

**Ultraseek**

**Ultraseek**

**Version 5.6**

# Implementation Guide

February 15, 2006  
Part Number DE0298

## Notice

This document is a proprietary product of Autonomy, Inc. and is protected by copyright laws and international treaty. Information in this manual is subject to change without notice and does not represent a commitment on the part of Autonomy, Inc. While reasonable efforts have been made to ensure the accuracy of the information contained herein, Autonomy, Inc. assumes no liability for errors or omissions. No liability is assumed for direct, incidental, or consequential damages resulting from the use of the information contained in this document.

The copyrighted software that accompanies this document is licensed to the End User for use only in strict accordance with the End User License Agreement, which the Licensee should read carefully before commencing use of the software. No part of this publication may be reproduced, transmitted, stored in a retrieval system, nor translated into any human or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual or otherwise, without the prior written permission of the copyright owner, Autonomy, Inc., 892 Ross Drive, Sunnyvale, California 94089.

This document may use fictitious names for purposes of demonstration; references to actual persons, companies, or organizations is strictly coincidental.

## Trademarks and Copyrights

Copyright 2006 Autonomy Systems plc. All rights reserved. Autonomy Dashboard™, Autonomy Desktop Suite™, DAH™, DIH™, DiSH™, IDOL™, IDOL server™, Portal-in-a-Box™, and Retina™ are trademarks of Autonomy Systems plc.

Copyright 2006 Verity, Inc. All rights reserved. Verity™, Cardiff™, the Verity logo, the LiquidPDF logo, KeyView™, Ultraseek™, Knowledge Organizer™, TOPIC™, Verity Portal One™, Verity Profiler™, LiquidOffice™, LiquidPDF™, Connect Agent™, HTML+Forms™, MediClaim™, PDF+Forms™, TeleForm™, Tri-CR™, RecoFlex™, AutoMerge Publisher™, TrueAddress™, and VersiForm™ are trademarks or registered trademarks of Verity, Inc., part of the Autonomy group of companies.

Portions of this product use Teragram Software.

This product may incorporate intellectual property owned by Microsoft Corporation. The terms and conditions upon which Microsoft is licensing such intellectual property may be found at

<http://msdn.microsoft.com/library/en-us/odcXMLRef/html/odcXMLRefLegalNotice.asp?frame=true>

Microsoft is a registered trademark, and MS-DOS, Windows, Windows 95, Windows NT, and other Microsoft products referenced herein are trademarks of Microsoft Corporation.

UNIX is a registered trademark of The Open Group.

Includes Adobe® PDF. Adobe is a trademark of Adobe Systems Incorporated.

*All other trademarks are the property of their respective owners.*

## Notice to Government End Users

If this product is acquired under the terms of a **DoD contract**: Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of 252.227-7013. **Civilian agency contract**: Use, reproduction or disclosure is subject to 52.227-19 (a) through (d) and restrictions set forth in the accompanying end user agreement. Unpublished-rights reserved under the copyright laws of the United States. Autonomy, Inc., 892 Ross Drive, Sunnyvale, California 94089.

# Contents

Figures, Tables, and Listings .....	5
<b>Preface</b> .....	7
Using This Book .....	7
Version .....	7
Organization of This Book .....	8
Stylistic Conventions .....	8
Related Documentation.....	9
Support and Contact Information .....	10
Downloading the Latest Documentation.....	10
Ultraseek Technical Support .....	10
Contact Autonomy.....	11
<b>1 Deploying Ultraseek</b> .....	13
Deployment Factors .....	13
Network Configuration .....	14
Repository Size .....	14
Query Volume .....	15
Request Latency .....	15
Query Origin .....	15
Deploying Intranet Search.....	16
Deploying Public Site and Intranet Search.....	17
<b>2 Configuring Ultraseek</b> .....	19
Integrating Ultraseek into a Portal.....	20
Assigning Multiple IP Addresses to a Single Host .....	20
Index Content and Service Queries on Different Hosts .....	21
Providing Failover with Mirror Collections .....	22
Accessing Ultraseek Through a Proxy Server .....	22
Segregating Content by Collection.....	23

## Contents

Restricting Access by IP Address for Extranet Installations .....	24
Localizing Your Search Application .....	25
Authenticating Users .....	26
Indexing Restricted Content .....	26
Filtering Search Results.....	26
Requiring User Login.....	27
Customizing Ultraseek’s Default Authentication.....	27
Providing Multiple User Interfaces.....	28
<b>3 Improving Search Usability .....</b>	<b>29</b>
Customizing Your Search Interfaces .....	30
Classifying Your Content.....	31
Creating Quick Links .....	32
<b>Glossary .....</b>	<b>35</b>
<b>Index .....</b>	<b>41</b>

# Figures, Tables, and Listings

<b>Figure 1-1</b>	Intranet Search Deployment.....	16
<b>Figure 1-2</b>	Mixed Public Site and Intranet Search Deployment .....	17
<b>Figure 2-1</b>	Separating Content Indexing from Query Response.....	21
<b>Figure 2-2</b>	Address-based Access Level Specification .....	24
<b>Table 2-1</b>	Ultraseek language modules .....	25
<b>Figure 3-1</b>	Ultraseek Topics.....	31
<b>Figure 3-2</b>	Quick Link to Java Documentation .....	32

## Figures, Tables, and Listings

# Preface

This guide is for administrators tasked with deploying and configuring an Ultraseek-based search application on their corporate network. It is intended for readers who have previously installed enterprise applications in a production environment and understand basic networking and administrative concepts.

This preface contains the following sections:

- [Using This Book](#)
- [Related Documentation](#)
- [Support and Contact Information](#)

## 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 Ultraseek version 5.6. The content was last modified February 15, 2006. Corrections or updates to this information may be available through the Customer Support site; see [“Support and Contact Information”](#) on page 10.

## Organization of This Book

This book includes the following chapters:

- Chapter 1, “Deploying Ultraseek” describes the issues that affect how you deploy Ultraseek, then describes sample deployments for intranet and public site search.
- Chapter 2, “Configuring Ultraseek” describes common configuration techniques for Ultraseek.
- Chapter 3, “Improving Search Usability” describes ways in which you can increase the usability of your search applications.

## Stylistic Conventions

The following stylistic conventions are used in this book.

Convention	Usage
Plain	Narrative text.
<b>Bold</b>	User-interface elements in narrative text: <ul style="list-style-type: none"><li>■ Click <b>Cancel</b> to halt the operation.</li></ul>
<i>Italics</i>	Book titles and new terms: <ul style="list-style-type: none"><li>■ For more information, see the <i>Ultraseek Customization Guide</i>.</li><li>■ A <i>crawler</i> is a software process that gathers documents for indexing.</li></ul>
Monospace	File names, paths, and code: <ul style="list-style-type: none"><li>■ The <code>name.ext</code> file is installed in: C:\Autonomy\Data\</li></ul>
<i>Monospace italic</i>	Replaceable strings in file paths and code: <ul style="list-style-type: none"><li>■ <code>user username</code></li></ul>
<b>Monospace bold</b>	Data types and required user input: <ul style="list-style-type: none"><li>■ <code>SrvConnect</code> A connection handle.</li><li>■ In the <b>User Interface</b> text box, type <code>user1</code>.</li></ul>

The following command-line syntax conventions are used in this book.

<b>Convention</b>	<b>Usage</b>
<code>[ optional ]</code>	Brackets describe optional syntax, as in <code>[ -create ]</code> to specify a non-required option.
<code> </code>	Bars indicate “either   or” choices, as in <code>[ option1 ]   [ option2 ]</code>  In this example, you must choose between <code>option1</code> and <code>option2</code> .
<code>{ required }</code>	Braces describe required syntax in which you have a choice and that at least one choice is required, as in <code>{ [ option1 ] [ option2 ] }</code>  In this example, you must choose <code>option1</code> , <code>option2</code> , or both options.
<code>required</code>	Absence of braces or brackets indicates required syntax in which there is no choice; you must enter the required syntax element.
<code>variable</code>	Italics specify variables to be replaced by actual values, as in <code>-merge filename1</code>
<code>...</code>	Ellipses indicate repetition of the same pattern, as in <code>-merge filename1, filename2 [, filename3 ... ]</code>  where the ellipses specify <code>, filename4</code> , 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.

## Related Documentation

---

The following guides provide more details on installing, administering, and customizing Ultraseek:

- *Ultraseek Installation Guide*

- *Ultraseek Administrator Guide*
- *Ultraseek Customization Guide*

## Support and Contact Information

---

Read this section if you want to contact Autonomy, request technical support, or obtain product documentation.

### Downloading the Latest Documentation

You can download documentation from the Ultraseek Download Center:

<http://downloadcenter.ultraseek.com/dlc/documentation.do>

### Ultraseek Technical Support

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

- Call, fax, or email the support group at the location nearest to you:

<b>Europe and Worldwide</b>	<b>The Americas</b>
Autonomy Systems Ltd. Cambridge Business Park Cowley Road, Cambridge CB4 0WZ Telephone: 00 800 4837 4890 (UK, France, Germany, Netherlands, Spain.) [Hours: 9:00 AM to 5:00 pm (GMT+1)] +1 403 294 1107 (Canada direct) Email: <a href="mailto:search-support@autonomy.com">search-support@autonomy.com</a>	Autonomy Inc. 892 Ross Drive Sunnyvale, California 94089 Telephone: +1 403 294 1107 877 483 7489 (U.S. and Canada) [Hours: 7:00 AM to 6:00 PM MST (GMT-7)] (or leave voice mail) Email: <a href="mailto:search-support@autonomy.com">search-support@autonomy.com</a>

- Access the Customer Support site, at

<https://customers.autonomy.com>

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 home page.

- Access the support site from the Ultraseek web page, at

<http://www.ultraseek.com/support/index.html>

## Contact Autonomy

Contact the location nearest to you for general information about Autonomy:

<b>Europe and Worldwide</b>	<b>The Americas</b>
Autonomy Systems Ltd. Cambridge Business Park Cowley Road, Cambridge CB4 0WZ Telephone: +44 (0) 1223 448 000 Fax: +44 (0) 1223 448 001 General information email: <a href="mailto:autonomy@autonomy.com">autonomy@autonomy.com</a>	Autonomy Inc. 892 Ross Drive Sunnyvale, California 94089 Telephone: +1 408 541 1500 Fax: +1 408 541 1600 General information email: <a href="mailto:info@us.autonomy.com">info@us.autonomy.com</a>

**Preface**  
Support and Contact Information

# 1 Deploying Ultraseek

This chapter describes the issues that affect how you deploy Ultraseek, then describes sample deployments for intranet and public site search.

- [Deployment Factors](#)
- [Deploying Intranet Search](#)
- [Deploying Public Site and Intranet Search](#)

---

**Note** For information on installation requirements, see the System Requirements chapter of the *Ultraseek Installation Guide*.

---

## Deployment Factors

---

The optimum architecture for your deployment depends on several factors:

- Network configuration
- Number of documents in your repository
- Volume of queries you will service
- Desired latency of requests
- Origin of queries, internal vs. external users

A common deployment technique is to separate indexing from query response, as described in [“Index Content and Service Queries on Different Hosts” on page 21](#). An index server is an instance of Ultraseek server that is dedicated to indexing content, but is not configured to respond to user queries. Conversely, a query server is an instance of Ultraseek server that is dedicated to responding to user queries, but does not index content. [Figure 1-2 on page 17](#) illustrates a deployment scenario with multiple index and query servers. You can install as many index and query servers as your deployment scenario requires. For small deployments, such as an internal department search, you may not need to separate indexing from query response. In this case, the index and query servers are the same.

## Network Configuration

Your network infrastructure affects how you deploy Ultraseek, especially with large or public sites. Security issues, such as whether to install Ultraseek inside or outside the firewall, depend more on how your network is configured than on Ultraseek. The default Ultraseek configuration is optimized for most use cases. However, your networking hardware and software may have limits or requirements that affect your Ultraseek deployment.

This guide does not provide network-specific instructions. However, it does refer to elements common to any network, such as a firewall and HTTP servers. See your networking documentation for details such as whether HTTP traffic is limited to predefined port numbers, or whether you can mirror content across your firewall.

## Repository Size

The size of your document repository determines how many index servers you should install for your deployment. As a guideline, you should install a second indexing server if your repository contains more than two million documents. If you must index more than four million documents, you should install four index servers.

A related issue is the frequency with which your content changes. If your repository contains mostly static documents, you may not need multiple indexing servers. However, if your repository contains mostly dynamic documents, such as web pages, you should consider a second indexing server as you approach the threshold of two million documents.

## Query Volume

The number of query servers you need depends on the following considerations:

- Whether you need to continue servicing queries while upgrading software, such as operating system patches, on the machine that hosts Ultraseek
- Whether you need to continue to service queries in the event of unscheduled down time
- The average and peak number of queries you receive per second and per day

If you must continue servicing queries when the machine that hosts Ultraseek is scheduled for downtime, you should install a second query server on a second machine. Doing so provides a backup server if the first server crashes or is temporarily unable to perform.

Ultraseek can process a maximum of 15 qps (queries per second) and 5 qps when passage-based summaries are displayed. If you receive more requests than this during your peak periods, you should consider adding a second query server to your deployment. As a general guideline, a single instance of Ultraseek can easily respond to 500,000 queries per day. However, you should also consider a second query server when your query volume exceeds one million queries per day.

## Request Latency

Request latency is the amount of time elapsed between when a user submits a search query and starts to receive results. If the same content is used to service queries from a wide geographic area, such as multiple continents, you should consider placing mirror collections of your content across multiple geographically-dispersed machines. This will decrease the latency of your users' search requests.

See [“Index Content and Service Queries on Different Hosts” on page 21](#) and [“Providing Failover with Mirror Collections” on page 22](#) for more information about mirror collections.

## Query Origin

If you will service queries from users outside your corporate network, you may consider installing a query server outside the corporate firewall. However, it is highly recommended that you administer such a server from a browser within the firewall. If you decide to administer Ultraseek from outside the firewall, you must configure your web server to allow Ultraseek to receive external requests.

# Deploying Intranet Search

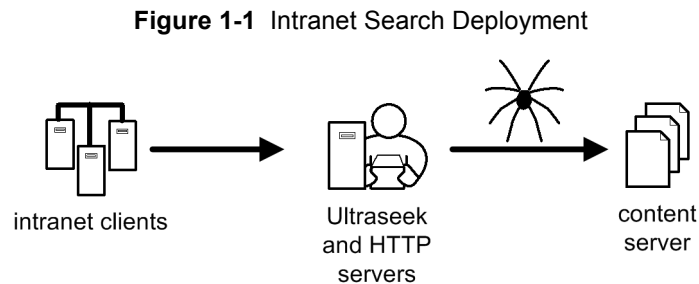
---

This section describes a typical Ultraseek deployment on a corporate intranet. The implementation will service queries for a large department within a company and makes the following assumptions:

- Repository with roughly 700,000 web-based documents
- Query volume of roughly 50,000 queries per day
- All search requests originate from within the company firewall

In this situation, it is not necessary to perform indexing and query response on separate hosts. In addition, the total number of documents is well below the 2 million document guideline for a second indexing server. There is a higher tolerance for variable request latency and scheduled downtime since the search application serves intranet users who are presumably accustomed to regular software upgrades.

Figure 1-1 illustrates the simple deployment architecture for this scenario:



Both Ultraseek and the HTTP server are installed on the same host machine in this scenario.

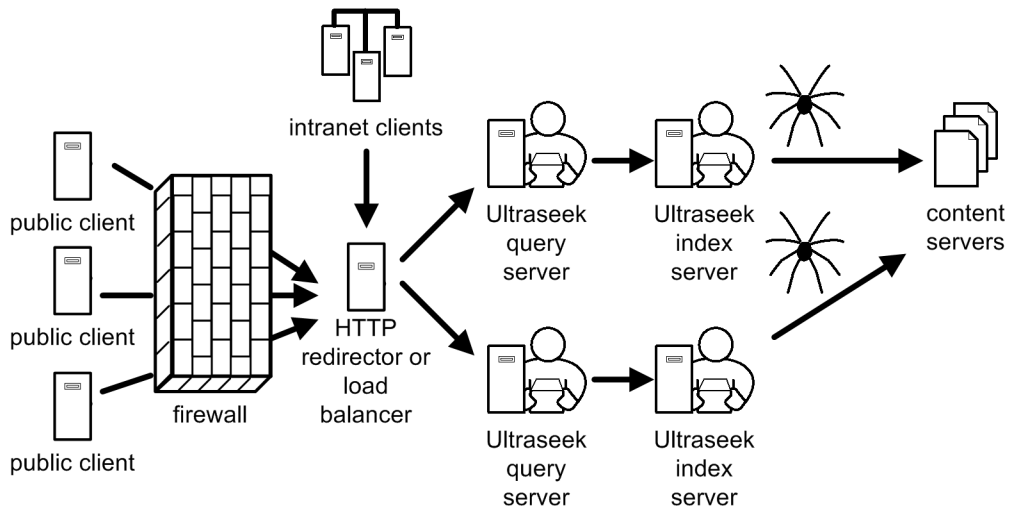
## Deploying Public Site and Intranet Search

This section describes the deployment architecture for a large public site with the following assumptions:

- Repository with roughly three million documents
- Public and intranet clients access the same documents
- Query volume of roughly 1 million queries per day, but with two thirds of this volume occurring within a six hour period
- Minimal request latency desired
- Search requests originate from external users and also from within the corporate intranet

Figure 1-2 illustrates a deployment architecture to meet the needs of this scenario:

**Figure 1-2** Mixed Public Site and Intranet Search Deployment



Ultraseek is installed on a total of four different hosts. The HTTP server is installed on a host separate from Ultraseek. The search administrator should use third-party HTTP redirection or load balancing software to distribute requests between the Ultraseek query servers. Be sure to configure your redirection or load balancing software to use only one Ultraseek server for an

entire client session. There are two Ultraseek query servers so that up to 30 queries per second can be processed during the peak six-hour period. In addition, the repository is large enough to require two indexing servers. The two query servers should mirror the collections created on the index servers.

---

**Note** You can also use Ultraseek XPA to distribute search requests among multiple Ultraseek servers. See the *Ultraseek XPA Programming Guide* for more information.

---

Whether you install Ultraseek inside or outside the firewall depends on your network infrastructure and the relative privacy of the documents in your repository. If you are serving the same content to public and intranet users, there is no reason to hide Ultraseek behind the firewall. However, if you are serving separate, private content to intranet users, then you may want to install one Ultraseek query and index server behind the firewall and another pair outside.

# 2

## Configuring Ultraseek

This chapter assumes that you have already installed Ultraseek. If you experience difficulty during installation, see the *Ultraseek Installation Guide*.

The default settings for Ultraseek server are already optimized for most use cases. It is highly recommended that you do not change them unless you have a specific reason for doing so. This chapter describes common configuration techniques related to deploying Ultraseek. Consult the *Ultraseek Administrator Guide* for information about other server configuration techniques.

- [Integrating Ultraseek into a Portal](#)
- [Assigning Multiple IP Addresses to a Single Host](#)
- [Index Content and Service Queries on Different Hosts](#)
- [Providing Failover with Mirror Collections](#)
- [Accessing Ultraseek Through a Proxy Server](#)
- [Segregating Content by Collection](#)
- [Restricting Access by IP Address for Extranet Installations](#)
- [Localizing Your Search Application](#)
- [Authenticating Users](#)
- [Providing Multiple User Interfaces](#)

## Integrating Ultraseek into a Portal

---

Please contact [software-sales@autonomy.com](mailto:software-sales@autonomy.com) if you want to integrate Ultraseek into third-party portal or application server software, such as the following:

- BEA WebLogic
- IBM WebSphere
- Yahoo! PortalBuilder

Ultraseek provides a portlet for each of these portals, but does not actively support them. Contact your Ultraseek sales representative for more information.

## Assigning Multiple IP Addresses to a Single Host

---

Many firewalls allow HTTP traffic only on port 80. If you run both Ultraseek and an HTTP server on the same machine behind such a firewall, you cannot run Ultraseek on its default port: 8765. A recommended solution is to assign two IP addresses to a single machine behind the firewall, then bind Ultraseek to one of these addresses and the HTTP server to the other.

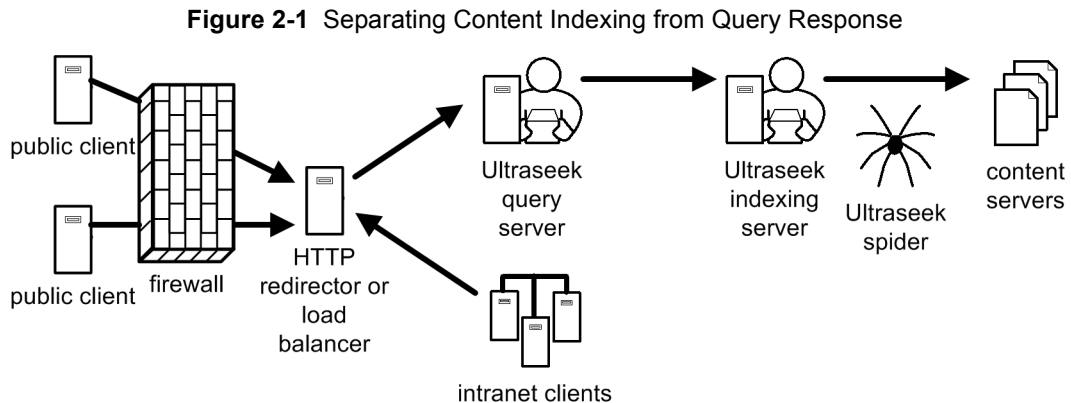
To bind Ultraseek to an IP address:

1. Click **Server>Parameters>Main** in the Ultraseek administrative interface.
2. Enter the IP address in the Binding Address text box.
3. Click **OK**.

Refer to your HTTP server documentation for information on how to bind your HTTP server to the other IP address on the host.

## Index Content and Service Queries on Different Hosts

Indexing content can be processor-intensive, and Ultraseek continuously allocates threads to this task. If your search application will handle a large query load, you may be able to improve your response time by indexing content and responding to queries on different machines, as shown in [Figure 2-1](#).



This approach requires that you install two instances of Ultraseek, one on each of the two machines. The indexing server on one host is dedicated to indexing content. The query server contains a mirror collection of each searchable collection on the indexing server. Once you create a mirror collection, Ultraseek automatically updates it.

Security considerations should determine whether you place corresponding index and query servers on the same side of the firewall. In general, if you want to hide your indexing server behind a firewall, you likely want to hide the replicated content on the query server as well. If you choose to mirror content across your firewall, you can choose the `HTTPS` protocol to transfer data between the multiple Ultraseek servers.

## Providing Failover with Mirror Collections

---

If you must guarantee that your search application is always available, you should replicate your content across multiple host machines. If you are already indexing content and servicing queries on different host machines, you can use the index server as your backup query server. Or, you can install an additional Ultraseek query server. Either way, you must create a mirror collection of your content on the backup query server. If the first query server fails, the backup responds to the query using the same content as the out-of-service server.

Mirror collections are automatically updated when the original collection changes. If your mirror collection contains dynamic content, you can manually configure how often an Ultraseek server revisits the original collection. However, the more frequently you revisit the original collection, the more time is spent replicating content. The needs of your search application should determine the proper trade-off between these demands.

---

**Note** You should use third-party HTTP redirector or load balancing software to direct all queries in an HTTP client session to the same query host. If HTTP client sessions are shared between the two Ultraseek query servers, users may receive duplicate results. In addition, the latency of a request is lengthened if a second query server must perform a search that another query server has already performed.

---

See the *Ultraseek Administrator Guide* for instructions on creating and configuring mirror collections.

## Accessing Ultraseek Through a Proxy Server

---

You can easily configure Ultraseek to provide search through a proxy server using the controls on the **Collections>Network** page of the Administrative Interface. Virtually all search functionality is available through a proxy server.

If you want to highlight query terms in search results obtained through a proxy server, you must specify an “in-document highlighting proxy server” on the **Server>Parameters>Advanced** page of the Administrative Interface. This proxy server can be the same proxy server you specified for Ultraseek on the **Collections>Network** page. See the *Ultraseek Administrator Guide* for instructions on how to configure a proxy server.

# Segregating Content by Collection

---

Consider using collections to segregate content based on the following criteria:

- Between user groups with varying access permissions
- Between sections of your intranet
- Between document languages

The advantage of this approach is that Ultraseek makes it easy to target specific collections with a query; you can easily customize your search page to query specific collections based on the identity of the user, whether the request originates from within or outside your intranet, or the language used in the search UI. For example, you might include all publicly available content in a collection that is used for public queries, but put all private content in another collection used only for queries originating from inside your organization. Depending on the number of documents in the prospective public and private collections, the performance impact of multiple collections may be outweighed by the ease of this configuration technique.

---

**Note** Ultraseek enforces a fifteen collection-limit for each instance of the server. The number of collections impacts the response time of the server.

---

Another technique is to create *virtual collections*, subsets of existing collections. You create a virtual collection by prepending a query prefix to each user query, which limits search results to documents that satisfy the query prefix and contain the query. In addition, you can assign an Ultraseek style to each virtual collection to provide multiple search presences from the same collection. See [“Providing Multiple User Interfaces” on page 28](#) more information.

See *Ultraseek Administrator Guide* for more information about creating virtual collections using the Administrative Interface. See *Ultraseek Customization Guide* for instructions on creating virtual collections by manually using the `qp` form variable.

## Restricting Access by IP Address for Extranet Installations

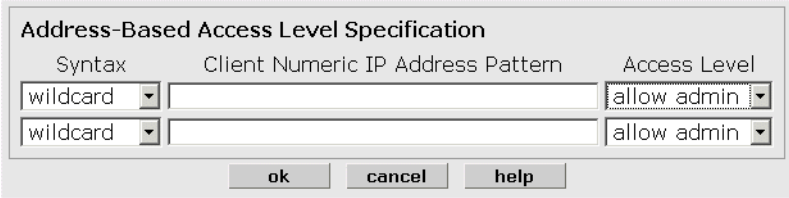
---

Because security is a greater concern when allowing access to external users, it's preferable to use IP address restrictions in addition to user accounts to secure Ultraseek on your extranet. For example, if you separated indexing from query response as described in [“Index Content and Service Queries on Different Hosts” on page 21](#) the query server needs to mirror the collections on the indexing server. You can restrict access to the indexing server on the extranet by configuring it to allow mirror-level access from the server responding to queries, but to deny requests from all other IP addresses.

To restrict access by IP address:

1. Click **Server>Users** in the Administrative Interface for the indexing server.
2. Enter the IP address of the query server in the “Address Based Access Level Specification” pane, as shown in [Figure 2-2](#).
3. Select **allow mirror** from the Access Level list box.
4. Click **ok**.

**Figure 2-2** Address-based Access Level Specification



The screenshot shows a dialog box titled "Address-Based Access Level Specification". It contains a table with three columns: "Syntax", "Client Numeric IP Address Pattern", and "Access Level". There are two rows in the table. The first row has "wildcard" in the Syntax column, an empty text box in the Client Numeric IP Address Pattern column, and "allow admin" in the Access Level column. The second row has "wildcard" in the Syntax column, an empty text box in the Client Numeric IP Address Pattern column, and "allow admin" in the Access Level column. Below the table are three buttons: "ok", "cancel", and "help".

Syntax	Client Numeric IP Address Pattern	Access Level
wildcard		allow admin
wildcard		allow admin

ok cancel help

Repeat this process for all query servers that need to mirror content on the index server.

# Localizing Your Search Application

---

It is highly recommended that you install one or more of the Ultraseek language modules if your documents are in multiple languages. The language modules automatically detect the language of a document, perform lexical analysis to improve the relevancy of search results, and provide a localized user interface for your search application. [Table 2-1](#) lists the languages supported by each Ultraseek language module:

**Table 2-1** Ultraseek language modules

Module	Supported Languages
<code>Ultraseek-platform-version-languages.exe</code>	<ul style="list-style-type: none"><li>■ English</li><li>■ Danish</li><li>■ German</li><li>■ Spanish</li><li>■ Finnish</li><li>■ French</li><li>■ Italian</li><li>■ Dutch</li><li>■ Norwegian</li><li>■ Portuguese</li><li>■ Swedish</li></ul>
<code>Ultraseek-platform-version-cjk.exe</code>	<ul style="list-style-type: none"><li>■ Japanese</li><li>■ Traditional Chinese</li><li>■ Simplified Chinese</li><li>■ Korean</li></ul>

You can download the language modules when you download Ultraseek.

# Authenticating Users

---

User authentication may occur when the Ultraseek spider indexes restricted content, when users attempt to log in to Ultraseek, or when Ultraseek filters search results based on user credentials.

## Indexing Restricted Content

The Ultraseek administrative interface allows you to specify credentials to index content at web sites that require a username and password. Similarly, you can create a form-based authentication specification if the web site requires this. We recommend that your search application use Ultraseek's SSL module for both these scenarios.

Ultraseek 5.3 supports the following authentication while crawling the web for documents:

- Basic + SSL (with or without X.509 client certificates)
- NTLM + SSL (with or without X.509 client certificates)
- Form-based + SSL

Contact your sales representative for more information about obtaining the SSL module.

## Filtering Search Results

Ultraseek can discard documents from search results that a user doesn't have permission to view. Before displaying search results to users, Ultraseek checks the permission requirements for each search hit. If the user has the required permission to view a document, Ultraseek displays a lock icon next to it in the search results. This is a server-level feature that applies to all searches.

To filter search results:

1. Click **Server>Parameters>Advanced**.
2. Check the "filter results based on user credentials" checkbox and then click **ok**.

---

**Note** There is performance penalty for enabling this feature, since Ultraseek determines what documents can be viewed when the query is submitted.

---

## Requiring User Login

You can require users to log in to Ultraseek before they submit queries. This style-specific parameter is limited to queries that use the specified *Ultraseek style*. An Ultraseek style is a set of server configuration parameters associated with a cascading style sheet (.css).

To require user login:

1. Click **Interface>Query**.
2. Select an Ultraseek style in the Style list box.  
Only queries that specify this style will require user login.
3. Check the “require login for result filtering” checkbox.
4. Click **ok**.

See *Ultraseek Administrator Guide* for more information on using the Style Editor to create and configure Ultraseek styles.

## Customizing Ultraseek’s Default Authentication

You can customize Ultraseek’s authentication scheme by modifying the `new_check_auth` function in the `patches.py` file, located in the `lib\python2.2` subdirectory of your Ultraseek installation. `patches.py` provides a Python interface to customizable functionality in the Ultraseek search engine. See the *Ultraseek Customization Guide* for more information on how to modify `new_check_auth` and other functions.

## Providing Multiple User Interfaces

---

Large companies frequently need to provide each department with its own search experience, such as a department-specific search interface and content. An easy way to provide a unique search interface is to create individual Ultraseek styles for each department with Ultraseek's Style Editor. An Ultraseek style consists of a set of Ultraseek server configuration parameters and a cascading style sheet (.css).

Using *query prefixes* and *suffixes*, you can set and associate user queries with an Ultraseek style. A query prefix is an initial search that is prepended to a query; the user sees only search hits that contain both the query prefix and the user query. A query suffix is a search appended to the end of a query. Unlike `qp`, `qs` expands search results; search hits may contain either the user query, the query suffix, or both. For example, the following query prefix specifies that all queries are made against indexed content from the Autonomy internet site:

```
site:www.autonomy.com
```

Query prefixes and suffixes provide an easy way to create multiple search presences, or virtual collections, from the same collection. You can use these style parameters to provide department-specific content without creating a new collection for each department.

---

**Note** Most server parameters, such as query prefix and suffix, can also be set manually by editing HTTP form variables within the HTML pages located in the `/docs` subdirectory, as described in the *Ultraseek Customization Guide*. However, this method requires administrators to re-implement their customizations each time they upgrade Ultraseek. It is highly preferable to use the Style Editor, so that your customizations are automatically migrated when you upgrade Ultraseek in the future.

---

See Chapter 4 of the *Ultraseek Administrator Guide* for more information on using the Style Editor.

# 3

## Improving Search Usability

This chapter describes ways in which Ultraseek administrators can increase the usability of their search applications.

- [Customizing Your Search Interfaces](#)
- [Classifying Your Content](#)
- [Creating Quick Links](#)

# Customizing Your Search Interfaces

---

Keep the following guidelines in mind when customizing your search and search results pages. Search pages contain the search box in which users enter their queries; search results pages list the search hits for those queries.

- Add a search box to the header of any HTML page served to a user. This allows users to search from anywhere in your search application. By default, Ultraseek places search boxes in its search and search results pages, but you need to add them manually to any other HTML pages that you serve to users.
- Display a thesaurus in all your search boxes. This helps users find appropriate keywords. In addition, edit your thesaurus to contain domain-specific terms used at your company.
- Send data to the Ultraseek server using HTML `GET` rather than `POST`. Using `POST` causes problems with expected browser behavior and caching. For example, clicking the browser's back button displays an annoying confirmation message to the user.
- Display related topics on your search page if you have a license for the Content Classification Engine (CCE). Research indicates that users are more successful at finding information when it is organized into topic hierarchies.
- Don't remove the search tips from your search page. They are likely the only search assistance your users will see. Instead, customize the tips to reflect local content.
- Enable spelling suggestions on your search page. The performance impact is minimal and the feature frequently helps users immediately correct a misspelled word without having to re-type the query.
- Simplify, rather than complicate your search results page. The smaller the footprint, the faster your users see results.
- Enable passage-based summaries in your search results pages. The performance impact is outweighed by the usability benefit of allowing users to quickly evaluate a document's content before opening it.
- Do not embed your search results pages within surrounding HTML page elements, such as frames. Doing so limits the number of search hits visible to the user on a single page.

See *Ultraseek Administrator Guide* for instructions on how to configure user interface and create virtual collections with the Style Editor.

# Classifying Your Content

Research indicates that users find information up to 50 percent faster when search is combined with an easy-to-navigate topic hierarchy. Ultraseek provides the Content Classification Engine (CCE) module for organizing your content into such browseable topics, or extending an existing organization into your search application. This is a cost-effective, basic categorization solution ideally suited to many departmental search applications. CCE topics are especially useful search aides when users don't know the right keywords, when they want to limit their search to a known subset of the available content, or just to make your search application look and feel more like a web portal.

Figure 3-1 Ultraseek Topics



Using the CCE Module, your application's 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.

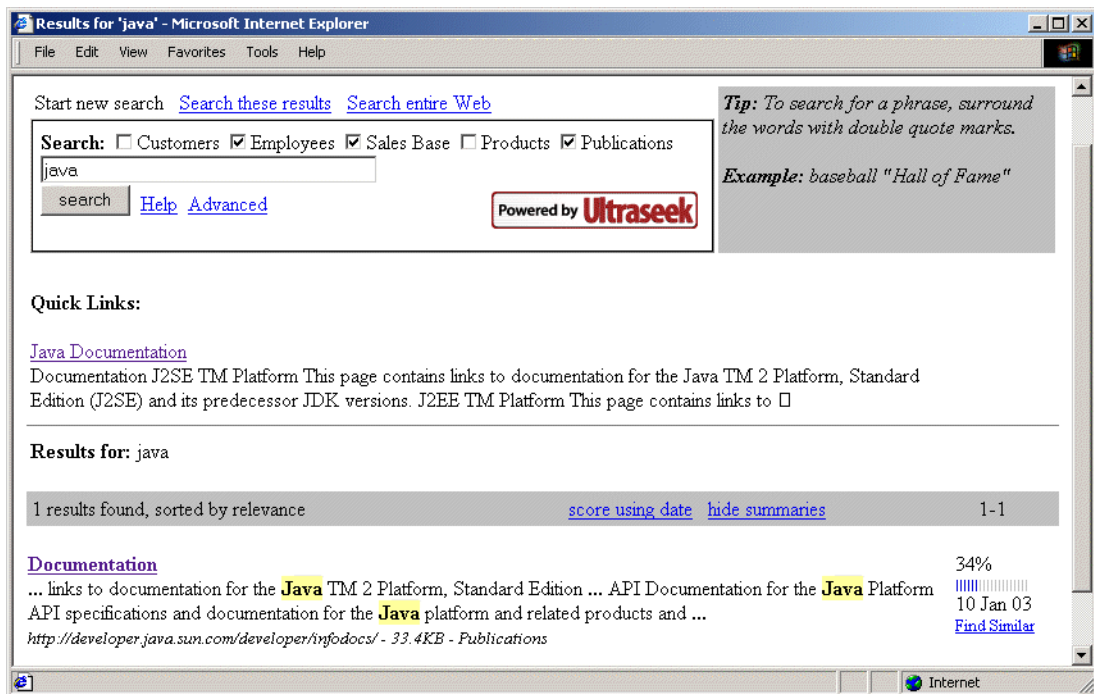
Creating a useful topic hierarchy requires familiarity with the content of documents in your repository. It is highly recommended that you obtain the assistance of a subject matter expert to create your topic tree.

You need to obtain a separate license to enable the CCE. After you have done so, click **Topics** in Ultraseek's administrative interface to see the panes used to create and edit topics. See the *Ultraseek Administrator Guide* for more information about creating topics with CCE.

## Creating Quick Links

A *Quick Link* is a URL that appears at the top of the search results page after a user searches for predefined keywords or phrases. You create these links in the **Server>Quick Links** page of the Ultraseek Administrative Interface by mapping keywords to URLs. When a user types one of these keywords or phrases, Ultraseek displays a Quick Links area above the search results, as shown in [Figure 3-2](#). 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 the query “Java” with a URL where Java documentation can be found, a search for “Java” will display a link to the URL at the top of the search results page.

**Figure 3-2** Quick Link to Java Documentation



Quick Links are particularly useful when users frequently submit the same query. Once you know which documents your user base is most interested in, use Quick Links to make it possible to retrieve the document without a query. Ultraseek 5.3 makes it easy for you to do this by placing a “Make a Quick Link” link next to each URL in a Top Requested Documents report.

## **Improving Search Usability**

### Creating Quick Links

You can generate this and other reports on the **Activity>Reports** page of the Administrative Interface. See *Ultraseek Administrator Guide* for more information about linking pre-defined keywords and phrases to specific documents and generating reports.

**Improving Search Usability**  
Creating Quick Links

# Glossary

<b>activity curfew</b>	Settings available from the <b>Collections&gt;Tuning</b> page that specify when a collection is updated.
<b>Administrative Interface</b>	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:  <code>http://hostname:port/admin</code>
<b>administrator</b>	Ultraseek user with administrative privileges. Administrators have permissions to modify and delete collections.
<b>binding address</b>	Associates a static IP address with an instance of 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.
<b>collection</b>	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.
<b>collection filter</b>	A URL pattern or site pattern on the <b>Collections&gt;Filters</b> page that specify which documents are allowed in a collection.

<b>Content Assistant</b>	<p>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:</p> <ul style="list-style-type: none"><li>Assign topics to documents</li><li>Filter documents</li><li>Add document metadata</li><li>Replace document titles and document description information</li></ul>
<b>Content Classification Engine (CCE)</b>	<p>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.</p>
<b>deduping</b>	<p>Removing from search results any duplicate search hits with the same URL, or with the same title and description.</p>
<b>document score</b>	<p>See <b>relevance score</b>.</p>
<b>Dublin Core</b>	<p>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 <a href="http://dublincore.org">http://dublincore.org</a> for more information.</p>
<b>HTTP keep-alive</b>	<p>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.</p>
<b>in-memory index</b>	<p>A search index for storing documents. When the in-memory index gets full, the spider saves it by creating an on-disk index.</p> <p>See also <b>on-disk index</b>.</p>

## Glossary

<b>index</b>	<p>1. <i>n.</i> A compilation of terms that can be searched. Each term in a Ultraseek collection index is stored together with a list of the spidered documents containing the term.</p> <p>2. <i>v.</i> To parse a document and enter its terms into a collection index. The Ultraseek <a href="#">spider</a> indexes documents.</p>
<b>indexer weight</b>	<p>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 <code>title:</code>, 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.</p>
<b>metadata</b>	<p>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.</p>
<b>on-disk index</b>	<p>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.</p> <p>See also <a href="#">in-memory index</a>.</p>
<b>Page Expert</b>	<p>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.</p>
<b>parse</b>	<p>To break down a document's content into its component parts. These components can then be used by Ultraseek to create index terms.</p>
<b>proxy server</b>	<p>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.</p>

## Glossary

<b>quality factor</b>	An integer between -16 and +15 used to adjust the term-dependent portion of a document's <a href="#">relevance score</a> for a query. Administrators assign a quality factor by specifying an <a href="#">URL pattern</a> 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.
<b>query term</b>	One part of a search query that users enter into the search input box.
<b>Quick Links</b>	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 <b>Server / Quick Links</b> pane in the administrative interface to define the keywords and phrases and specify the documents that are returned when searches contain these terms.
<b>regex</b>	A regular expression used to filter content in collections. A selection option for many of the Ultraseek Administrative interface fields. For example, administrators specify regular expressions to configure Ultraseek to work with specific IP addresses.
<b>relevance score</b>	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.
<b>root</b>	The URL of the first page visited by the <a href="#">spider</a> when building a collection.
<b>score</b>	See <a href="#">relevance score</a> .
<b>search</b>	To match query terms against terms in the collection index, and return links to documents containing the matching terms.
<b>spelling suggestion</b>	Ultraseek server parameter that, when checked, suggests re-spellings of query terms when a re-spelling may result in a more useful query.

## Glossary

<b>spider</b>	<p>[noun] The Ultraseek code that builds HTTP-based collections.</p> <p>[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.</p>
<b>spider throttle</b>	<p>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.</p>
<b>Style Editor</b>	<p>A feature in Ultraseek 5.1 that provides a graphical user interface for customizing the Ultraseek search and search results pages.</p>
<b>topic</b>	<p>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.</p> <p>You must license the CCE (Content Classification Engine) module to include topics in your search application.</p>
<b>thesaurus</b>	<p>A language-specific XML file that relates sets of query terms. When users enter a search query specified in a <code>&lt;show&gt;</code> element in the xml file, all other query terms in <code>&lt;show&gt;</code> elements within that set are displayed to the user as checkboxes that expand their search. By default, Ultraseek contains an english language thesaurus, <code>thesaurus_en.xml</code>, located in the <code>language</code> 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.</p>
<b>thread</b>	<p>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).</p>
<b>Ultraseek Query Language</b>	<p>The language defining the query syntax used by the Ultraseek server. See the <i>Ultraseek Administrator Guide</i> for more information about using plus, minus, and wildcard operators, vertical and double-vertical bars, and field searches in queries.</p>

## Glossary

<b>URL database</b>	A subset of an Ultraseek collection that holds information about URLs and links.
<b>URL pattern</b>	A URL path with <i>/*</i> and <i>*.extension</i> wildcards used to filter HTTP requests. The <i>/*</i> wildcard matches all URLs from that point forward, and the <i>*.extension</i> matches any file name with the specified extension.
<b>User Agent</b>	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 Ultraseek <a href="#">spider</a> 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.
<b>wildcard</b>	An operator that expands the results of a query search. Ultraseek supports two wildcards ( <i>*</i> and <i>?</i> ) in individual search terms, but not in search phrases.

# Index

## A

authenticating users 26

## B

BEA WebLogic 20

## C

classifying content 31  
content segregation 23  
creating Quick Links 32  
customizing authentication 27  
customizing search interfaces 30

## D

deploying Ultraseek 13  
deployment factors 13  
/docs subdirectory 28

## F

failover 22  
filtering search results 26

## I

IBM WebSphere 20  
improving search usability 29  
index server 21  
indexing restricted content 26  
intranet search 16  
IP address restrictions 24

## L

localized search 25

## M

mirror collections 22  
multiple IP addresses 20

## N

network configuration 14  
new\_check\_auth function 27

## P

portal integration 20  
providing multiple UIs 28  
proxy server 22  
public site search 17

## Q

query origin 15  
query prefix 28  
query server 21  
query suffix 28  
query volume 15  
Quick Links 32

## R

repository size 14  
request latency 15  
requiring user login 27

## S

search usability 29

## U

Ultraseek language modules 25  
user authentication 26

## Index

user login 27

## Y

Yahoo! PortalBuilder 20