Display the keyfile (see “Displaying Distinguished Names for Public or Private Key Pairs” on page 198) to confirm that the key pair and certificates have been deleted.

   As with many keyfm commands, if a Common Name or Distinguished Name is not provided at the command line, keyfm will display a list of available keys, and prompt you to select one.
Managing Trusted Roots

This section describes how to use the keyfm demonstration application to display, delete, and install trusted roots.

Displaying Trusted Roots

Use keyfm to display the pre-compiled trusted roots.

Command line: keyfm -T

Example: keyfm -T

This example assumes that the key file does not initially contain any trusted roots. When no trusted roots are found in the keyfile, the toolkit uses a set of hard-coded roots.

The output should be similar to the following:

The "mykeys.kdb" keyfile contains 2 trusted roots.
OU=Commercial Certification Authority,O="RSA Data Security, Inc.",C=US
OU=Secure Server Certification Authority,O="RSA Data Security, Inc.",C=US

The output above reflects the toolkit's hard-coded trusted roots:

- RSA's commercial ca
- RSA's secure server ca

Deleting Trusted Roots

To delete a trusted root:

1. Delete a root key from the list of trusted roots.

   Command line: keyfm -u

   Example: keyfm -u

   The u flag deletes a trusted root. A list of trusted roots will be displayed by keyfm, and you can select the desired root key by number. When the root has been deleted, the remaining trusted roots will be written to the keyfile.

   The keyfm code snippet for this command can be found at "case READ_TRUSTED_ROOT:"

2. Display the keyfile (see “Displaying Trusted Roots” on page 201) to confirm that the trusted root has been deleted from the keyfile.
Installing Trusted Roots

To install a trusted root:

1. Obtain the CA’s certificate.

   For the Verisign Demo certificate authority mentioned above, the X.509 certificate can be obtained by saving the data returned by:

   https://digitalid.verisign.com/cgi-bin/getcacert

   `keyfm` uses the CA’s certificate to install the new trusted root. Note that `keyfm` analyzes the input data and removes any necessary BEGIN/END lines and decodes any Base64 data.

   **Command line:** `keyfm -t -i infile`

   **Example:** `keyfm -t -i certfile`

   The `t` flag installs the trusted root.

2. Display the `keyfile` to confirm that the trusted root has been re-installed in the `keyfile`. (See “Displaying Trusted Roots” on page 201.) Including the `l` flag causes `keyfm` to display any available certificates for the trusted root.

   `CN=Terisa CA,O=Demonstration Issuer,C=US`

   Since this key is not a hard-coded trusted root, the four hard-coded roots were written to the `keyfile` with this new key when it was originally installed.

Using the ca Tool for Quick Testing

Instead of using a certification authority on the Internet, you can use the `ca` tool for developing and debugging your search engine.

**Example:** `keyfm -kvfr -o csrfile "William Shakespeare"`

**Example:** `ca -K ~/tsw/tools/ca/keydata/mykeys_rsa.kdb -i csrfile -o certfile`

**Example:** `keyfm -cX -i certfile`

Since base64-decoding of `csr` data currently does not work correctly in the `ca` program, use the binary form of the `csr`. In other words, do not set a `-b` flag to `keyfm` when you create a `csr`. 
Installing an UnknownCAHook

A certificate can be accepted by a call to `TSW_install_certificates()` even when a trusted path cannot be found between the certificate issuer and a trusted root recognized by the toolkit. In this case, the new key pair is not fully certified until a path to a trusted root, such as the certificate for the issuer, is created. However, this may not work if the ca has modified the key pair’s Distinguished Name.

In the example below, the `keyfile` does not contain the trusted root for the Terisa Test ca:

Example: `keyfm -cX -i certfile`

Couldn't find a path to a trusted root for a key pair or certificate.

Here is the last certificate found in the chain:

Subject Distinguished Name:
CN=testing,OU=Engineering,O="Terisa Systems, Inc.", ST=California,C=US

Issuer Distinguished Name:
OU=For Testing Only,OU=Terisa Test CA (RSA key), O="Terisa Systems, Inc",C=US

Certificate
Serial=546875204665622032352031323a31343a32342031393939390a
Valid Begin: Thu Feb 25 12:14:24 1999
Valid End: Sat Feb 24 12:14:24 2001
Status: UNVERIFIED
No Certificate Extensions Exist

Do you want to install this certificate as a new trusted root (y/n)? y
Using Content Assistants

*Content Assistants* are external programs that you write to classify, filter, and enhance information about documents when they are indexed by Ultraseek. During the indexing process, a Content Assistant can assign one or more topics to any document, filter documents, add document metadata, or replace document title and description information.

This chapter contains the following sections:

- Understanding Content Assistants
- About the Content Assistant Interface
- Configuring Your Service
- Creating a Topic Tree
- Pushing the Topic Tree
- Document Evaluation
- Modifying Document Contents
- Example GetFullDesc Message
- DTD for Topics.xml
Understanding Content Assistants

Content Assistants are programs that you create to interact with Ultraseek to define and assign topics to indexed documents, filter unwanted or inappropriate documents, or add metadata to indexed documents. The Service Provider Interface (SPI) described in this chapter defines the interaction between Ultraseek and a Content Assistant service. Two sample Content Assistants, one implemented in Python and one in Java, are available on Verity’s Technical Support site:

https://customers.verity.com

A license key is required in order to access Content Assistant functionality. The Content Assistant feature will not be enabled in the Ultraseek Administration UI if you do not have the required license. Please contact sales@verity.com for more information.

Note  You cannot use a Content Assistant to inject new documents into the Ultraseek index.

Evaluating Documents

When Ultraseek begins indexing its documents, it sends each document to the Content Assistant. Your service can use criteria of your own choosing to evaluate each document. The result of each evaluation of a document will always result in one of the following outcomes:

- The Content Assistant takes no action on the document and indexing of that document is not affected.

- Ultraseek is instructed by the Content Assistant to filter the document due to inappropriate, unwanted, or irrelevant document contents.

- Any topics and metadata assigned to the document by a Content Assistant will be stored by Ultraseek.
Example Uses

- Specify topic information for a document, which is stored by Ultraseek.
- Define a hierarchy of topics and make that topic tree known to Ultraseek.
- Identify documents containing obscene or pornographic material that should be filtered by Ultraseek.
- Add metadata containing information such as reviews or ratings to a document, based on the document’s content.
- Replace the title and description information for an indexed document.

About the Content Assistant Interface

Content Assistants and Ultraseek communicate using a sequence of interactions, as shown in Figure 9-1.

1. Content Assistants may optionally choose to transmit a document topic hierarchy to Ultraseek using HTTP PUT. This can be done at any time.

2. Ultraseek stores or updates the topic hierarchy. Any new topics are merged with the existing hierarchy. Topic information that already exists in the hierarchy will be replaced by the new information.

3. When Ultraseek begins re-indexing documents, it invokes a SOAP procedure asking your Content Assistant to evaluate each document. You pre-configure Ultraseek to send the type of information your Content Assistant needs to evaluate each document. The document re-indexing process is controlled and configured using the Ultraseek Administration UI.

4. Your Content Assistant processes the SOAP request, evaluates the document using criteria you define, and then invokes another SOAP procedure to send the results of the document evaluation back to Ultraseek.

5. Ultraseek processes the response and updates the information for the document. This process continues until all documents have been evaluated or until an error occurs. If an error occurs, the collection suspends and email is sent to the Ultraseek administrator.
The following diagram describes the transactions that occur between Content Assistants and Ultraseek.

**Figure 9-1** Transactions between Content Assistants and Ultraseek
Configuring Your Service

You use the Content Assistant Service page in the Ultraseek Administration UI to define information about your service, such as which collections to search, the URI for the Content Assistant, and the level of information to send for each document in the collection.

To configure your service:

1. Click the **CAS** tab in the Ultraseek administrative interface.
   The Content Assistant Service page displays.

   **Figure 9-2  Content Assistant Service Page**

   ![Content Assistant Service Page](image)

   - **Content Assistant Service**
     - Enable Content Assistant Service: ☐
     - Service URI: 
     - Which data level: GetURLDesc
     - Dump Debug Info: None
     - Which collections should use Content Assistant?
       - [econ] The Economist
       - [nytimes] New York Times
       - [python] python.org
       - [swf] Sun Freeware
       - [veritysp] Verity Spidered

2. Check “Enable Content Assistant Service.”

3. Enter the location where the external Content Assistant is running. This includes the host name, port and path name, if any.

4. Set the data level for each document that is sent to the Content Assistant. See “Document Evaluation” on page 216 for more information.
## 9 Using Content Assistants
### Configuring Your Service

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>GetURLDesc</td>
<td>Sends the URL name, content type, size, collection name, and date of modification for each document being indexed to the external Content Assistant.</td>
</tr>
<tr>
<td>GetHdrDesc</td>
<td>In addition to the information sent for GetURLDesc, sends the document title and description.</td>
</tr>
<tr>
<td>GetFullDesc</td>
<td>Sends all available information about a document that is known to Ultraseek, including META tags and body text.</td>
</tr>
</tbody>
</table>

5. Set the level of debugging information. You can choose from the following options:

| None                         | Prints only the URL of the document sent to the Content Assistant.                                               |
| Ingoing/Outgoing             | Prints the contents of the fields of each message sent to the Content Assistant and the fields of each reply message received from the Content Assistant. |
| Data Fields                  |                                                                                                                  |
| Ingoing/Outgoing             | Prints the complete content and HTTP headers of each SOAP message sent and received from the Content Assistant.    |
| SOAP Messages                |                                                                                                                  |
| Both SOAP Messages and Fields| Prints both the fields and SOAP messages exchanged between Ultraseek and the Content Assistant.                  |

6. Check the check box next to each collection that you want the Content Assistant to use.

7. Click **OK**.
Creating a Topic Tree

The topic tree used by your Content Assistant can be as simple or as complex as you need it to be. You define your topic tree using an XML format made up of `<subject>` elements. Figure 9-3 shows a sample topic tree; the following page shows how that topic tree could be defined in XML.

Your topic tree definition must adhere to the following rules:

- Each topic is defined using a `<subject>` element.
- Each `<subject>` element has a unique ID number (positive integer less than $2^{31}$). The values 0, 1, 2, and 3 are reserved by the system.
- Each `<subject>` element also has a `<name>` element and a `<superTopic>` ID. The top-level `<superTopic>` ID must have a value of zero. The values 1, 2, and 3 are reserved by the system.
- A topic may cross-reference another topic by including an `<xref>` element containing the ID number. A topic may define cross-references to multiple topics.

See “DTD for Topics.xml” on page 225 for the complete Document Type Declaration of the topic tree format used by Ultraseek.
Sample Topic Tree

This section describes a sample topic tree and gives its corresponding XML.

Figure 9-3 Sample Topic Tree

```
Energy
   Alternative Fuel
     Solar
     Wind
     Geothermal
   Fossil Fuel
     Coal
     Oil
     Natural Gas
   Nuclear
     Fusion
     Fission
   ...
```
Sample XML Definition

```xml
<subject>
  <id>100</id>
  <name>Energy</name>
  <superTopic>0</superTopic>
</subject>
<subject>
  <id>110</id>
  <name>Alternative Fuel</name>
  <superTopic>100</superTopic>
</subject>
<subject>
  <id>111</id>
  <name>Solar</name>
  <superTopic>110</superTopic>
</subject>
<subject>
  <id>112</id>
  <name>Wind</name>
  <superTopic>110</superTopic>
</subject>
<subject>
  <id>113</id>
  <name>Geothermal</name>
  <superTopic>110</superTopic>
</subject>
<subject>
  <id>120</id>
  <name>Fossil Fuel</name>
  <superTopic>100</superTopic>
</subject>
<subject>
  <id>121</id>
  <name>Coal</name>
  <superTopic>120</superTopic>
</subject>
<subject>
  <id>122</id>
  <name>Oil</name>
  <superTopic>120</superTopic>
</subject>
<subject>
  <id>123</id>
  <name>Natural Gas</name>
  <superTopic>120</superTopic>
</subject>
```
Cross-References

A subject element can cross-reference another subject element by including an `<xref>` element that contains the `<id>` of another subject element. In the following example, the topic **Fuel Cell** contains a cross-reference to the topic **Natural Gas**.

```xml
<subject>
  <id>143</id>
  <name>Fuel Cell</name>
  <superTopic>140</superTopic>
  <xref>123</xref>
</subject>
```

Pushing the Topic Tree

Your service may optionally push the topic tree that it will use to Ultraseek using the HTTP PUT mechanism. If the topic tree has already been pushed before, any existing topic data will be replaced with the new topic data. Any new topic data will be merged with the existing topics. The topic data that is transmitted includes a service identification block and the topic tree.

CAS Identification Block

The identification block contains the version identifier and the creator name elements.

The code sample below shows a sample identification block and topic tree:

```xml
<UltraseekTopics version="5.0">
  <creator>ExternalCE</creator>
  <subject>
    <id>100</id>
    <name>Energy</name>
    <superTopic>0</superTopic>
  </subject>
  <subject>
    <id>110</id>
    <name>Alternative Fuel</name>
    <superTopic>100</superTopic>
  </subject>
  ...
</UltraseekTopics>
```
**HTTP PUT Details**

Your Content Assistant should follow the HTTP PUT conventions described below.

- Specify the destination on the Ultraseek server that the topic will be pushed to using the following syntax: "PUT /admin/topics.xml HTTP/1.0".

- Specify additional HTTP headers that need to be included with the PUT request, as follows:
  - Include the “Content-Type” header which should be “application/xml”.
  - Include an “Authorization” field which should have the value “Basic” followed by a base64 encoding of the string `<Ultraseek Admin UserName>:<Password>`.
  - The PUT message should be sent to the same port number that the Ultraseek server is listening on.

An example of a PUT message sent to an Ultraseek server is shown below. The authorization string is a base64 encoding of the string “tiger:scottdb”.

```
PUT /admin/topics.xml HTTP/1.0
Content-type: application/xml
Content-length: 2553
Authorization: Basic dGlnZXI6c2NvdHRkYg==

<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE UltraseekTopics [ ...
```

(continued)
Document Evaluation

Ultraseek will begin sending document evaluation requests to your service for each document it needs to index. Ultraseek may send multiple, simultaneous requests to your service.

Your service may only require Ultraseek to send each document’s URL, just the document’s header information, or the entire document content. The type of evaluation request that is sent to your service by Ultraseek will be consistent with the way you configured your service, described on “Configuring Your Service” on page 209.

The URLDesc, HDRDesc, and FullDesc data formats are used to describe the document characteristics to your service.

URLDesc Evaluation

If your service has been configured to receive document URLs, Ultraseek uses the GetURLDesc procedure to deliver a data structure of type URLDesc to your service. The URLDesc data structure, shown below in WSDL format, contains the document’s URL, content type, size, date of last modification, and collection name.

```xml
<complexType name="URLDesc">
    <all>
       <element name="url" type="xsd:anyURI" />
       <element name="contentType" type="xsd:string" />
       <element name="size" type="xsd:int" />
       <element name="lastModified" type="xsd:dateTime" />
       <element name="collection" type="xsd:string" />
    </all>
</complexType>
```

After evaluating the URL, your service uses a GetHdrDescResp message to send a reply to Ultraseek.
HdrDesc Evaluation

If you configured your service to receive complete header information for each document, Ultraseek uses the `GetHdrDesc` procedure to send a data structure of type `FullDesc` to your service. The `HdrDesc` structure, shown below in WSDL format, contains the document’s URL, content type, size, date of last modification, collection name, title, and description.

```xml
<complexType name="HdrDesc">
  <all>
    <element name="url"          type="xsd:anyURI" />  
    <element name="contentType"  type="xsd:string" />  
    <element name="size"         type="xsd:int" />    
    <element name="lastModified" type="xsd:dateTime" />  
    <element name="collection"  type="xsd:string" />  
    <element name="title"        type="xsd:string" />  
    <element name="description"  type="xsd:string" />  
  </all>
</complexType>
```

After evaluating the header information, your service uses a `GetHdrDescResp` message to send a reply to Ultraseek.

FullDesc Evaluation

If you configured your service to receive complete information on each document, Ultraseek uses the `GetFullDesc` procedure, described on “Example GetFullDesc Message” on page 224, to deliver a data structure of type `FullDesc` to your service.

The `FullDesc` structure contains an array of `DocItem` elements. The `FullDesc` message contains the same information contained in `HdrDesc` as well as the complete document contents.

A `DocItem` element will be present for each tag contained in the document. The body text of the document will have a field that is empty. The body text is free of any formatting such as carriage returns, linefeed characters, and tabs. Additionally, the body text is stripped of any characters that are not legal in XML, even if they are legal HTML characters.

Other possible `DocItem` elements include:

- **link**—represents a link present in the document.
- **imagelink**—represents a link to an image present in the document.
Note The `ArrayOfDocItem` element may contain multiple occurrences of body text fields, links to other documents, links to images, an ISO language identifier, or other document tags.

The code sample below shows the `DocItem` element used in the `FullDesc` data structure:

```xml
<!-- Used by FullDesc -->
<complexType name="DocItem">
<all>
<element name="field" type="xsd:string" />
<element name="text" type="xsd:string" />
</all>
</complexType>
```

The code sample below shows the `FullDesc` data structure in WSDL format:

```xml
<complexType name="ArrayOfDocItem">
<complexContent>
<restriction base="soapenc:Array">
<attribute ref="soapenc:arrayType" wsdl:arrayType="tns:DocItem[]"/>
</restriction>
</complexContent>
</complexType>

<complexType name="FullDesc">
<all>
<element name="url" type="xsd:anyURI" />
<element name="contentType" type="xsd:string" />
<element name="size" type="xsd:int" />
<element name="lastModified" type="xsd:dateTime" />
<element name="collection" type="xsd:string" />
<element name="title" type="xsd:string" />
<element name="description" type="xsd:string" />
<element name="ArrayOfDocItem" type="tns:ArrayOfDocItem" />
</all>
</complexType>
```

After evaluating the document information, your service uses a `GetFullDescResp` message to send a reply to Ultraseek.
**IndexAsstDesc**

The **IndexAsstDesc** data type is sent by your service to Ultraseek to indicate how the document was evaluated. This message identifies the evaluated document by its URL and indicates whether or not the document should be indexed. In addition, this message can provide a new title and description for the document which will overwrite the existing title and description. Optionally, this message may contain a list of one or more topic identifiers that define the topics with which the document should be associated. The message may also contain optional metadata to be stored by Ultraseek. If, for example, the document contains information on a book or an author, the metadata might be used to contain reviewer ratings of the book or author.

The code sample below shows the **IndexAsstDesc** data structure in WSDL format:

```xml
<complexType name="ArrayOfInt">
<complexContent>
<restriction base="soapenc:Array">
    attribute ref="soapenc:arrayType" wsdl:arrayType="xsd:int[]"/>
</restriction>
</complexContent>
</complexType>

<complexType name="IndexAsstDesc">
<all>
    <element name="url" type="xsd:string" minOccurs="0" />
    <element name="doIndex" type="xsd:boolean" minOccurs="0" />
    <element name="title" type="xsd:string" minOccurs="0" />
    <element name="description" type="xsd:string" minOccurs="0" />
    <element name="lastModified" type="xsd:dateTime" minOccurs="0" />
    <element name="ArrayOfInt" type="tns:ArrayOfInt" minOccurs="0" />
    <element name="ArrayOfDocItem" type="tns:ArrayOfDocItem" minOccurs="0" />
</all>
</complexType>
```

<!-- List of TopicIds -->
<element name="ArrayOfInt" type="tns:ArrayOfInt" minOccurs="0" />

<!-- Additional metadata -->
<element name="ArrayOfDocItem" type="tns:ArrayOfDocItem" minOccurs="0" />
</all>
</complexType>
IndexAsstDesc Contents

The following rules apply to the contents of the IndexAsstDesc message your service returns to Ultraseek. Large SOAP messages sent by Ultraseek to your Content Assistant use the multi-reference style of encoding documented in the SOAP 1.1 Protocol description available at http://www.w3.org/TR/SOAP.

- If an element originally sent by Ultraseek is not contained in the IndexAsstDesc message returned by your service, Ultraseek assumes the original value is to remain unchanged.

- If the doIndex element is not included in the IndexAsstDesc message, Ultraseek will assume the document is to be indexed.

- If your service returns a doIndex element set to false, Ultraseek will not index the document.

- If your service returns new values of the <url>, <title>, and <description> elements originally sent by Ultraseek, the new values will overwrite the old values. If the document appears as a query result, its summary information will display the updated values. However, these updated values will not be searchable. For more information, see “Modifying Document Contents” on page 221.

- If your service returns new values of the <lastModified> element originally sent by Ultraseek, the new values will overwrite the old values. If the document appears as a query result, its summary information will display the updated values. In addition, the <lastModified> element needs to be encoded in ISO 8601 format.

- If your service returns a DocItem element that was not originally sent by Ultraseek, the new element will be added to the end of the original DocItem array. See “Modifying Document Contents” on page 221 for details.

- The ArrayOfDocItem type may contain multiple occurrences of body text fields or other document fields.

- Body text represented within the ArrayOfDocItem element always has an empty <field> element.

- If an invalid TopicId is specified in the IndexAsstDesc message by your Content Assistant, it will be ignored by Ultraseek.
Your service can overwrite the `<url>`, `<title>`, `<description>` elements and can add new elements to the `DocItem` array. The values of existing `DocItem` elements cannot be overwritten. If the document appears as a query result, its summary information will display the updated values.

Due to the difference between how Ultraseek handles the `<url>`, `<title>`, and `<description>` elements (which are part of the “title” record) and the elements in the `DocItem` list (which are part of the “index” record), title record elements are not searchable. If you want the `<title>` and `<description>` elements of a document to be searchable, you should add the contents of these fields to the body text via entries in the `DocItem` array. For example, if you want to associate a new `<url>` with the document, return an entry in the `DocItem` array with the field name “url” and value set to the new URL.

Assume that Ultraseek sends your service the elements shown below within a FullDesc message.

```xml
...<description xsi:type="xsd:string">Readme File</description>
<url xsi:type="xsd:anyURI">www.stuff.net/booklist</url>
<title xsi:type="xsd:string">Read these books</title>
...<ArrayOfDocItem SOAP-ENC:arrayType="xsd:SOAPStruct[5]"
    xsi:type="SOAP-ENC:Array">
  <xsd:item>
    <field/>
    <text>Add these books to your reading list.</text>
  </xsd:item>
  <xsd:item>
    <field xsi:type="xsd:string">author</field>
    <text xsi:type="xsd:string">Jack Kerouac</text>
  </xsd:item>
  <xsd:item>
    <field xsi:type="xsd:string">title</field>
    <text xsi:type="xsd:string">Dharma Bums</text>
  </xsd:item>
  <xsd:item>
    <field xsi:type="xsd:string">author</field>
    <text xsi:type="xsd:string">David James Duncan</text>
  </xsd:item>
  <xsd:item>
...
After reviewing the document, your service determines it wants to overwrite the original <url>, <title>, and <description> value elements. Further assume that your service wants to add elements to the original DocItem element values. To do this, your service could return a FullDesc message containing the elements shown below.

```
<description xsi:type="xsd:string">Reading List</description>
<url xsi:type="xsd:anyURI">www.greatstuff.net/mustread.html</url>
<title xsi:type="xsd:string">Help file</title>
```

The final values that would be stored by Ultraseek are shown below. Note that the <description> and <url> elements have been overwritten by the new values specified by the service and the two new DocItem elements have been appended to the original list.

```
<description xsi:type="xsd:string">Reading List</description>
<url xsi:type="xsd:anyURI">www.greatstuff.net/mustread.html</url>
```

```
<ArrayOfDocItem SOAP-ENC:arrayType="xsd:SOAPStruct[7]" xsi:type="SOAP-ENC:Array">
<xsd:item>
    <field xsi:type="xsd:string">author</field>
    <text xsi:type="xsd:string">Arturo Perez-Reverte</text>
</xsd:item>
<xsd:item>
    <field xsi:type="xsd:string">title</field>
    <text xsi:type="xsd:string">The Club Dumas</text>
</xsd:item>
...```
<field xsi:type="xsd:string">author</field>
<text xsi:type="xsd:string">Jack Kerouac</text>
</xsd:item>
<xsd:item>
<field xsi:type="xsd:string">title</field>
<text xsi:type="xsd:string">Dharma Bums</text>
</xsd:item>
<xsd:item>
<field xsi:type="xsd:string">author</field>
<text xsi:type="xsd:string">David James Duncan</text>
</xsd:item>
<xsd:item>
<field xsi:type="xsd:string">title</field>
<text xsi:type="xsd:string">The Brothers K</text>
</xsd:item>
<xsd:item>
<field xsi:type="xsd:string">author</field>
<text xsi:type="xsd:string">Arturo Perez-Reverte</text>
</xsd:item>
<xsd:item>
<field xsi:type="xsd:string">title</field>
<text xsi:type="xsd:string">The Club Dumas</text>
</xsd:item>
</ArrayOfDocItem>
Example GetFullDesc Message

Ultraseek invokes the SOAP procedure GetFullDesc to provide your service with the full contents of a document so that your service can evaluate the document. A sample GetFullDesc SOAP message sent by Ultraseek is shown below.

```xml
<SOAP-ENV:Envelope
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/1999/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/1999/XMLSchema"
>
  <SOAP-ENV:Body>
    <m:GetFullDesc xmlns:m="http://ultraseek.com/indexingassist.wsdl ">
      <FullDesc>
        <lastModified xsi:type="xsd:dateTime">2001-10-10T18:35:37.764782Z</lastModified>
        <size xsi:type="xsd:int">222567</size>
        <title xsi:type="xsd:string">Heart of Darkness</title>
        <description xsi:type="xsd:string">One of Conradís most enduring works</description>
        <contentType xsi:type="xsd:token">text/html</contentType>
        <collection xsi:type="xsd:string">Books</collection>
        <ArrayOfDocItem SOAP-ENC:arrayType="xsd:SOAPStruct[7]"
          xsi:type="SOAP-ENC:Array">
          <xsd:item>
            <field xsi:type="xsd:string">title</field>
            <text xsi:type="xsd:string">Heart of Darkness</text>
          </xsd:item>
          <xsd:item>
            <field/>
            <text xsi:type="xsd:string">The Nellie, a cruising yawl, swung to her anchor without a flutter of the sails, and was at rest.</text>
          </xsd:item>
        </ArrayOfDocItem>
      </FullDesc>
    </m:GetFullDesc>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```
The following DTD illustrates how to write well-formed topics.xml file:

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE UltraseekTopics (View Source for full doctype...)
<!-- Note: Topic Id 0 is reserved by UltraSeek. Using this -->
<!-- as topicid/xrefid can lead to strange behavior.-->
<UltraseekTopics version="3.1">
<creator>janem</creator>
<date>2004-01-30T19:30:40Z</date>
<location>apollo.mycompany.com:8765</location>
  __<subject>
    <id>1</id>
    <name>Home</name>
  </subject>
  __<subject>
    <id>670429875</id>
    <name>Business Conditions</name>
    <superTopic>0</superTopic>
  __<rule>
    <required>zzzzzzzzzzzzzzzzzzzzzz</required>
  </rule>
</subject>
</UltraseekTopics>
```
9 Using Content Assistants
DTD for Topics.xml
Search Syntax

Knowing how to create refined, efficient searches is important for both users and administrators. Leveraging the Ultraseek search syntax to its full potential yields efficient, specific results. Users are helped along with this through rotating tips, the help interface, and the Advanced Search form.

This appendix describes the Ultraseek syntax and contains the following sections:

- Search Strategies
- Plus and Minus Operators
- Vertical Bar and “Search these results”
- Double-Vertical bars
- Wildcard Search
- Field Searches
Search Strategies

It is important to identify phrases and terms when constructing a query. In each case, Ultraseek is instructed to match words in consecutive order.

Spell Checking

Query, or spelling suggestions are not made for every query. Ultraseek insists on a reasonable level of confidence before making the suggestion. This confidence is based on the typographic difference between the query and the suggestion as well as the frequency of the suggested query terms relative to the frequency of the original query terms.

Suggestion, not correction: Ultraseek does not correct the spelling of queries. Instead Ultraseek proposes alternate queries based on re-spelling the terms of a query. For instance, in the following English language queries, each query term is correctly spelled:

silicone valley
silicon implant

However, the following suggestions can be made:

silicone valley
Did you mean: silicon valley?
silicon implant
Did you mean: silicone implant?

After the user has submitted a query, if Ultraseek has an alternate query to suggest, it will be presented under the query box, and above the search results. For instance, if the original query is:

britneey speers rock music

The following suggestion appears:

Did you mean: britney spears rock music

The phrase “britney spears” is hyperlinked back to the query page, with the adjusted spelling. Words with adjusted spellings from the original query are presented with the HTML <I> emphasis.

Spell Checking can be disabled by commenting out the following line of Python in query.html located in the Ultraseek /docs directory:

```python
if enable_spell and st == 1: exec self.file("spell.html");
```
The commented version looks like this:

```python
# if enable_spell and st == 1: exec self.file("spell.html");
```

After you turn off spell checking, you need to reload the query.html for the changes to take effect. To reload the file, click the `reload` button next to the Document directory field under the Server | Parameters | Main page.

### Phrases and Terms

Consider the following query:

white house cat

This query will produce better results if phrases are identified. Surround phrases with quotation marks to identify them:

"white house" cat

white “house cat”

In the first example “white house” cat, Ultraseek will recognize the two phrases. In the first example, the search results will contain the phrase “white house” as a primary phrase and the results will be ranked higher. The search will also return “cat” as a secondary phrase, or word, and will rank the documents containing that word/phrase with lower relevancy.

**Note** If there is a document that contains both the phrase “white house” and the term “cat”, then that document will be ranked by the phrase “white house”. Therefore, it is possible to have a document ranked very high which contains the term “cat”, even though this is the secondary search term.

In the second example, white “house cat”, Ultraseek will again recognize both phrases, but with differing results. In the second result set, the color white will be ranked higher than “house cat” because it appears first in the search terms. The results set will have all documents with the term “white” first, “house cat” second, and white “house cat” third.
Commas and Separating Phrases

Ultraseek allows you to search phrases without quotation marks. By doing this, Ultraseek interprets the entire statement as one phrase. If you want to search multiple phrases and enter the example below, you will get search results which contain all of the terms because Ultraseek interprets “white house cat” as a single phrase.

white house cat

Note: Long phrase searches may return documents that contain a portion of the phrase.

However, if you want to search primarily for “white house”, and secondly for “cat”, then you would separate the terms by using a comma as below:

white house, cat

Your search results would first contain documents with “white house” in them, and then documents with “cat”.

Alternatively, if you want to search primarily for the term white, and secondly for “house cat”, the search entry would look like this:

white, house cat

Your search results would first contain documents with the term white in them, and then the documents with house cat. If you had intended to search for documents containing “white house cat” as a single phrase, then you would enter the terms without a comma, and in parenthesis:

“white house cat”

Capitalization

If a word is submitted entirely in lower case, case is ignored in the resulting hits. However, if any capitalization is used, case is exactly matched.

For example, a query for next will return documents containing the words “next”, “Next”, and “NeXT”. However, a query for NeXT will only return documents containing the word with exactly matching case: “NeXT”.
Plus and Minus Operators

The plus and minus operators are useful tools that improve the control you have over search results by allowing you to exclude or include terms.

**Plus**

If a term must appear in each search result, put a plus in front of it:

```
product price +module
```

In the above example, the words `product` and `price` might be included in the search results. However, the word `module` must occur in each hit. The plus sign can be used with other search syntax tools:

```
+site:www.verity.com +"new and improved"
```

In the above example, the plus sign is used in conjunction with the site field search. The plus sign is also used with quotation marks to require a phrase.

**Note** Using the “+” excessively is not recommended. Ultraseek will scan each document for the entire term `product price module` when it determines relevance. Therefore, if you need to see the term `module` in each of the results, enter the terms without a comma, and Ultraseek will interpret them as a phrase and rank all documents with all three terms higher in relevancy then those without.

**Minus**

Put a minus sign in front of terms that are not desired in search results:

```
bats -baseball
```

Search results will prefer documents that contain information about `bats (flying rodents)`, over `bats (baseball bats)`. However, the minus operator does not necessarily exclude terms altogether. Documents that contain terms marked with a minus operator will be excluded from the search results.
Vertical Bar and “Search these results”

The vertical bar and “Search these results” are useful tools that allow users to create specific, refined queries. The vertical bar will tell Ultraceek to include content on both sides of the vertical bar, where the string to the right of the vertical bar is a subset of the string to the left of the vertical bar. The vertical bar behavior mimics “Search these results” behavior. For example, suppose a user were to type the following query:

restaurant bistro

Resulting hits will contain the words “restaurant” and/or “bistro”. The user then clicks “Search these results” and enters the subsequent query:

pizza spaghetti

Results are now narrowed to those hits which also contain the words “pizza” and/or “spaghetti”. This process is called query refinement.

This same behavior can be duplicated in a single query by using the vertical bar, as shown in the following example:

restaurant bistro | pizza spaghetti

Resulting hits will contain the words “restaurant” and/or “bistro”, and also contain the words “pizza” and/or “spaghetti”.

Note that the vertical bar can be used repeatedly in the same query. For example, the above query can be refined even further:

restaurant bistro | pizza spaghetti | San Francisco "bay area"

The Difference Between Vertical Bar and Plus

There is often confusion between the vertical bar operator and plus operator. In many cases, both operators produce the same effect. For example, this query

+restaurant +pizza

and this query

restaurant | pizza

will return the same results. In each case, both the word “restaurant” and the word “pizza” must appear in the search results. However, by adding several more query terms to the mix, the difference becomes more apparent.
Examples

+restaurant +bistro +pizza +spaghetti

restaurant bistro | pizza spaghetti

In the first example, the search results must contain all of the terms listed. In the second example, the search results must contain at least one term to the left and one term to the right of the vertical bar.

Double-Vertical bars

In search syntax, the double-vertical bar has exactly the same function as the vertical bar, but with one difference: relevance ranking. In searches that contain a double-vertical bars, only the content to the right of the double-vertical bars is used in relevance ranking. Content to the left of the double-vertical bars has no impact on relevance scores.

For example, consider the following query:

restaurant bistro || pizza spaghetti

The resulting hits will include content from both sides of the double-vertical bars, but only the words “pizza” and “spaghetti” will impact relevance scores. Hits containing these words will rank higher in the search results.

There is little use for the double-vertical bars in user queries, however administrators may want to know this syntax in order to better understand the qp form variable. When an administrator sets a value for the qp form variable, that value is prefixed to the user’s query with a double-vertical bar. For example, suppose the administrator has set a qp value of “language modules” in a hidden form element:

<input type="hidden" name="qp" value="language modules">

Now suppose a user types in the following query:

parts accessories

Ultraseek interprets the combination of qp and the user query with the following syntax:

language modules || parts accessories

Resulting hits will contain the word “language” and/or “modules”, and will contain the word “parts” and/or “accessories”. Relevance scores will be ranked based on the words parts and accessories.
Wildcard Search

Wildcard search expands the results of a search query to terms that match a wildcard operator in a search term. Ultraseek expands the wildcard term to matching terms in the searched collections, and appends the term matches to the query string. For example, a search for the term “stan*” expands to include the following results: “stand”, “standalone”, and “stanford”.

Ultraseek supports wildcard expansion in individual terms, but not in phrases. For example, wildcards used within quotes (“Stanford undergrad*”) are ignored, and not expanded. Searching for Stanford Undergrad* will search for “Stanford” “Undergrad”, but not the phrase “Stanford undergrad”.

To increase the relevance of results, only terms that appear in three documents will appear as a wildcard expansion. In addition, wildcard searches are case sensitive: a search for “stan*” produces different results than a search for “Stan*”. Ultraseek expands each wildcard search to a maximum of 40 terms in the search results. This number may be less depending upon the terms and their frequency within the collections.

If the wildcard search expands to too many terms to be useful, you may have difficulty locating the term that you are searching for. In this case, try a more specific search term. For example, you may have better results searching for “install*” instead of “inst*”.

### Supported Operators

The following wildcard operators are supported:

<table>
<thead>
<tr>
<th>Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>Matches any number of characters in the search term, including the term itself. For example, a search for undergrad* returns results for “undergrad”, “undergrads”, and “undergraduate”.</td>
</tr>
<tr>
<td>?</td>
<td>Matches a single character. For example, a search for “st?n” returns results for “stan” and “stun”. Note: If you use a question mark (?) in a query, such as in ‘What’s up Doc?’, Ultraseek will expand the results for the word “Doc”, which may not return the intended search results. In this situation, enclose the search phrase in quotes.</td>
</tr>
</tbody>
</table>
Wildcard Position

The wildcard character can be at any position in the query term (leading, middle, or trailing). For example, “*stan”, “stan*”, “*stan*”, “st?n”.

**Note** Using wildcard operators at the beginning of a search term requires a full search of the term database in the collections you are searching, which will reduce search performance.

Rare Term Exclusion

The rare term exclusion threshold is configurable. In patches.py, you add the following information:

```python
import wildcard
wildcard.MIN_DOC_FREQ = 0 # Does not check for existence of term
wildcard.MIN_DOC_FREQ = 1 # Checks that the term exists in at least one doc
wildcard.MIN_DOC_FREQ = 3 # default value
<etc.>
```

Note that “deprecated” terms can persist in the term database for a time after they have been deprecated. For example, say you had a document containing “xyzzy”, but the document has been modified so it no longer contains “xyzzy”. The term “xyzzy” will persist in the terms database. If you set `MIN_DOC_FREQ` to 0, it will return xyzzy as an expansion of xy*. If you set `MIN_DOC_FREQ` to 1, it will check to see if xyzzy still (appears to) exist before expanding xy* to include xyzzy.

Using Modifiers in Wildcard Searches

Ultraseek supports the use of exclude modifiers. For example, `modifier -exclud*` expands to `modifier -exclude -excluded -excluding`. However, required modifiers (+) are not supported.
Field Searches

A field search is a search that looks for certain elements within a document. These elements and how to search on them are discussed in the following sections:

- “Searching by Document Title” on page 237
- “Searching by URL” on page 237
- “Searching by Description or Keywords” on page 238
- “Searching by Document Type” on page 238
- “Searching by Language” on page 239
- “Searching by Link” on page 240
- “Searching by Image Link” on page 240
- “Searching by Metadata” on page 240
- “Searching by Dublin Core Metadata” on page 241
- “Searching by Site” on page 241
- “Searching by Topic” on page 242
- “Searching by Encoding” on page 242

For example, suppose a user wanted to search for the phrase “New Year’s Resolution”. However, the user only wants documents returned which contain that phrase in the title. Documents containing the phrase in the body or elsewhere should be ignored. The user could construct a query that looks like this:

title:“new year’s resolution”

Users can also use the Advanced Search form to create the query, in which they do not need to know the exact syntax to create the search. However, knowing this syntax is extremely useful for the administrator who is populating a form variable that restricts searches to certain document titles.

This same format of field: text is followed for all field searches. Field searches can become completely customized by entering specific metadata within the HTML tag for use in field searches (see “Searching by Metadata” on page 240). For example:

<meta name="author" content="Edgar Allen Poe">

Using author: edgar, or author: allen, or author: poe will return all of the documents with this field value.
Searching by Document Title

You can search for an HTML document’s title by typing all or part of a title as your search query. The format is the same as for other field searches:

title:year

or:

title:year title:2000

The first query yields a broad search; the second query yields a more narrow search. To ensure that a particular phrase is found within a title field search, enclose the phrase in quotes:

title:“year 2000 compliance”

The above example yields the most narrow pool of search results.

Searching by URL

You can search on any document with a given URL. The URL can include any subdirectories. This enables you to perform a broad search:

url:www.verity.com

or a more narrowed search:

url:www.verity.com/products/search

You would get the same results if you typed in a more complete version of the same URL search:

url:http://www.verity.com/products/search

These queries yield all documents within the specified directories. However, you can also search for a particular word as part of the URL. For example, the query url:faqs returns all documents with “faqs” as a component of the document’s URL.
Searching by Description or Keywords

Users have the ability to search for elements within a document’s description or keywords. Both of these elements are created within the META tag of the document. For more information on the META tag, see “Searching by Metadata” on page 240.

The document’s description is the summary that appears in search results. Keywords are inserted to highlight certain topics within a document. Both are created by each document’s author.

Field searches by description or keywords follow the same format as other field searches:

description:modules
keywords:“press release”

Searching by Document Type

To search on a specific document type, use the doctype field search. For example, to search only Microsoft Word documents, type the query:

doctype:msword

This syntax will match the doctype of Microsoft Word documents specified by the MIME type “application/msword”. Other supported MIME types include, but are not limited to:

Adobe Acrobat application/pdf
Autocad Interchange application/x-dxf
Corel WordPerfect application/wordperfect5.1
HTML text/html

The document type is based on the MIME type of the document, depending on the configuration of your web server. If your web server has unique MIME type configuration, you can easily modify the interpreted MIME type through the Administrative Interface. See “Supported MIME Types” on page 174 for a complete listing of supported MIME types. For more information about specifying unique document types to a format that Ultraseek recognizes, such as text, see “Document Type Specification” on page 173.
Searching by Language

If you have multiple language support, you can perform searches based on a document’s language. To perform a search for documents in a particular language, use the following syntax:

`language:fr`

Resulting hits for the above example would be recognized by Ultraseek as being in French. Acceptable languages for the `language` field search are listed below. For detailed information about multiple language support, view the language FAQ on the Verity Technical Support site.

<table>
<thead>
<tr>
<th>Language</th>
<th>Field search variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese Simplified</td>
<td>zh_cn</td>
</tr>
<tr>
<td>Chinese Traditional</td>
<td>zh_tw</td>
</tr>
<tr>
<td>Danish</td>
<td>da</td>
</tr>
<tr>
<td>Dutch</td>
<td>nl</td>
</tr>
<tr>
<td>English</td>
<td>en</td>
</tr>
<tr>
<td>Finnish</td>
<td>fi</td>
</tr>
<tr>
<td>French</td>
<td>fr</td>
</tr>
<tr>
<td>German</td>
<td>de</td>
</tr>
<tr>
<td>Italian</td>
<td>it</td>
</tr>
<tr>
<td>Japanese</td>
<td>ja</td>
</tr>
<tr>
<td>Korean</td>
<td>ko</td>
</tr>
<tr>
<td>Norwegian</td>
<td>no</td>
</tr>
<tr>
<td>Portuguese</td>
<td>pt</td>
</tr>
<tr>
<td>Spanish</td>
<td>es</td>
</tr>
<tr>
<td>Swedish</td>
<td>sv</td>
</tr>
</tbody>
</table>
**Searching by Link**

Search for documents that link to a URL by using the `link` field search. This can be a useful tool for finding cross-references. For example, suppose you want to find how many documents from the `sales.verity.com` site are pointing to the `products.verity.com` site. This can be accomplished with a simple link field search:

```
+link:products.verity.com +site:sales.verity.com
```

Notice the use of the plus signs as requiring operators. This makes both components of the query required in the search results.

Ultraseek also returns documents with links to substrings of the above field search. If you want to limit your field search to a single URL, use the `linkexact:` field search operator.

**Searching by Image Link**

You can search for documents containing a specified image with the `imagelink:` field search operator. For example, suppose you want to search for documents containing `board.gif`, a picture of your company’s board of directors. The following query accomplishes this:

```
imagelink:http://companyXYZ//shared/img/board.gif
```

The `imagelink:` field search also returns documents with images containing a substring of the specified field search. For example, a document with an image link of `http://companyXYZ/corporate/board.gif` would also be returned by this query.

**Searching by Metadata**

You can also use the META HTML tag to customize field searches. For example, to create a collection that lists books and provides searches by author, embed the following META tag into the HTML:

```
<meta name="author" content="Edgar Allen Poe">
```

To find a document, you can then search `author:poe, author:allen, author:“edgar allen poe”`, and so forth. You can set the content to any string.

Another use of the META tag is to create keywords that help users find your document. The more keywords available for a document, the more likely users will find the document when searching. Below is an example of keywords you might list for a document containing information on Edgar Allen Poe:
<meta name="keywords" content="literature, poems, poetry, book, American, classic, Raven, Edgar Allan Poe, Edgar, Poe, Usher, black, cat, dark, stormy">

**Searching by Dublin Core Metadata**

Dublin Core meta searches are very similar to standard META searches, but with a slightly different format. All Dublin Core meta tags are case sensitive. In the two examples below, the first example shows all lower case, while the second shows varied case as well as lower case. A Dublin Core meta tag search might look something like this:

```plaintext
dc.subject:finance
```

Resulting hits will contain the word “finance” in the Dublin Core subject element.

If you would like a particular META tag to be treated like the `dc.subject`, in that words in the META tag are found in non-field qualified queries, then you add the name of that META tag to the “keywords” list in the HTML META Tag Names region (see “Specifying HTML META Tag Names” on page 109). For example, normally to search the `DC.Coverage.PlaceName` Meta tag, the user could enter queries with:

```plaintext
DC.Coverage.PlaceName:Library
```

or

```plaintext
dc.coverage.placename:library
```

However, if you add `dc.coverage.placename` to the Keywords field on the HTML Meta Tag Names page (see “Specifying HTML META Tag Names” on page 109), and re-index your data, the queries “Library” and “library” will return the same document.

**Searching by Site**

You can search on any web site within a collection. Unlike searching by URL, this search is limited to the host portion of a URL and does not return results for matches in the file portion of a URL. Search by site to list all pages in a given site. For example, the query:

```plaintext
site:sun.com
```


The query does not return results for sun.co.uk because uk does not match com. The query also does not return results for http://www.corp.net/sun.com since sun.com is in the file part of the URL.
Searching by Topic

To search on a specific topic, set the field search to the topic ID number for that topic. Note that you have topics only with the Content Classification Engine (CCE) add-on module.

You can find the topic ID for a topic under the Topics | Edit page of the Ultraseek administrative interface. Under the Current topics section of the page, click the topic in question and look for the ID value at the top of the Edit Topic page.

Suppose you wanted to limit your search to the “Scottish History” topic. You have determined that the topic ID for “Scottish History” is 1575973028. Set your field search as follows:

```
topic:1575973028
```

This query will only return results from within the “Scottish History” topic.

Searching by Encoding

Use the encoding: field search operator to search for documents based on how their characters are encoded. This can be useful if you want to determine whether your document collections contain certain kinds of encoded documents. For example, the following query returns all documents encoded with the ISO-8859-1 (Latin 1) character set:

```
encoding:ISO-8859-1
```
PART II

Content Classification Engine

- Chapter 11: Getting Started with CCE
- Chapter 12: Creating and Editing Topics
- Chapter 13: Creating Topic Rules
- Chapter 14: Viewing Reports
- Chapter 15: Mirroring
Getting Started with CCE

CCE, or Content Classification Engine, is an optional module to Verity Ultraseek that provides your users with the convenience of browseable topics combined with the power of full-text searches. CCE works with collections that you create using Ultraseek. Using CCE features in the Ultraseek administrative interface, you can create and manage topic hierarchies, apply rules to automate the classification of documents, and generate reports analyzing document populations in existing topics.

Multiple administrators can collaborate on a single instance of CCE. However, you should establish procedures to avoid any problems that might arise with this practice.

This chapter explains how to get started using CCE and contains the following sections:

- About CCE
- Obtaining a CCE License Key
- Starting CCE
- Sample CCE Topics Pages
- Setting CCE Display Parameters
- Customizing CCE in the Search Interface
About CCE

CCE organizes your site's content into a hierarchy of topics for easy searching by applying search rules that you create using the CCE administration tools. Using familiar point-and-click navigation, users can “drill down” through topics and subcategories that appear on the main search interface page with the query box. This helps them locate documents of interest.

With the CCE Module, search and browse capabilities are tightly integrated. Users can work with search functionality while viewing the topic hierarchy, conduct focused queries by searching within an individual topic, or simply browse through the topics. In response to queries, CCE displays search results and topics related to the returned documents. The result: faster, more accurate searches. The benefit: satisfied users.

The table below describes the benefits of using CCE.

<table>
<thead>
<tr>
<th>Search alone</th>
<th>Search + CCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ When a keyword or phrase can help find the right document.</td>
<td>■ When you don’t know the right keywords.</td>
</tr>
<tr>
<td>■ When you know exactly what to look for.</td>
<td>■ When you know it must be “in here somewhere.”</td>
</tr>
<tr>
<td></td>
<td>■ To help end users confine a search to a subset of a site or network.</td>
</tr>
<tr>
<td></td>
<td>■ When you want to create a more portal-like look and feel.</td>
</tr>
</tbody>
</table>

Manage Topics with Flexible Automation Features

The flexible automation features of CCE make the addition of browseable topics a manageable and cost-effective option for any organization.

Creation and maintenance of the topic hierarchy is simple and straightforward, occurring in real time. “Hands-off” administration ensures the most up-to-date topics possible by automatically adding new documents and removing links to deleted content. CCE easily handles your expanding needs through its high performance and scalability, capable of categorizing millions of documents.

Maintain Control With Flexible Automation Features

Maintain a high degree of control over every aspect of CCE using flexible administrative features. You can customize the display of topics to match your web site's look and feel, manage CCE from remote locations using a standard web browser, and control which documents display and the topics in which they are found.
Obtaining a CCE License Key

CCE is installed when you install Ultraseek. However, you must obtain a license key to enable the server’s CCE functionality. You can obtain a valid license key by contacting Verity sales at 1-888-328-SEEK. When you have your license key, use the Administrative Interface to install it.

Starting CCE

You use the CCE Topics page in the Ultraseek administrative interface to create new topics, import topics from existing HTML site maps, and edit topic rules.

To access the CCE topics page:

1. Launch the Administrative Interface in a web browser.
2. If necessary, log on to the Administrative Interface with the administrator ID and password, or your user account. The Ultraseek administrative interface displays.
3. Click Topics to access the CCE topics page.

The CCE topics page displays, shown below. This is the main page that you use to create and manage your topic hierarchy. Below the Current topics heading you’ll see a list of tasks you can perform.
Figure 11-1  CCE Topics Page

Ready Topics: 4
Edits pending: 0
Last edit by admin at Thu Apr 7 17:23:04 2005
Total URLs: 25122
URLs queued for processing: 1 (0%)
URLs completed: 25121 (100%)

Reload topics from topics.xml file

Current topics
Click on a topic title to edit the topic definition.

- Create new top-level topic
- Create new top-level cross-reference
- Import subtopics at top level
- Change top-level name and blurb
- Export topics as Verity TAX file

Format: $<$title> $<$level> <id>

- htxtA
- htxA1
- htxtB
- htxB1
Sample CCE Topics Pages

This section provides you with examples of the different ways you can create and display topics pages on your site by setting CCE parameters. These are just a few examples of the many layouts the parameter page provides.

**Figure 11-1** shows a regular, two-column, listing of CCE topics.

**Figure 11-2** CCE Topics

- **Search:**
  - Collections
  - Search the Internet

- **Categories:**
  - Danish
  - German
  - English
  - Spanish
  - Finnish
  - French
  - Italian
  - Dutch
  - Norwegian
  - Portuguese
  - Swedish

- **Topics:**
  - **Books**
    - Mystery, Romance, Travel
  - **Home Hobbies**
    - Cooking, Gardening, Sewing/Quilting
  - **Movies**
    - Action, Adventure, Children, Foreign Films
  - **Music**
    - Classical, Country, Popular, Rock 'n' Roll
  - **Pets**
    - Birds, Cats, Dogs, Fish
  - **Sports**
    - Baseball, Basketball, Football, Soccer, Tennis
  - **Television**
    - Comedies, News, Soap Operas, Talk Shows

---
Figure 11-3 shows a three-column view of CCE topics.

**Figure 11-3 Three-Column View**

Tip: Search terms are stemmed.

Example: network matches networks as networked.

Search: the collections

- Arthima
- Northwind Database
- File System
- Heroic Age Publication
- Labyrinth
- Microsoft Exchange
- Nicene Fathers
- Old English

- the Internet

<table>
<thead>
<tr>
<th>Search</th>
<th>Help</th>
<th>Advanced</th>
</tr>
</thead>
</table>

Languages: German - English - Spanish - French - Italian - Dutch - Norwegian - Portuguese - Swedish

Expanded all topics

Topics:

**Books**
- Mystery, Romance, Travel

**Home Hobbies**
- Cooking, Gardening, Sewing, Quilting

**Movies**
- Action, Adventure, Children, Foreign Films

**Music**
- Classical, Country, Popular, Rock'n Roll

**Pets**
- Birds, Cats, Dogs, Fish

**Sports**
- Baseball, Basketball, Football, Soccer, Tennis

**Television**
- Comedy, News, Soap Opera, Talk Show
Figure 11-4 shows search results grouped by topic:

**Figure 11-4** Results Grouped by Topic

**Books**

- *Pride and Prejudice (Eng/Chi)*
  Bilingual English/Chinese text of Jane Austen's novel *Pride and Prejudice* (1813). Texts are side by side for easy comparison and language study.

- http://shoophoria.lakwml.com/81/jane_austen/PRIDPREJ.TXT
  This e-text is in the public domain, and has been corrected against the 1925 R.W. Chapman edition, with slight punctuation modernization, by churchb@us.eastwes.edu. Chapman's chronology and...
  - http://shoophoria.lakwml.com/81/jane_austen/PRIDPREJ.TXT - 70.7KB - Books

- *Pride & Prejudice Chapter 18 (Vol. I, Chap. XVIII)*
Figure 11-5 Results with Inline Topic Path

Start new search. Search these results. Search entire Web.

Search EDB:

search Help Advanced


Topic: Home > EDB

12 results found, sorted by relevance.

EDB Login
Login to the Editorial DB. Have you registered before? Username: Password: You can login as user guest but you will be unable to insert or delete items in the EDB. Otherwise, register here ... http://internal.sitel.com/cachery/EDB/View/ - 1.0KB - Scan's Internal

EDB Schema
In Oracle, all tables contain a pseudocolumn called rowid. Doing a "select rowid" on a table will return the rowid of all the rows in the table in string form. Hence, the rowid column is not included in the table ... http://internal.sitel.com/cachery/EDB/EDBSchema.htm - 5.2KB - Scan's Internal

EDB
Detailed schema is here. Integrate current lists into EDB. Transfer the current repositories on juggling and insert to the new db. Interface Tools Development: three weeks: Develop the basic server side interface for adding and downloading ... http://internal.sitel.com/cachery/EDB/EDB.htm - 15.8KB - Scan's Internal

EDB Schedule
1. Initial Preparation Scrub existing metadata Install/Setup new EDB server Incorporate database schema Basic EDB Access Integrate/Migrate data to new EDB Scan/Ops Scan/Ops ... http://internal.sitel.com/cachery/EDB/EDBSchedule.htm - 4.7KB - Scan's Internal
Before you get started, you can format where and how your topics display. Use the Interface | Topics page to define the layout and content of queries and topics, including the type of search results you want to display and how to format and organize search results for topics. Alternatively, you can accept the default layout.

To set CCE display parameters:

1. Select Interface.
2. Select the style that you want to edit in the Style list box on the Style page.
3. Select the Topics tab.

Figure 11-6 shows the Interface | Topics page, and the following table describes each CCE display parameter. See “Sample CCE Topics Pages” on page 249 for examples of how these settings can change the look and feel of your topics pages.

**Figure 11-6  CCE Topic Parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Parameters for style &quot;Simple&quot;</td>
<td></td>
</tr>
<tr>
<td>CCE Initial Page</td>
<td></td>
</tr>
<tr>
<td>Home topic</td>
<td>Default</td>
</tr>
<tr>
<td>Number of topics columns</td>
<td>3</td>
</tr>
<tr>
<td>Show query form above the topics</td>
<td>✔</td>
</tr>
<tr>
<td>Show query form below the topics</td>
<td></td>
</tr>
<tr>
<td>Maximum number of topics children</td>
<td>3</td>
</tr>
<tr>
<td>Indent topics children</td>
<td>✔</td>
</tr>
<tr>
<td>Show all topics on search home page</td>
<td></td>
</tr>
<tr>
<td>CCE Topic Browse Pages</td>
<td></td>
</tr>
<tr>
<td>Number of subtopics columns</td>
<td>3</td>
</tr>
<tr>
<td>Show subtopics above the hits</td>
<td>✔</td>
</tr>
<tr>
<td>Show subtopics below the hits</td>
<td></td>
</tr>
<tr>
<td>Maximum number of subtopics children</td>
<td>3</td>
</tr>
</tbody>
</table>
CCE Initial Page

These settings affect the format and appearance of the initial topics page that displays to users.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Home topic</strong></td>
<td>Specifies the top-level topic to use for the selected interface style. For example, you can design the interface so that the Human Resources department only sees the “HR” topic tree, and the Engineering department only sees the “Development” topic tree.</td>
</tr>
<tr>
<td><strong>Number of topics columns</strong></td>
<td>Specifies the number of topic columns to display on the initial CCE page.</td>
</tr>
<tr>
<td><strong>Show query form above or below the topics</strong></td>
<td>Controls whether the search box appears above topics, below topics, or both.</td>
</tr>
<tr>
<td><strong>Maximum number of topics children</strong></td>
<td>Specifies the maximum number of subtopics to display in the search interface. A topic’s children are those subtopics which are one level below the active topic.</td>
</tr>
<tr>
<td><strong>Indent topics children</strong></td>
<td>Controls whether the subtopics are indented. Indenting subtopics makes them stand out, making the interface more intuitive for users.</td>
</tr>
<tr>
<td><strong>Show all topics on search home page</strong></td>
<td>Determines whether or not all topics are displayed on the initial search page.</td>
</tr>
</tbody>
</table>
CCE Topic Browse Pages

These settings affect the look of all topic browsing pages subsequent to the initial page when users click through initial CCE topics.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subtopics columns</td>
<td>Specifies the number of topic columns that display on the CCE topic browse pages.</td>
</tr>
<tr>
<td>Show subtopics above or below the search results</td>
<td>Controls whether subtopics display above the search results, below the search results, or both.</td>
</tr>
<tr>
<td>Maximum number of subtopics children</td>
<td>Specifies the maximum number of subtopics to display in the search interface. A topic’s children are those subtopics which are one level below the active topic.</td>
</tr>
<tr>
<td>Indent topics children</td>
<td>Controls whether the subtopics are indented. Indenting subtopics makes them stand out, making the interface more intuitive for users.</td>
</tr>
<tr>
<td>Show query form above or below the search results</td>
<td>Controls whether the search box appears above the search results, below the search results, or both.</td>
</tr>
</tbody>
</table>

CCE Search Pages

These settings affect the look of the pages that have both topics and search results. Related topics appear after a user has performed a query and suggest topics related to that query. Related topics children are subtopics that are one level below the active related topic.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of related topics columns</td>
<td>Specifies the number of topic columns to show on the CCE search pages.</td>
</tr>
<tr>
<td>Show related topics above or below the search results</td>
<td>Controls whether the subtopics appear above the search results, below the search results, or both.</td>
</tr>
<tr>
<td>Maximum number of related topics children</td>
<td>Specifies the maximum number of related topics children to display in the search interface.</td>
</tr>
<tr>
<td>Indent related topics children</td>
<td>Controls whether the related topics children are indented. Indenting related topics makes them stand out, making the interface more intuitive for users.</td>
</tr>
<tr>
<td>Maximum number of related topics</td>
<td>Specifies the maximum number of related topics to show in the search interface.</td>
</tr>
</tbody>
</table>
You can customize the look and feel and search functionality of the CCE topics, subtopics, and related topics pages in the search interface. Refer to the *Verity Ultraseek Customization Guide* for more information.
Creating and Editing Topics

Once you’ve created a collection, you can create and edit topics. Using the topics page, you set up a CCE topic hierarchy by creating a new top-level topic or by importing an existing set of topics. You can also create cross-references, edit topic blurbs, and edit existing topic hierarchies. To display the topics page, follow the procedure explained in “Starting CCE” on page 247.

Once you’ve created your topic hierarchy, you can edit the topics in it. You can view topics and browse them as your users will, delete a topic, and create subtopics and cross-references. You can also specify more detail about how to display a topic, such as adding a blurb and choosing which topics should appear in the category.

This chapter contains the following sections:

- Creating Topics
- Creating Cross-References
- Importing Topics
- Changing the Top-Level Name and Blurb
- Editing Topics
- Understanding the Topic Definition File

Note  CCE categorizes documents and indexes them in the same manner as Ultraseek. Before you create and test a topic hierarchy you should first create an Ultraseek collection with filters that include all the documents you want to categorize.
Creating Topics

You can set up a CCE topic hierarchy by creating a new top-level topic. You can also set one up by importing an existing set of topics, as explained in “Importing Topics” on page 261.

To create new topics:

1. From the Topics page (“Starting CCE” on page 247), click Create new top-level topic. The Create New Topic page displays; this is the main page that you use to create new topics. You may want to supply a name. If you are creating a super topic, you may not want to enter any rules, because super topics inherit documents from their subtopics.

Figure 12-1  Creating New Topics

2. Enter a name for the topic. It should be as short as possible for display purposes; however, it can be up to 256 characters long. Enter characters as you want them to appear, not as HTML syntax. For example, if you want to use an ampersand, use the ampersand character (&), not the HTML ampersand syntax (&amp;).

3. Fill in the Blurb field (optional). This free-form, HTML-formatted text area lets you add additional information about the topic to the end user display. This is interpreted as HTML source, so you can enter HTML syntax for any text. For example, if you want to use an ampersand, use the HTML ampersand syntax (&amp;) not the ampersand character (&).
The benefit of this field is that you can reference specific sites with `<href>` tags, or include images with `<img>` tags. Relative URLs are resolved starting from the Ultraseek /docs directory.

Put the images you intend to use in the root directory for Ultraseek (C:\Program Files\Verity\Ultraseek\docs), or use absolute URLs to map to their location on another server.

For example, you may want your HR topic site to contain the blurb:

```html
<b>Remember, Open Enrollment Ends January 31!</b>
```

Or your public site may say:

```html
<p>Read what the <a href="http://www.wsj.com">Journal</a> wrote about our new product!</p>
```

In addition to standard text, all valid HTML is acceptable. Text is displayed in a `<td>` element, so there is no need to start or end with a line break `<br>` tag.

4. Define keywords in the **Keywords** field (optional). By defining useful keywords for a topic, you ensure that it is shown as a related topic in a query. Separate multiple keywords with commas.

   For example, you might want to list your Human Resources topic as a related topic whenever a user searches for “benefits.” If you put the word “benefits” in the keyword field, you can ensure the Human Resources topic is always found.

5. Assign URLs to a topic and rank them in order of importance (optional). You can do this by specifying a particular URL and assigning a star number, from 0 to 3, as shown below.

   ![Figure 12-2 Assigning URLs to a Topic](image)

6. You can also define rules for a topic. Rules are explained in Chapter 13.

7. Click **ok**.
Creating Cross-References

Cross-references are topics that point to other, related categories. You may want to create them to give your end users multiple entry points to the same information. For example, a subtopic of Roses could appear under two different parent topics: Gardening and Flowers.

To create a cross-reference:

1. From the Topics page, click Create new top-level cross-reference.

   A cross-reference topic page displays, as shown below.

   Figure 12-3 Creating Cross-References

   ![Create New Topic]

2. Enter a name in the Name field and click Choose new target.

   A list of all topics in the hierarchy appears. If the list of topics is long, you may need to scroll to the bottom of the list of topics to locate the topic you want to cross-reference.

3. Click on the topic or topics to which you want this topic to be cross-referenced, and click ok.

   The Create New Topic page appears again, reflecting the target to which the cross-reference topic is aimed.

4. If you are satisfied with the target, click ok. If you are not, click cancel.
Importing Topics

Site maps are often organized into categories, such as Products, News, Support, Jobs, and About. If this organization resembles your planned topic hierarchy, you can import it into CCE as topics.

The document to which you point is scanned for links (\texttt{<a>} and \texttt{<area>} tags), and each link is converted to a new topic. The target of each \texttt{<a>} tag (the \texttt{href} attribute) is converted into a \texttt{url:} or \texttt{site:} rule, and the text enclosed by the tag becomes the topic name. In \texttt{<area>} tags, the text in the \texttt{alt} attribute is used as the name. See Chapter 13 for more information.

Topics are not created for \texttt{<a>} tags that have no associated text (for example, the tags might enclose an image), or for URLs with protocols other than HTTP.

\textbf{Import Topics} automatically creates topics with the Topic Name the same as the link name, and the rule is based on the URL to which the HREF points.

\textbf{To import a topic list:}

1. From the Topics page, click \textbf{Import subtopics at top level}. The Import Topics page displays.

2. Enter the URL of the file that contains the topic list. For example:

   \url{http://www.pausd.palo-alto.ca.us/district/schools.html}

3. Click \textbf{Import Topics}. The topics that reside within the URL you entered in the Import Topics window are spidered.

4. CCE displays a page listing the topics that are to be imported, along with the derived rules. A check box appears next to each topic name. If the box is checked, that topic is created. If the check box is cleared, that topic is not created. For example, a page may contain several links that do not belong in the hierarchy, such as “Home,” “Back,” or “copyright.” To create a topic, click its check box. To exclude a topic, clear its check box.

5. Click \textbf{ok} to create the topics you have selected, or \textbf{cancel} to return to the Topics Status page.

   While the topics are created, your browser is returned to the Topics Status page, where you can track the topic creation process. Click the top banner to refresh the status page. When the status is “ready,” the import process is complete.

   When spidering is complete, the names of the pages belonging to the topics you selected display. If you encounter an error fetching the URL, the page with the error (“404”) displays.
Changing the Top-Level Name and Blurb

At any point, you can change the name and blurb you assigned to your top-level topic. You might do this if, for example, you want to expand the topic’s meaning or use the blurb to share new material. The change does not appear in the default interface, but will feed through to any alternative views.

To change the top-level name and blurb:

1. From the Topics page, click Change top-level name and blurb. The Edit Topic page displays.

2. Enter the new text you want in the Name field.

3. Enter the new text you want in the Blurb field.

4. Enter the keywords that you want to associate with this topic.

5. Click ok.
Export Topics as Verity TAX File

At any point, you can export your topics as a Verity TAX file, allowing for import into other compatible Verity products, such as Verity Collaborative Classifier.

To export topics:

1. From the Topics page, click Export topics as Verity TAX file.
2. When prompted, allow the specified TAX file to be opened in the application selected, or save the TAX file to disk.

Editing Topics

After you create or import topics for your hierarchy, you may want to change some portion of their definition or their position relative to other topics. For example, when you import topics from an existing page, all topics appear at the same level. You can edit topics to place them beneath their appropriate super topic.

You might also want to change topic names. The names generated for topics match the URL title to which they link. Link names tend to be long and descriptive, but topic names should be concise. For example, the link name “Planning and Community Environment Office” could be shortened to “Planning.”

You edit the topic information using the Edit Topic page. To display this page, click on the topic for which you want the status, definition, or rules. The Edit Topic page is similar to the Create Topic page. For more information on fields on this page, see “Creating Topics” on page 258.
Figure 12-5  Editing Topics

The links at the top of the Edit Topic page allow you to perform the following actions:

- Browse a Topic
- Delete a Topic
- Create a Subtopic
- Create a Cross-reference Subtopic
- Import Subtopics
- Make a Copy of a Topic
- Create New Super Topic
Browse a Topic

Once you’ve created a topic, you may want to review its results as they will be seen by users. You can do this by clicking Browse topic. CCE displays the results of the current state of the spidering of the topic list. As results are added to a topic, they are displayed; if you browse the topic before it completes spidering, the results you see may be incomplete. Browse topic does not predict results.

Delete a Topic

If you want to delete a topic you are viewing, click Delete. A confirmation window appears, asking if you are sure you want to delete the topic and all its subtopics. Click ok to delete, or cancel to return to the editing page without deleting the topic. After you choose to delete a topic, all documents in that topic are re-indexed.

IMPORTANT Be careful when using Delete. When you delete a top-level topic, all of its subtopics are also deleted.

Create a Subtopic

When you click Create subtopic, the Create New Topic page opens with the supertopic assigned. This is the same page you used to create topics, as explained in “Creating Topics” on page 258.

Create a Cross-reference Subtopic

You can create cross-references to subtopics by clicking Create cross-reference subtopic in the existing subtopic’s Edit Topic page. The Import Topics page appears with the super topic identified. This is the same page you used to create cross-references, as explained in “Creating Cross-References” on page 260.

Import Subtopics

You can import topics to reside beneath an existing subtopic by clicking Import subtopics in the existing subtopic’s Edit Topic page. The Import Topics page appears with the super topic identified. This is the same page you used to import your topic lists at the top level, as explained in “Importing Topics” on page 261.
Make a Copy of a Topic

You might want to create a copy of a topic if you plan to create another one with a similar structure; by making a copy, you can save typing effort by entering only the modifications. To create a copy of a topic, click Make a copy in the existing subtopic’s Edit Topic page. The Create New Topic page appears with the fields filled in for the topic you are copying. This is the same page you used to create your topics at the top level, as explained in “Creating Topics” on page 258.

Create New Super Topic

Sometimes you are not satisfied with the topic hierarchy. The super topic field provides you with the ability to modify a topic’s position in the hierarchy. When you first access the Edit page, the super topic field shows the parent topic. Click the Create new super topic button to choose a new super topic for the topic.

Understanding the Topic Definition File

When you complete your topics definition and modification process, the definitions reside in a file called topics.xml. You should only modify topic definitions from within CCE, but if for any reason you need to modify the topics.xml file, you can reload it within CCE by clicking Reload in the Edit Topic page.

For example, you might have a set of topics that are managed in an existing database or application. If you want to edit topics within that application instead of through CCE, you should write the topics.xml file from your application’s data and then do an HTTP request to the page that reloads the topics.xml file. When the file is reloaded, changes are recognized and handled as if they are made through the CCE.

You might also find the reload capability useful if you want to revert to older copies of the topics tree, or if you want to update a new set of topics from a staging server to a production server.

The HTTP request is a GET of this URL: /admin/topicsreloadgo.html?ok=+++ok+++ The request must also include authentication (login) information in the headers. See the Verity Ultraseek Customization Guide for more information on performing administrative operations with scripts.
Creating Topic Rules

Topic rules are the foundation of CCE because they determine which documents are included or excluded from topics. By setting up CCE to autoclassify documents, you reduce time in both creating and maintaining your hierarchy. Ultraseek applies your rules when it indexes new content.

CCE provides an interface to create rules by specifying document search parameters, which resemble a normal document search, and field rules, such as URLs and document titles.

This chapter describes how to set up and test rules. It contains the following sections:

- Getting Started with Rules
- Specifying Rules
- Testing Rules
- Sample Rules
You use the Create Topic or Edit Topic page to specify rules. To display the Edit Topic page for an existing topic, follow the procedure explained in “Starting CCE” on page 247, then click on a topic name. The bottom half of the Edit Topic page contains fields you fill in to specify rules, following the guidelines explained in this chapter.

**Figure 13-1** Create Rules Region

![Create Rules Region](image)

By default, each topic contains two sets of rule fields. If you need more fields for a given topic, click the **Show more rule and URL entries** button on the upper section of the page. If you run out of entries again, click for more entries again.

### Specifying Rules

A topic is defined by Rules 0 through n. All indexed content satisfying any rule is included in a topic. For example, if document A satisfies Rule 1 or Rule 2, and document B satisfies Rule 3, then both document A and document B appear in the results for the topic, even though document B does not satisfy Rule 1 or 2, and document A does not satisfy Rule 3. The requirement is a document must satisfy any one rule to be included in the topic.
Rule Terms

Within each topic can be multiple rules, and each rule can contain an unlimited number of terms. A term consists of a “Must Contain” or “Must Not Contain” qualifier, and a value, such as:

url:http://www.corp/finance/

Values can include locations, sites, META tags, title specifications, raw query terms, or document type. Within each rule you can combine the usage of terms to obtain the desired content.

Since topics inherit rules from subtopics, it may not be necessary to define rules for higher-level topics.

Rule Syntax

When a field qualifier is a phrase, the phrase must be enclosed within quotation marks:

- Legal syntax: title:“year 2000”
- Illegal syntax: title:year 2000

If you specify a phrase without quotation marks, Ultraseek inserts quotes for you within the document itself.

To set up rules:

1. In the edit page, select the topic you want to work with, and double click to launch it.
2. Enter each rule in its field.
3. Click ok. This sets off the categorization process.
4. To see your topics and their associated documents, go to the search interface at:
   
   http://host:adminport/admin/

   Where host is the name of the host where Ultraseek is installed and adminport is the number assigned to the administrative and search interface port (the default value for adminport is 8765).
Field Rules

A field search term is a rule that looks for certain elements within a document. These elements may be in metadata, or in other elements of the document.

The order in which you define rules establishes their document order. For example, if you know you want the 401(k) benefits page to be first within the HR Benefits rule, specify that page’s title or URL rule first.

The field rules available to you are listed below, in the order of their usefulness, and explained on the following pages:

- `url`: (Specifying Rules by URL)
- `urlexact`: (Specifying Rules by Exact URL)
- `site`: (Specifying Rules by Site)
- `collection`: (Specifying Rules by Collection)
- `title`: (Specifying Rules by Document Title)
- `raw terms`: (Specifying Rules by Raw Terms)
- `doctype`: (Specifying Rules by Document Type)
- `metadata`: (Specifying Rules by Metadata)

The format for field rules is

`field:phrase`

where `field` is the document’s element, and `phrase` is the search phrase within that element.

A title rule would be a word or phrase, enclosed in quotation marks, and could look like this:

`title: “new years resolution”`

In this example, an administrator would be looking for documents with the words “New Years Resolution” in the title.

Fields should be specified in lowercase so they will match any combination of case in the documents being indexed.
Specifying Rules by URL

You can include documents within a given URL. The URL can include any subdirectories. This enables you to create a broad topic:

url:software.verity.com

or a more narrowed one:

url:software.verity.com/products/ultraseek/faqs

You would get the same results if you typed in a more complete version of the same URL rule:

url:http://software.verity.com/products/ultraseek/faqs

These rules yield all documents within the specified directories. However, you can also define rules for a particular word as part of the URL.

For example, the rule:

url:faqs

returns the same documents as the previous rule, and also returns documents from other directories that contain the word faqs in the URL.

Specifying Rules by Exact URL

You can create rules for an exact document. Unlike rules by URL, this rule allows only the exact page or screen listed.

For example, the rule:


returns only that page.

Specifying Rules by Site

You can create site rules based on any site. Unlike rules by URL, this rule is limited to the host name portion of a URL and suffixes of the host name. For example, the rule,

site:sun.com

yields results for www.sun.com and java.sun.com, but not sun.co.uk.

Use rules by site to list all pages in a given site.
**Specifying Rules by Collection**

You can set a rule to include or exclude a particular collection by typing the internal name of the collection as your rule:

```
collection: collection_name
```

**Specifying Rules by Document Title**

You specify a rule for an HTML document’s title by typing all or part of the title as your rule:

```
title:year
```

or:

```
title:“year 2000”
```

The first rule yields a broad topic: all titles that contain the word “year.” The second rule yields a more narrow topic: all titles that contain the phrase “year 2000.”

However, if you wish to ensure that a specific phrase is found within a title field search, enclose the rule in quotes:

```
title:“year 2000 compliance”
```

This example yields the most narrowed pool of results.

**Specifying Rules by Raw Terms**

You can specify raw terms that either force a document into or out of your topic. To do so, simply specify the term the document needs to include (or exclude in case you are using the “must not contain” option) in order to match the rule. For example, the rule:

```
sun
```

matches documents that contain the word “sun” in their body text.
Specifying Rules by Document Type

The document type rule specifies a document’s form, not its content. Its usefulness is fairly limited, unless you only want to search within documents that have been saved as PDF format, for example.

To match a specific document type, use the `doctype` field rule. For example, to display only Adobe Acrobat documents, use a rule:

```plaintext
doctype:application/pdf
```

The document type specifies the MIME type of the document to be found, and is available only if your web server’s configuration allows it.

The table below lists the common, officially-registered MIME types that Ultraseek accepts. For a detailed list of all supported MIME types, check the Doc Types tab under the Server button in the Administration interface.

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Mime Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe Acrobat</td>
<td>application/pdf</td>
</tr>
<tr>
<td>Adobe PostScript</td>
<td>application/postscript</td>
</tr>
<tr>
<td>Autocad Interchange</td>
<td>application/x-dxf</td>
</tr>
<tr>
<td>Corel WordPerfect</td>
<td>application/wordperfect5.1</td>
</tr>
<tr>
<td>HTML</td>
<td>text/html</td>
</tr>
<tr>
<td>Ichitaro</td>
<td>application/x-js-taro</td>
</tr>
<tr>
<td>Lotus 1-2-3</td>
<td>application/vnd.lotus-1-2-3</td>
</tr>
<tr>
<td>Lotus Freelance</td>
<td>application/vnd.lotus-freelance</td>
</tr>
<tr>
<td>Lotus WordPro</td>
<td>application/vnd.lotus-wordpro</td>
</tr>
<tr>
<td>Microsoft Excel</td>
<td>application/vnd.ms-excel</td>
</tr>
<tr>
<td>Microsoft Powerpoint</td>
<td>application/vnd.ms-powerpoint</td>
</tr>
<tr>
<td>Microsoft Word</td>
<td>application/msword</td>
</tr>
<tr>
<td>Multipart</td>
<td>multipart/alternative</td>
</tr>
<tr>
<td></td>
<td>multipart/digest</td>
</tr>
<tr>
<td></td>
<td>multipart/mixed</td>
</tr>
<tr>
<td></td>
<td>multipart/parallel</td>
</tr>
<tr>
<td></td>
<td>multipart/related</td>
</tr>
<tr>
<td></td>
<td>multipart/signed</td>
</tr>
</tbody>
</table>
The variation of additional MIME types depends on the configuration of your Web server, and this affects how you compose searches.

### Specifying Rules by Metadata

You can also define Dublin Core Subject terms in `META` tags. For example, if you defined a `META` tag:

```html
<meta name="dc.subject" content="Water Quality"/>
```

You could then define rules such as

```
dc.subject:"Water Quality"
```

to populate the topic.

You can use the `meta` HTML tag to customize field rules.

For example, to create a topic that lists books by author, embed the following `META` tag into the HTML:

```html
<meta name="Author" content="James Fenimore Cooper"/>
```

You could then define rules such as

```
Author:"James Fenimore Cooper"
```
Testing Rules

You may want to test your rules before applying them to determine whether they perform the correct topic selections. Testing rules allows you to see how many documents match each rule you’ve defined. If you are not satisfied with the results, edit the rule and test it again.

To test rules, click the **Test Rules** button on the Edit Topic page.

**Figure 13-2** Test Rules Page

<table>
<thead>
<tr>
<th>Testing Rules for &quot;Arts and Culture&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>These are the number of documents at this server that match each rule. A document may match multiple rules.</td>
</tr>
</tbody>
</table>

**If you are editing rules**, these results are for the **changed** rules, not the currently installed rules.

**Document count for topic**: 0 (before edits)

<table>
<thead>
<tr>
<th>Rule 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>arts and culture</td>
</tr>
</tbody>
</table>

**Document count**: 4

Show matching documents

The document counts in Test Rules may be larger than the current topic document count. Some reasons this may occur include:

- You may not have saved the rules.
- Duplicates in other collections may not have been removed for this count.
- Documents may not have been re-indexed yet.
Sample Rules

Sample #1

Include all indexed content in the HR 401k folder.

Rule 0
Must Contain url:http://www.achp.corp/hr/401k/

Sample #2

Include all indexed content on two different sites:

Rule 0
Must Contain site:www.achp.corp
Rule 1
Must Contain site:www.achp.eng

Sample #3

Include all indexed content in the finance folder, except those with the word “Events” in the title.

Rule 0
Must Contain url:http://www.achp.corp/finance/
Must Not Contain title:“Events”

Sample #4

Include content with a custom META tag defined.

For example, if the HTML Contained <META NAME="SBU" CONTENT="Buena Vista Home Entertainment">, then the rule would be:

Rule 0
Must Contain SBU:”Buena Vista Home Entertainment”

Sample #5: Include content with the following Dublin Core META tag:

<meta name="dc.subject" content="Environmental research -- Minnesota ">

The following rule applies:

Rule 0
Must Contain dc.subject:”Environmental research -- Minnesota”
Sample #6: Include all content in the Home Video section, with the term “Lion King” and “Return to Pride Rock”, but not with the following META tag:

<meta name="Status" content="not-released"/>

The following rule applies:

Rule 0
Must Contain url:http://www.disney.com/HomeVideo/
Must Contain “Lion King”
Must Contain “Return to Pride Rock”
Must Not Contain Status:”not-released”

Sample #7: Include three specific documents.

Rule 0
Must Contain urlexact:www.rebis.com/information/index.html
Rule 1
Must Contain urlexact:www.rebis.com/press/index.html
Rule 2
Must Contain urlexact:www.rebis.com/press_releases/010197.html
13 Creating Topic Rules
Sample Rules
Viewing Reports

CCE allows you to generate reports that you can use to monitor the viability of your topic list. These reports summarize information about topics or queries. Many of the reports allow you to edit a topic by clicking on the topic name link, providing you with a convenient place to maintain topics, check for empty topics, and analyze topics with too few or too many documents.

The following sections explain each report:

- Getting Started with Reports
- Viewing Documents per Topic, Sorted by Hierarchy
- Viewing Documents per Topic, Sorted by Population
- Viewing Queries, Sorted by Frequency
- Viewing Topics, Sorted by Browse Frequency
- Viewing Query Terms, Sorted by Frequency
You use the Reports page to choose the report to view. To display the Reports page, follow the procedure explained in “Starting CCE” on page 247, then click the Reports tab.

**Figure 14-1** Reports Page

Links on the page display the available reports. The following pages describe each report in detail.
Viewing Documents per Topic, Sorted by Hierarchy

The Documents per Topic report tells you how many documents reside in each topic, and sorts the results by hierarchy.

You can use this report to determine the balance within your topic hierarchy. You can also determine if some topics have no documents in them, and then redefine your search rules.

Each topic name in the report is linked to the topic’s corresponding Edit Topic page. To edit a topic, click on the corresponding topic link.

**Figure 14-2** Viewing Documents Sorted by Hierarchy

<table>
<thead>
<tr>
<th>topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents per Topic</td>
</tr>
<tr>
<td>0 Total documents</td>
</tr>
<tr>
<td>75 Total documents in topics</td>
</tr>
<tr>
<td>75 Arts and Culture</td>
</tr>
<tr>
<td>10 Books</td>
</tr>
<tr>
<td>23 Dance</td>
</tr>
<tr>
<td>33 Music</td>
</tr>
<tr>
<td>26 Theater</td>
</tr>
<tr>
<td>3 Science and History</td>
</tr>
</tbody>
</table>
The Topics by Population report tells you how many documents are contained within each topic, and sorts the results by population.

You can use this to determine the relative size of various topics. You can also determine if some topics have no documents in them, and thus redefine your search rules.

Each topic name in the report is linked to the topic’s corresponding Edit Topic page. To edit a topic, click on the corresponding topic link.

**Figure 14-3** Viewing Topics by Population
Viewing Queries, Sorted by Frequency

The Queries by Frequency report lists all queries made to Ultraseek and sorts them by the number of times the queries have been made. You can use this report to see which queries are repeatedly made, and consider creating a CCE topic to encompass the queries. Each query in the report is a link that performs that query search, which allows you to see the results obtained for that query.

This report shows raw queries, including the terms that are used for browsing topics. See “Viewing Query Terms, Sorted by Frequency” on page 285 for queries with the topics browse terms filtered out.

**Figure 14-4** Viewing Queries Sorted by Frequency

---

```
queries

Queries by Frequency

Results from past 2 days

<table>
<thead>
<tr>
<th>query</th>
</tr>
</thead>
<tbody>
<tr>
<td>27 Total queries</td>
</tr>
<tr>
<td>3 +topic:280942060, topic:280942060-0, topic:280942060-1, topic:280942060-2, topic:280942060-3</td>
</tr>
<tr>
<td>2 +topic:1212375062, New name</td>
</tr>
<tr>
<td>2 modern dance</td>
</tr>
<tr>
<td>1 +arts , +culture ,</td>
</tr>
<tr>
<td>1 +inktom1 , +coe , +rule ,</td>
</tr>
<tr>
<td>1 +ontology , -taxonomy ,</td>
</tr>
<tr>
<td>1 +topic:1212375062, New Product</td>
</tr>
<tr>
<td>1 +topic:280942060, CCE</td>
</tr>
<tr>
<td>1 ballet</td>
</tr>
<tr>
<td>1 books</td>
</tr>
<tr>
<td>1 modern dance</td>
</tr>
</tbody>
</table>
Viewing Topics, Sorted by Browse Frequency

The Topic Visits by Frequency report tells you how many times each topic has been browsed, and lists the topics by browse frequency.

You can use this to determine the popularity of parts of the hierarchy you’ve designed.

Each topic name in the report is linked to the topic’s corresponding Edit Topic page. To edit a topic, click on the corresponding topic link.

Figure 14-5 Viewing Topics Sorted by Browse Frequency

<table>
<thead>
<tr>
<th>topics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic Visits by Frequency</strong></td>
</tr>
<tr>
<td>Results from past 2 days</td>
</tr>
<tr>
<td>22 Total visits</td>
</tr>
<tr>
<td>11 Total topic browses</td>
</tr>
<tr>
<td>4 Arts and Culture &gt; Books</td>
</tr>
<tr>
<td>3 Arts and Culture &gt; Music</td>
</tr>
<tr>
<td>2 Arts and Culture</td>
</tr>
<tr>
<td>1 Arts and Culture &gt; Dance</td>
</tr>
<tr>
<td>1 Science and History</td>
</tr>
</tbody>
</table>
Viewing Query Terms, Sorted by Frequency

The Query Terms by Frequency report shows you which terms users have been searching for the most often and sorts them by the number of times the queries have been made.

You can use this report to see which queries are repeatedly made, and consider creating topics containing those terms to make them easier to reach using top-level links or cross-references. Each query in the report is a link that performs that query search, which allows you to see the results obtained for that query.

This report shows terms only, not raw queries, which include the terms that are used for browsing topics. See “Viewing Queries, Sorted by Frequency” on page 283 for queries with the topics browse terms.

Figure 14-6  Viewing Query Terms, Sorted by Frequency

<table>
<thead>
<tr>
<th>queries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query Terms by Frequency</strong></td>
</tr>
<tr>
<td>Results from past 2 days</td>
</tr>
<tr>
<td>2 +taxonomy, ~ontology,</td>
</tr>
<tr>
<td>2 modern dance</td>
</tr>
<tr>
<td>1 ballet</td>
</tr>
<tr>
<td>1 +arts, +culture,</td>
</tr>
<tr>
<td>1 +linkto:ml, +oce, +rule,</td>
</tr>
<tr>
<td>1 theater</td>
</tr>
<tr>
<td>1 +ontology, ~ontology,</td>
</tr>
<tr>
<td>1 books</td>
</tr>
<tr>
<td>1 modern dance</td>
</tr>
</tbody>
</table>
14  Viewing Reports
Viewing Query Terms, Sorted by Frequency
Mirroring

Mirroring allows you to create a copy of your collection and topic hierarchy on another server. Mirror collections are used at large sites to separate the spidering and query-serving functions for more throughput, to secure data, and to allow redundant query-serving hosts.

In order to mirror topics, you must license both servers for CCE. See “Obtaining a CCE License Key” on page 247 for information on getting a license.

The Verity Ultraseek Administrator Guide contains information on mirroring. This chapter explains mirroring issues as they apply to topics and contains the following sections:

- Enabling and Disabling Topic Mirroring
- Multiple Host Issues
Enabling and Disabling Topic Mirroring

Note that you cannot create, edit, or delete topics on the mirrored system; you can only do so on the master system.

**Enabling Topic Mirroring**

The configuration parameters for the mirror collection include a check box to enable mirroring the topics along with the collection. If that box is checked, the topics will be checked for changes whenever the mirror collection polls the remote server.

If a host has several mirror collections, you may want to check the topics mirroring box on all mirror collections. The topics are only updated when they change, regardless of how many collections are polling them. Checking the box on all collections ensures that topics continue to be mirrored if some of the collections are deleted.

If two or more spidering systems are configured, they will need to share the same topics definitions. One system will be the master for the topics, and the others will mirror the topics. To do this, mirror one collection from the master Ultraseek/CCE host to the other spidering servers. This may be an existing collection, or a new collection may be created specifically to carry the topics (perhaps a file scanning collection with one document).

When the mirror collection copies an updated set of topics, the current set and the new set are compared to detect any changes. When changes are detected, the changed rules cause matching URLs to be queued in any local spidering or file scanning collections.

**Disabling Topic Mirroring**

To disable topic mirroring, un-check the mirroring box on all collections. The local copy of topics is removed from the system. Topics are no longer shown on the user interface.
Multiple Host Issues

Mirroring with Multiple Spidering Hosts

In a system with multiple spidering hosts and one or more query-serving hosts (hosts with only mirror collections) it is safe to mirror topics on all collections. There is no need to limit topic mirroring to just the host with the master copy of the topics definitions. The mirroring protocol will always use the most recently edited copy of the topics definitions.

On a host other than the topics master, the Topics Status page will indicate that the topics are mirrored. Since the topics cannot be edited locally, the edit links point to the Edit Topics pages on the topics master host. Other links, like the Reports pages, are evaluated on the local system. This allows the administrator to run reports against all collections, then update topics on the master host, all without juggling multiple browser windows.

Mirroring on a Multiple-Host System

On a multiple-host system, there will be some additional delays between editing the topics and seeing the results in reports. The edits on the master system will queue URLs locally, but other spidering systems will not queue URLs until they mirror the updated topic definitions. After URLs are queued, each spidering or file scanning collection will need to re-index the documents to assign them to topics. After the documents are re-indexed, the mirror collections will copy the changed collection indexes. When the updated collections are available on the mirroring host, the reports will reflect the results of the topics changes.

Blurbs in multi-host systems also require a bit more attention. If the blurb includes HTML with relative URLs (in an `<img>` tag, for example), the URL will be evaluated with respect to the host serving the page. So, the targets of those URLs will need to be manually copied to each CCE host. Alternatively, the URLs can be made absolute, and all refer to the same page or image.
15 Mirroring
Multiple Host Issues
## Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>activity curfew</td>
<td>Settings available from the Collections</td>
</tr>
<tr>
<td>Administrative Interface</td>
<td>The web interface used to administer the Ultraseek server. To access this interface, point your web browser to machine running Ultraseek with the following URL: <a href="http://hostname:port/admin">http://hostname:port/admin</a></td>
</tr>
<tr>
<td>administrator</td>
<td>Ultraseek user with administrative privileges. Administrators have permissions to modify and delete collections.</td>
</tr>
<tr>
<td>binding address</td>
<td>Associates a static IP address with an instance of Verity Ultraseek. You can use address binding to run multiple instances of Ultraseek on the same port, each bound to a different IP address. By default, Ultraseek binds to all of the IP addresses; therefore, if you bind it to a specific IP address this will free up the other IP addresses on the port.</td>
</tr>
<tr>
<td>collection</td>
<td>A set of searchable indexes and settings for a group of documents. Some collection types include a mechanism for adding individual documents to and updating the indexes.</td>
</tr>
<tr>
<td>collection filter</td>
<td>A URL pattern or site pattern on the Collections</td>
</tr>
</tbody>
</table>
Content Assistant
An external program that allows administrators to classify, filter, and enhance information about documents as they are indexed by Ultraseek. A Content Assistant can perform the following actions during the indexing process:
- Assign topics to documents
- Filter documents
- Add document metadata
- Replace document titles and document description information

Content Classification Engine (CCE)
The Ultraseek module that allows administrators to create and manage categories, specify rules to automate the classification of documents, and generate reports analyzing documents linked to existing topics. CCE allows users to point-and-click their way through categories and subcategories that appear on the search page.

deduping
Removing from search results any duplicate search hits with the same URL, or with the same title and description.

document score
See relevance score.

Dublin Core
A set of 15 standard HTML tags for tracking and cataloging Web pages and creating metadata. These tags are: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier (URL), Source Language, Relation, Coverage, and Rights (copyright information). See http://dublincore.org for more information.

HTTP keep-alive
An HTTP protocol option that keeps a TCP connection open for connections. When the option is “off,” a new TCP connection is opened for each HTTP request.

in-memory index
A search index for storing documents. When the in-memory index gets full, the spider saves it by creating an on-disk index.
See also on-disk index.

index
1. n. A compilation of terms that can be searched. Each term in a Verity Ultraseek collection index is stored together with a list of the spidered documents containing the term.
2. v. To parse a document and enter its terms into a collection index. The Verity Ultraseek spider indexes documents.
indexer weight
An integer that defines the importance of a portion of a document
relative to the body text of the document, which is weighted at 1.
The greater a weight for a section, the more a word in that section
impacts the relevance score of the document. However, field
specifiers in search queries, such as title:, cause Ultraseek to
disregard the indexer weight assigned to that portion of returned
documents. Words in a section with an indexer weight of zero are
ignored in determining a document’s score.

metadata
Non-content information about documents. Metadata can be used
to document data elements or attributes (name, size, data type), data
about records or data structures (length, fields, columns), or data
about data (where it is located, how it is associated, ownership).
Ultraseek examines document and protocol headers and performs
internal index analysis to determine document metadata.

on-disk index
A search index for storing documents. Ultraseek creates an on-disk
index when the in-memory index gets full. A collection will have
several on-disk indexes, which are eventually merged.
See also in-memory index.

Page Expert
A feature in Ultraseek that increases the relevance of search results
by letting the administrator specify portions of a document that
should not be indexed. This is an excellent way to ignore
unimportant content, such as navigation, copyrights, and legal
disclaimers.

parse
To break down a document’s content into its component parts.
These components can then be used by Verity Ultraseek to create
index terms.

proxy server
An HTTP server between an HTTP client and Ultraseek.
Administrators can configure Ultraseek to use a proxy server to
provide security, administrative control, and caching service.

quality factor
An integer between -16 and +15 used to adjust the term-dependent
portion of a document’s relevance score for a query. Administrators
assign a quality factor by specifying an URL pattern and assigning it
an integer. To increase the relevance score of documents from a
URL, assign a positive value. To decrease the relevance score of
documents from a URL, assign a negative value. Ultraseek’s indexer
assigns a final quality factor to each document it indexes based on
internal algorithms and the quality factor you specify.

query term
One part of a search query that users enter into the search input box.
Glossary

Quick Links
Search results that appear at the top of the search results page after a user searches for predefined keywords and phrases. Ultraseek administrators use the Server / Quick Links pane in the administrative interface to define the keywords and phrases and specify the documents that are returned when searches contain these terms.

regex
A regular expression used to filter content in collections. A selection option for many of the Verity Ultraseek Administrative interface fields. For example, administrators specify regular expressions to configure Ultraseek to work with specific IP addresses.

relevance score
An estimate by the Ultraseek server of how closely a document in search results pages matches the search query. Relevance scores are expressed as percentages.

root
The URL of the first page visited by the spider when building a collection.

score
See relevance score.

search
To match query terms against terms in the collection index, and return links to documents containing the matching terms.

spelling suggestion
Ultraseek server parameter that, when checked, suggests re-spellings of query terms when a re-spelling may result in a more useful query.

spider
[noun] The Verity Ultraseek code that builds HTTP-based collections.
[verb] To obtain documents from an HTTP server, parse each document, feed terms from each document into a collection index, and store each document URL in a collection database.

spider throttle
A URL pattern and setting that specifies the number of seconds to wait before sending another HTTP request to an HTTP server. The spider waits the specified time between document requests submitted to all URLs matching a pattern, which reduces the load on web servers.

Style Editor
A feature in Ultraseek 5.1 that provides a graphical user interface for customizing the Ultraseek search and search results pages.
**Glossary**

**topic**
A link on a search page that enables users to quickly find important documents without having to compose sophisticated queries. Administrators arrange topics in a hierarchy to classify important content. You must license the CCE (Content Classification Engine) module to include topics in your search application.

**thesaurus**
A language-specific XML file that relates sets of query terms. When users enter a search query specified in a `<show>` element in the xml file, all other query terms in `<show>` elements within that set are displayed to the user as checkboxes that expand their search. By default, Ultraseek contains an english language thesaurus, `thesaurus_en.xml`, located in the `language` subdirectory of your Ultraseek installation. Administrators must edit this file to add new thesaurus entries, or create a new file for a thesaurus in another language.

**thread**
The computing resources allocated by the spider to creating or revisiting an index (indexer thread). Thread can also refer to the computing resources allocated by the search engine to maintain simultaneous HTTP connections (server thread).

**Ultraseek Query Language**
The language defining the query syntax used by the Ultraseek server. See the *Verity Ultraseek Administrator Guide* for more information about using plus, minus, and wildcard operators, vertical and double-vertical bars, and field searches in queries.

**URL database**
A subset of an Ultraseek collection that holds information about URLs and links.

**URL pattern**
A URL path with /* and *.extension wildcards used to filter HTTP requests. The /* wildcard matches all URLs from that point forward, and the *.extension matches any file name with the specified extension.

**User Agent**
A header line in an HTTP message which tells the receiver of the message what application sent the message. This information is most commonly used by administrators viewing access logs to distinguish human users from spiders. The Verity Ultraseek spider uses HTTP messages to obtain web documents for indexing. Web servers receiving these messages might check the User Agent header to determine whether the spider is allowed to examine web content.

**wildcard**
An operator that expands the results of a query search. Ultraseek supports two wildcards (*) in individual search terms, but not in search phrases.
A
about
  mirroring 287
  reports 279
  rules 268
  the topic definition file 266
access log files
  viewing 155
add-in modules 26
administrative user
  changing passwords 167
  creating accounts 166
  deleting accounts 166
allowing directories in collections 61
ArrayOfDocItem 218
assigning
  rules to a topic 259
  URLs to a topic 259
authentication, sites
  controlling access to 93
automatic refresh 147
automatic URL adder (cooperative spidering) 99
automatically check for updates 156

B
binding address 156
BLOBs
  parsing data 165
blog entries, adding to spider collections 43
blurb
  of topic, changing 262

Blurb field 262
browsing topics 265
bug reports
  email address 157

C
cache size
  title cache size 162
capitalization
  search syntax 230
CAS identification block 214
case 105
case sensitivity
  search syntax 230
CCE
  defined 245
  initial page parameters 254
  license keys 156
CGI scripts, allowing URLs to 66
changing
  topic blurbs 262
  topic names 262
character encoding 158
clearing collections 76
client access, restricting 170
collections
  allowing directories 61
  clearing 76
  creating 39
  data directory 104
  defining filters 59
  deleting 75
  display name 115
  HTML meta tag names 109
importing from Verity Information Server 70
internal name 115
managing URLs 80
merging 58
mirroring 56
primary language 104
reindexing 76
revisiting 76
URL status 80
view log 78
view sites 77
viewing indexing status 146
configuration file (collection) 69
configuring Content Assistants 209
connection timeout 160
content
evaluating 217
Content Assistant
license keys 156
Content Assistants
configuring 209
defined 205
document evaluation 216
evaluating documents 206
example uses 207
identification block 214
overview 206
sample programs 206
topic tree creation 211
cookie, HTTP header entries 96
copying topics 266
creating
cross references
to subtopics 265
cross-references 260
page types 86
subtopics 265
supertopics 266
topics 258
creating collections 39
cross references
creating
to subtopics 265
cross-references
creating 260
defined 260
in topic trees 214
D
data directory 27, 104
data level 209
databases
indexing 48
parsing binary data 165
supported databases 48
date override 110
debug mode 165
debugging level 209
deleting collections 75
deleting documents 82
deleting topics 265
description
customizing searches by 238
description override 110
direct file access 98
directories
allowed in collections 61
directory listing documents 66
discover sites 104
display name 115
displaying
topics page 247
DocItem 217
document
revisit interval 102
document directory
setting 157
document evaluation 216
by entire content 217
by header description 217
by URL 216, 217
modifying contents 221
response 219
document scores, improving 84
document title
customizing searches by 237
document type, customizing searches by 238
document types
  managing 172
  parsing 177
  supported highlight types 176
  supported MIME types 174
documents
  default age estimate 103
  deleting from indexes 82
  order of, in topic 270
double-pipe
  search syntax 233
dublin core metadata
  customizing searches by 241
duplicate hits
  removing results with the same URL 165

e
editing topics 263
e-mail address
  bug reports 157
  important notifications 157
  encoding 158
error log files 161
  log file level 161
  number of files to keep 161
  operating system log level 160
  viewing 155
error messages
  debug mode 165
  evaluating documents 206

f
field rules 270
  usefulness 270
field searches
  by description 238
  by document title 237
  by document type 238
  by dublin core metadata 241
  by keywords 238
  by language 239
  by link 240
  by meta info 240
  by site 241
  by topic 242
  by URL 237
fields
  Blurb 262
  Name 260, 262
file log level
  errors 161
  logging errors 161
files, accessing directly 98
filter number of hits 163
filters
  defining 59
firewalls, and indexing 98
form based authentication 94
FullDesc 217

G
GetFullDesc
  example message 224
GetFullDesc procedure 217
GetHdrDesc procedure 217
GetURLDesc procedure 216, 217

H
HdrDesc 217
headers
  evaluating 217
  heap size 163
highlightable document types 176
home topic 254
hostname 156
  primary 104
Index

HTML Meta tag names 109
  keywords 111
HTML tag override
  date 110
  description 110
  keywords 111
  publisher 111
  title 110
  URL 111
HTTP
  allowing keep-alives 164
  connection timeout 102, 160
  keep-alive timeout 160
  keep-alives 104
  server port 157
HTTP header entries 96
HTTP PUT conventions 215
HTTP server
  and spider traffic 98
  number of threads 160
  threads, viewing status 149
HTTPS
  server port 158

I
ignore case differences in URLs 105
images
  HTTP requests 164
implementing page types 88
importing
  collection configurations 68
  subtopics 265
  topic lists 261
  Ultraseek collection configurations 68
  Verity Information Server collection configurations 70
improving document scores 84
IndexAsstDesc data type 219
indexes
  merging 79
  merging automatically 103
indexing
  adding a URL 81
  and firewalls 98
  databases 48
  deleting URL 82
  enabling stemming 165
  excluding a subdirectory 63
  filtering documents by size 103
  filtering text from directory listings 66
  HTTP keep-alives 104
  improving performance 104
  including CGI scripts 66
  in-memory size 103
  limiting by document type 172
  new sites 104
  number of spider threads 102
  parse CPU time 102
  reducing redundancy 104
  removing documents from the index 82
  restricting by authorization 93
  restricting by user agent 93
  resuming 76
  revisit interval 102
  revisit queues 102
  scoring documents 89, 108
  setting weights 108
  suspending 76
  viewing status 146
indexing collections 76
in-memory index size 103
installation instructions 247
installing Search/CCE 247
interleaved sites 165
interleaving search results 165
internal name 115
IP address
  resolving client addresses 164
Index

J
java example 206

K
keep-alives
  HTTP connection 164
  HTTP timeouts 160
keywords
  creating quick links 181
  customizing searches by 238
  overrides 111

L
language
  character encoding 158
  customizing searches by 239
  default language 158
  primary 104
license keys
  enabling add-in modules 156
  expired license keys 156
  for mirroring 287
  invalid license keys 156
  obtaining 156, 247
  restrictions 156
link
  customizing searches by 240
log files
  access log file 155
  debug mode 165
  error log file 155
  file log level 161
  number of files to keep 161
  operating system log level 160
  query log file 155
  viewing 155
logging
  requests for images 164

M
malloc, allocating memory to 163
mapping
  XML elements 177
maximum document download size 103
maximum parse CPU time 102
memory, allocating to malloc 163
merging collections 58
merging indexes 79
meta searches, using to customize searches 240
meta tag
  using to create keywords 240
  using to customize field searches 240
MIME types 273
  supported MIME types 174
mirrored collections
  creating 56
mirroring
  about 287
  disabling 288
  enabling 288
  multiple host issues 289
mirroring a remote collection 56
most requested documents reports 140
Must Contain qualifier 269
Must Not Contain qualifier 269

N
Name field 260, 262
name, of topic, changing 262
network
  direct file access 98
  proxy servers 98
  settings 92
no-clickthrough reports 138
no-results reports 137
Index

O
operating system
  logging errors 160
  viewing status 152
otherwise allowed sites 64
override HTML meta tag names 110

P
Page Expert 85
page types 85
pages
  topics page 247
parameters
  CCE search pages 255
  CCE topic browse pages 255
  main server parameters 154
  server parameters 154
parsing
  document types 177
passwords
  changing 167
performance
  and spider threads 102
  reducing server load 97
  viewing information on 148
phrases
  search syntax 229
pipes
  search syntax 232
plus and minus operators
  search syntax 231
port
  specifying on Solaris 27
primary language 104
proxy servers 98
publisher override 111
pushing topic trees 214
python example 206

Q
quality factor 89
queries
  enabling stemming 165
query
  number of hits 163
  removing duplicate hits 165
query cache size 161
query log files
  viewing 155
Query parameters 128
query performance
  HTTP keep-alives 164
query strings
  allowing in URLs 66
Quick Links
  creating 181
  defined 181

R
ranking URLs 259
recent queries 143
reindexing collections 76
removing documents 82
reports
  about 279
  documents per topic
    sorted by hierarchy 281
    sorted by population 282
  most requested documents 140
  no-clickthrough 138
  no-results 137
  queries, sorted by frequency 283
  query terms, sorted by frequency 285
  top queries 136
  topics, sorted by browse frequency 284
  trend 141
resolving IP addresses 164
resume indexing 76
Index

revisit queues   102
revisit URL   76
revisiting collections   76
rule terms   269
    and subtopics   269
rules
    about   268
    assigning to a topic   259
document count   275
field rules   270
    usefulness   270
getting started   268
sample rules   276
show more rule and URL entries button   268
specifying
    by collection   272
    by document title   272
    by document type   273
    by metadata   274
    by raw terms   272
    by site   271
    by URL   271
syntax of   269
testing   275
running
    on Solaris   27

S
sample rules   276
scanning the filesystem   44
scoring documents   89, 108
search results
    number of hits   163
search these results   232
Search/CCE
    and collections   257
    benefits of   246
    installation instructions   247
    installing   247
starting   247
searches
    case sensitivity   230
    plus and minus operators   231
    using capitalization   230
    using double-pipe   233
    using phrases   229
    using pipes   232
    using wildcard search   234
searches, customizing
    by description   238
    by document title   237
    by document type   238
    by dublin core metadata   241
    by language   239
    by link   240
    by meta searches   240
    by site   241
    by topic   242
    by URL   237
security
    HTTPS port   158
    license keys   156
server
    binding address   156
    hostname   156
    HTTP threads   160
    limiting client access   170
    limiting virtual image size   163
    main server parameters   154
    parameters   154
    performance information   148
    proxy   98
    restarting   152
    shutting down   153
    total memory usage   163
service location   209
site
    customizing searches by   241
sitelist file   105
sites
  allowing and disallowing  64
  restricting indexing  93
SMTP mail forwarding host  157
SOAP
  procedures  207
spider
  and local HTTP server  98
  pausing between requests  97
  revisit URL  76, 81
  setting filter options  65
  threads  102
  throttle  97
spider filter options  65
spidering the network  40
starting Search/CCE  247
stemming
  enabling  165
storage heap size  163
Style Editor
  Ultraseek styles  117
Style parameters  125
Stylesheet properties  120
subtopics
  creating  265
  importing  265
supertopics
  creating  266
suspend indexing  76
syslog facility  160
test sites
  allowing and disallowing  64
  restricting indexing  93
SMTP mail forwarding host  157
SOAP
  procedures  207
spider
  and local HTTP server  98
  pausing between requests  97
  revisit URL  76, 81
  setting filter options  65
  threads  102
  throttle  97
spider filter options  65
spidering the network  40
starting Search/CCE  247
stemming
  enabling  165
storage heap size  163
Style Editor
  Ultraseek styles  117
Style parameters  125
Stylesheet properties  120
subtopics
  creating  265
  importing  265
supertopics
  creating  266
suspend indexing  76
syslog facility  160
testing rules  275
threads
  HTTP server  160
  viewing status  149
timeouts
  HTTP connection  160
  keep-alives  160
title cache size  162
title override  110
top queries reports  136
topic definition file, about  266
topic lists
  and rules  261
  importing  261
topic trees
  CAS identification block  214
  creating  211
  cross-references  214
  pushing to Ultraseek  214
  XML definition sample  213
topic, customizing searches by  242
topics
  browsing  265
  copying  266
  creating  258
  deleting  265
  editing  263
  export as TAX file  263
home topic  254
reports  284
topics page
  displaying  247
trend reports  141

T
TAX file
  export topics as  263
TCP port
  connect time for queries  164
termlist
  cache size  161
  maximum length  162
testing page types  88

testing rules  275
threads
  HTTP server  160
  viewing status  149
timeouts
  HTTP connection  160
  keep-alives  160
title cache size  162
title override  110
top queries reports  136
topic definition file, about  266
topic lists
  and rules  261
  importing  261
topic trees
  CAS identification block  214
  creating  211
  cross-references  214
  pushing to Ultraseek  214
  XML definition sample  213
topic, customizing searches by  242
topics
  browsing  265
  copying  266
  creating  258
  deleting  265
  editing  263
  export as TAX file  263
home topic  254
reports  284
topics page
  displaying  247
trend reports  141

U
Ultraseek
  Administrative interface  28
  features  25
  restarting  152
  shutting down  153
starting 27
  system status 152
Unix username 160
updates 156
  automatically check 156
URI location 209
URL
  allowing query strings 66
  allowing to CGI scripts 66
  customizing searches by 237
  deleting 82
  duplicate filter specification 84
  duplicate preference 83
  filtering sequence 62
  filters 59
  hostname 104
  revisit queues 102
  revisiting 102
  scoring documents 89
  using existing 64
  viewing status 80
URL override 111
URLDesc 216, 217
URLs
  adding 81
  assigning to a topic 259
  collection management 80
  evaluating 216, 217
  ranking 259
  revisit queues 102
  usage summaries 139
user agent 93
using search these results 232

W
  wildcard search
    search syntax 234

X
  XML
    conflicting root elements 180
    generic mappings 178
    mappings 177
    specifying attributes 180

V
  Verity Information Server 70
    importing collections from 70
view log 78
view sites 77
viewing log files 155
Index