Wrensoft Zoom Search Engine Indexer



Zoom Indexer is an easy to use Windows application that scans your website and indexes the information it finds, in order to generate the index data that will be required by the Zoom search script. It is a part of the *Zoom Search Engine* package, that provides custom search functionality for your website.

For more information visit our website: http://www.wrensoft.com/zoom

We recommend all users to <u>download the Zoom Users Guide</u> (available in PDF format as a free download from our website), as it provides more detailed information on how to use Zoom, and how the Zoom Indexer works in the context of the Zoom Search Engine package.

First-time users should try the Zoom Search Engine Wizard to step them through the main steps. Click on the magic wand in the toolbar to start the wizard.

The Zoom Indexer has two main modes of operation: *spider mode* and *offline mode*, for more information on each of these modes, click on the following topic:

<u>Scanning modes (spider mode and offline mode)</u> Differences between spider and offline indexing modes

<u>Spider indexing mode</u> Crawl your website online, by following HTML links

<u>Offline indexing mode</u> Index a copy of your website on your local hard disk

Other topics available:

<u>Search Platform (PHP, ASP, JavaScript, or CGI)</u> Which is the right platform for you?

<u>Configuring Zoom Indexer</u> How to configure the Zoom Indexer

<u>Defining your website categories</u> Enabling and defining categories and sub-sections of your website

<u>What to do after indexing</u> You've indexed your website, but what now? (Customizing appearance, etc.)

<u>Plugins</u> Indexing and searching PDF and DOC files (Standard and Pro only)

<u>Scheduler</u> Scheduling Zoom to automatically index or generate reports on a regular basis

<u>FTP</u>

Configuring Zoom to FTP upload the search files to your server.

Search statistics

Analyzing search logs and generating search reports such as "Top 10 searched words", etc.

Command-line parameters

How to run the indexer from the command-line using pre-specified config files.

System requirements

What you need to run the Zoom indexer application

Purchasing information

Information on the Standard or Professional licenses of Zoom available for larger websites, commercial development, PDF and DOC support, amongst other extra features.

Contacting WrenSoft

For more information, support, queries, or custom development questions

Scanning modes (spider mode and offline mode)

The Zoom Indexer is a Windows application that scans your website and indexes the information it finds, in order to generate the data files that will be required by the search script. It can operate in two distinct modes: *spider mode* and *offline mode*.

Spider mode

Spider mode indexes a remote copy of your website already uploaded and hosted on a web server. It does this via the use of a 'spider', which starts from a given start page, and follow the links it finds on each page. This allows the indexer to thoroughly index a website containing both static content (.htm and .html files which do not change) and dynamically generated content (such as websites with a PHP or ASP driven backend, message boards, etc.). This mode requires an established Internet connection.

To find out more, click here about Spider indexing mode

Offline mode

Offline mode indexes a local copy of your website, stored on your hard disk. This is effective for static web pages and allows the user to index a website without uploading it to a web server, maximising indexing speed and convenience. It can also be used for web pages that will be published on a disk or CD-ROM, where a web server will not be available and does not require an Internet connection.

To find out more, click here about Offline indexing mode

Spider indexing mode

Spider mode indexes a remote copy of your website already uploaded and hosted on a web server. It does this via the use of a 'spider', which starts from a given start page, and follow the links it finds on each page. This allows the indexer to thoroughly index a website containing both static content (.htm and .html files which do not change) and dynamically generated content (such as websites with a PHP or ASP driven backend, message boards, etc.). This mode requires an established Internet connection.

Start spider URL (spider mode only)

In spider mode, you are required to specify the URL from which the indexer will start the spider scanning from. Typically, you would point this to the entrance page of your website, (such as index.html) so that it will be able to find links to other pages on your website by following the links it finds on each page (as a visitor would).

Also note that the spider mode automatically skip links to external web sites, ie: those that are outside of the base URL defined (see below). This is to prevent indexing pages outside of the specified website.

Advanced spider URL options: Clicking on the "More >>" button will bring up a window which allows you to add more spider URLs or specify advanced spider crawling options. This is particularly helpful when indexing across multiple websites or domains. With each spider URL you can specify the following options:

- Index page and follow internal links: index the content of the specified page and follow any internal links (links to pages beginning with the base URL) found.
- Index page and follow internal and external links: index the content of the specified page and follow any internal and external links. However, external links are only followed up to one level. For example, an external page linked from an internal page is scanned, but an external page linked from an external page is not.
- *Index single page only:* index the content of the specified page and not follow any of the links found on it.
- *Follow links only:* only follow the links found on this page but will not index any of the page content.

You can also override the automatic base URL determined from this window, if necessary.

The *Import* and *Export* buttons allow you to mass-insert a list of start points from a text file. See chapter 6 in the Users Guide for more information.

Base URL

This is the URL where your website will be published and uploaded. For example, if your website will be published at http://www.myisp.com/~bob/index.html, then http://www.myisp.com/~bob/ will be the base URL of your website. This is used to determine the base location of each file on your website, so do not specify the filename of the main page (ie: index.html, home.htm, etc.). In spider mode, this is automatically determined for you based on the URL of the start page specified. However, you can override it from the Advanced spider options window (by clicking on "More>>", and selecting the URL and clicking "Edit").

Output directory

This is the directory in which the index files generated will be saved. Usually you would put this in the same directory as your website files so that they can be uploaded together. Note that it must be uploaded to the same directory as where the search script files will be located, so it is recommended you keep them in the same directory as these two files.

If you have "PHP/ASP/CGI" selected in the configuration (default), four data files will be generated in the

output directory: zoom_dictionary.zdat, zoom_pagetext.zdat, zoom_pages.zdat, zoom_titles.zdat, and zoom_descriptions.zdat.

If you have "JavaScript" selected in the configuration dialog, only one data file will be generated in the output directory: zoom_index.js

Platform

This setting selects the script platform, which you will host or distribute your search engine with. Select from either PHP, ASP, or CGI format for server-side searches, or JavaScript format for client-side searches. If you do not know which format best suits your method of distribution, please see <u>Output</u> <u>Format</u> for more information.

Verbose mode

Verbose mode displays information on all files the indexer comes across, including files that are not scanned and indexed. This can be useful in determining which files are excluded from the index of your search function or refining your search engine's effectiveness.

Zoom Indexer can be further configured by clicking on the *Configure* button (or the *Configure* option in the *Indexer* menu). Click here for more information: <u>Configuring Zoom Indexer</u>

Offline indexing mode

Offline mode indexes a local copy of your website, stored on your hard disk. This is effective for static web pages and allows the user to index a website without uploading it to a web server, maximising indexing speed and convenience. It can also be used for web pages that will be published on a disk or CD-ROM, where a web server will not be available and does not require an Internet connection.

Start directory (offline mode only)

The start directory specifies the local directory for the offline scanning to begin in. All sub-directories under the start directory will also be scanned. In other words, you should point this at the folder in which you have created the files for your website.

You can also specify multiple folders to index from, by clicking on the "More>>" button.

Base URL

This is the URL where your website will be published and uploaded. For example, if your website will be published at http://www.myisp.com/~bob/index.html, then http://www.myisp.com/~bob/ will be the base URL of your website. This is used to determine the base location of each file on your website, so do not specify the filename of the main page (ie: index.html, home.htm, etc.). In spider mode, this is automatically determined for you based on the URL of the start page specified.

For CD/DVD distribution: For offline/local distribution, the base URL should be the relative path from where the search script will be. It does not have to be a HTTP address. For example, if the files you have indexed will be placed in the same directory as the search.html page on the CD, use "./" (which refers to the current directory) as your base URL. If all the files to be searched will be in a folder named "data" on the CD, then you would use "./data" as the base URL. This is generally recommended over the use of absolute paths such as "/" or "C:\myfiles\", which will raise compatibility issues between different operating systems (such as Macs).

Output directory

This is the directory in which the index files generated will be saved. Usually you would put this in the same directory as your website files so that they can be uploaded together. Note that it must be uploaded to the same directory as where the search script files will be located, so it is recommended you keep them in the same directory as these two files.

If you have "PHP/ASP/CGI" selected in the configuration (default), four data files will be generated in the output directory: zoom_dictionary.zdat, zoom_pagetext.zdat, zoom_pages.zdat, zoom_titles.zdat, and zoom_descriptions.zdat.

If you have "JavaScript" selected in the configuration dialog, only one data file will be generated in the output directory: zoom_index.js

Platform

This setting selects the script platform, which you will host or distribute your search engine with. Select from either PHP, ASP, or CGI format for server-side searches, or JavaScript format for client-side searches. If you do not know which format best suits your method of distribution, please see <u>Output</u> <u>Format</u> for more information.

Verbose mode

Verbose mode displays information on all files the indexer comes across, including files that are not scanned and indexed. This can be useful in determining which files are excluded from the index of your search function or refining your search engine's effectiveness.

Zoom Indexer can be further configured by clicking on the *Configure* button (or the *Configure* option in the *Indexer* menu). Click here for more information: <u>Configuring Zoom Indexer</u>

For more information and troubleshooting, refer to the User's Guide (available in PDF format as a free

download from our website <u>http://www.wrensoft.com</u>). Or go back to the <u>Help overview</u>

Configuring Zoom

Zoom Indexer can be further configured by clicking on the *Configure* button (or the *Configure* option in the *Indexer* menu). Here you can define the behavior and operation of the indexer.

The configuration window consists of the following tabs:

General Search Page Scan Options Skip Options Indexing Options Results Layout Categories Limits Authentication FTP Languages Status Log Advanced

Configuring – General settings

Do not show wizard on startup

Stop the wizard from being displayed on startup.

Beep at the end of indexing

Play the system alert sound when indexing completes.

Spider downloading options

These are options which control the way in which Zoom will download files when it is in spider mode. Note that it does not apply for offline mode indexing.

Single-threaded downloading

This option uses only one dedicated thread for downloading files. This is typically the slower option but provides reasonable speed when indexing a site with a fast connection. It is also recommended when you are trying to follow the spider's crawling path, to determine if it is scanning the pages you are expecting.

Multiple threads

This option allows you to specify more than one dedicated thread for downloading files in spider mode. This is recommended to increase the speed of indexing. It allows Zoom to download multiple files in the background whilst indexing at the same time.

Reload all files (do not use cache)

Check this option to ensure that all files are downloaded from the site and that the cached copy of pages are not used.

Configuring – Search Page

Search Form Appearance

The search form is the actual search box where a user enters a search query. Here you can choose from "Advanced" (with full options) to "Basic" (simple search box), to "Do not generate". With the latter option, the search box is not generated by the script, allowing you to create your own search form in your search template HTML file (or provide the search interface elsewhere, such as in a separate frame). Note that the generated search form is useful because it can "remember" the entries of the previous search query.

Results linking

This option defines how clicking on a search result link will open the page. The options available are:

- The current window (default)
- A new window
- A frame or window with the specified name.

Note that this is effectively equivalent to a target= attribute in an A HREF link in HTML.

Default to "match all search words"

This will set the default searching method to "match all search words" instead of "match any search words".

Show time taken to perform search

When checked, this will display a "Search took x seconds" line at the bottom of each search page.

Show Zoom info line

When checked, this will display a "Powered by Zoom Search Engine" line at the bottom of each search page.

Exact phrase

This allows users to search for words in a specific order as they appear on a page by surrounding the words in quotes (eg. a search for "wrensoft zoom" will not match "zoom wrensoft" or "wrensoft something zoom").

Note: Exact phrase searches can be significantly exhaustive (and thus, slow) depending on how common the words in the search phrase are, on your site. Because of this, it is a user-defined option whether to allow users to enter exact phrase searching (or ignore them and treat them as multiple keywords, as in previous versions).

The *Max Context Seeks* value allows you to eliminate the chances of slow, exhaustive searches. By lowering this value, you decrease the chances of exhaustive searches. When a phrase is entered that is too slow to search thoroughly, a limited set of results are returned and a note is displayed advising the user to specify a more precise search phrase (usually by adding some less common words to the existing phrase). The recommended value for *Max Context Seeks* is around 500.

This option is not available for the JavaScript version, or when context descriptions are disabled.

Provide option to sort results by date or by relevance

When checked, your search page will have the option to switch between "Sort by relevance" or "Sort by date".

Provide spelling suggestions when less than x results found

When checked, your search page will offer spelling suggestions when less than a certain number of results were found.

Configuring – Scan options

Scan Extensions

This is a list of the filename extensions that will be scanned by the indexer. If a file has an extension that is not in this list then it will be skipped. Note that each file extension must begin with a '.' character or it will be ignored. Up to 20 extensions can be defined in this list.

Checking the "Scan files with no extensions" options allows you to index files with out any extension (files such as "Home" and "index").

In Zoom Standard and Professional Edition, you can specify file extensions such as ".pdf" and ".doc" which are supported by external plug-ins. When you do so, you can tick the option to "*Use .desc files for plugin extensions*" to use custom meta description files for these documents. See the section titled Zoom Indexer Plug-ins – Using custom description (.desc) files in this document for more information.

Duplicate page detection

Checking this option enables the use of CRC-32 signatures to ensure that only pages with unique content are indexed. This is particularly useful for spidering websites with links to pages without a filename, for instance, to a directory (eg: http://mywebsite.com/). These links may otherwise be indexed twice if there is another link somewhere else on the website which points to the same place, but with the actual filename specified (such as http://mywebsite.com/index.html, http://mywebsite.com/home.htm, etc.). It is best to avoid this on your website and use a consistent linking method. However, you can also prevent this by turning on this option.

Plugin options

Use meta information from plugins

This feature extracts meta information (where available) from the plugin supported files. For example, Author, Subject, and Keywords from PDF files.

Configuring – Skip options

Page and folder skip list

This is a list of pages and folders that will not be scanned during the indexing process. Note that filenames and paths are case sensitive. Typically you would want to filter pages that the user should never be able to get to directly via the search function. Note that if the path to a page partially or fully matches any entry in this list it will be filtered. For example, an entry of "\private\" will filter "\private\ file1.htm", "\private\file2.htm" and "photos\private\athome.htm".

Skip files or directories that begin with an underscore

When this is enabled, files or directories beginning with an underscore (such as those kept by Dreamweaver and FrontPage) will be ignored and skipped in the indexing.

Word skip list

This is a list of words that will be filtered during the indexing process. Typically you would want to filter some small words that appear on every page such as, "and", "or", "the", etc. Depending on your site you may also want to skip other words. For example, if your site had a thousand pages about different aircrafts, you might want to remove the word "aircraft" from the index, as it would match every page on your site.

If a word is included with a '*' as the first letter then it will match and filter any word that contains this word. eg. "*fun" would filter "fun", "funny" and "nofun".

Note that directories and files that have names that start with underscores are always skipped.

Skip words less than x characters: You can now specify the minimum number of characters that a word must be before it is indexed. It is defaulted to two characters so all single-character words are skipped.

Configuring – Indexing options

What to index

You can specify the parts of a page that should be included or excluded from indexing here. This includes the page title, content, and filename. Meta information can also be indexed such as meta descriptions, keywords, and author information.

By excluding certain sections of pages, you can make index data files smaller and the indexing procedure faster, and less memory intensive.

Word Boosting

These drop-downs allow you to increase or decrease the importance of words found, depending on whether it was found within:

- The title of the page
- The meta description of the page
- The Zoomwords or meta keywords tag of the page

The boosting can be specified between 10 levels, ranging from -5 Deboost to Normal, to +5 Boost.

For example, if the Title is set to have a word weighting of "+2 Boost", then each words which appear within the page title would be treated as though have appeared twice on that page.

This makes it possible for you to give priority to pages (placing it higher up in the results), when the word was matched in the title, description or keyword tag..

You can also boosts ALL words found on specific pages, by use of the ZOOMPAGEBOOST meta tag, eg.

<meta name="ZOOMPAGEBOOST" content="5">

Putting this on the most important page on your site would help make it appear higher up in the search results. You can probably use less than 5 to do the same thing on a small site. Similarly, a negative value would decrease the weight of words on that page.

Indexing words

This allows you to specify which characters should be allowed to act as a join character between two words. Otherwise, these characters will act as separators of words (for example, if the 'dash/hyphen' character is a join character, words such as "web-based" will be indexed as one word. Otherwise, it would be split into two words, "web" and "based"). Note that the character must be immediately preceded and followed by another valid character to be indexed.

Configuring – Results layout

Search results layout

This set of options allows you to modify the appearance of the search results. You can select or disable the elements that should be displayed per search result.

Result number

Displays the rank number of the search result, with the most relevant page being numbered 1.

Title of page

Displays the link to the result with the page title as the link text. When disabled, the link text will simply be the URL.

Meta description

Displays the the meta description when available.

Context description

Displays the surrounding words from the content of the page where the searched word was found. Note that this is not available when Page Content is not being indexed. It is also not available for the Javascript version due to technical constraints with the scripting language.

Context size

The number of words surrounding the searched word that should be displayed as part of the context description.

Terms matched

The number of user searched terms that were matched on this page.

Score

A score representing the relevance of the page found, depending on the number of words matched on the page, and where the word was found on that page (can be configured with Word boosting parameters set in the Indexing Options tab).

Date

Displays the date for the page. This is determined by the Last Modified date and time for that particular file. When this information is not available, the file is given the date and time that the indexing took place.

Highlighting options

Words matched in search results

This option enables the highlighting of keywords found in the search results. You can customize the appearance of this highlighting by changing the CSS of the search page template. See "Fonts and colours" below.

Jump to match and highlight within document

This will enable support for highlighting and scrolling to the matched word on the actual web page when you click on a search result link. Note that this only works for HTML web pages, and does not work for Word, PDF, XLS and other document types. It also will not work for Javascript disabled browsers.

Once you have enabled this option and re-indexed your website, you must locate the "highlight.js" file included with Zoom. It can be found in the Zoom program folder (usually "C:\Program Files\Zoom Search Engine 4.0\highlight.js"). Copy this file to your website, as you will need to link to the "highlight.js" file from each page of your site that requires the feature.

For example, you could paste the following HTML in your site's header or footer:

```
<style>.highlight { background: #FFFF40; }</style>
<script type="text/javascript" src="highlight.js"></script>
```

Note that you will need to specify the correct path to "highlight.js" depending on where the file is located relative to the page.

You will then need to modify the BODY tag on your page to include an "onLoad" attribute, such as:

<body onload="highlight();">

If for some reason you can not modify the body tag of your page, an alternative would be to put the following line after the </body> tag of your page:

<script type="text/javascript">highlight();</script>

Once you have successfully applied the above changes, you should find that clicking on a search result will go straight to the word found and all matching words will be highlighted.

Fonts and colours

You can change the fonts, text colors, and styles of your search results by editing the CSS styles defined at the top of the template HTML file. Clicking on the Edit HTML template button will open up the search template source file in your default HTML editor.

See chapter 5 in the Users Guide for more information and a definition of the CSS styles you can use to change the appearance of the search results.

Configuring - Limits

Here you can set the limits for the indexer, to control the size of your index files and also to use memory more efficiently.

In the free edition of Zoom, these limits are restricted to a typical size of a free, personal website (50 pages, 15,000 unique words, 100,000 bytes, and 150 characters respectively).

For larger personal sites, or small businesses, we recommend the Standard Edition upgrade which is capable of indexing websites up to 100 pages and provide <u>plugins</u> support for PDF, DOC and other file types.

For commercial developers or very large sites, there is the Professional Edition which allows you to modify the limits as you require, effectively allowing you to index websites of unlimited sizes (limited only by the amount of memory in your indexing computer). The Professional Edition also supports <u>plugins</u> and additional file types.

For more information on the differences between these three editions, please see our webpage at http://www.wrensoft.com/zoom/editions.html

Max pages to scan

This specifies the number of web pages or files that can be scanned and included in the index.

Max unique words

This specifies the number of unique words (NOT total words) that can be indexed. Note that the English dictionary generally has around 40,000 unique words.

Max. file size scanned

This is the maximum size in KB (kilobytes) of a file that can be scanned. This is *not* the total size of all files indexed, just the size of the largest file to index. Also note that it is specified in KB, which means 1024 KB = 1 MB.

The notice message that appears when you specify over 10 MB is only there to warn users who accidentally over-specify this amount, due to confusing KB with bytes.

Max Description length

The number of characters used for the description of each page, as shown in the search results. By increasing this amount, you can include a larger portion of text as the page description in your search results.

Limit pages per start point

Allows you to specify a limit for each start point of the indexer, before it stops, and moves on to the next start point.

Configuring – Authentication

This tab allows you to enter login and password details to spider a secure site that requires authentication.

Note: Authentication information is saved in the .zcfg file when you select Save configuration from the file menu. The password is obfuscated, but not heavily encrypted. For sites with security sensitive information, we recommend creating a special user account for indexing on the web server where possible, so you can disable this user account after indexing.

Configuring – FTP

Zoom provides built-in FTP capabilities so that it can upload the files required for your search function, directly to your website.

Automatically upload files at the end of indexing

This option is only accessible from the FTP tab of the configuration window. When enabled, Zoom will automatically attempt to upload your files every time you finish indexing. We recommend you only use this when you have confirmed that your FTP settings are correct, and that the files are uploaded to the correct folder or path.

This option is particularly useful for scheduled indexing, allowing you to configure Zoom to update your website without any user interaction, and on an automated, regular basis.

FTP server

Specify the IP address or domain name of your ftp server, eg. "<u>ftp.myserver.com</u>". You can also specify the port number (default port for FTP is 21).

Username and Password

Specify the login information to your FTP account.

Folder or path on server

This should be the relative or absolute path on your FTP server where the files are to be uploaded to. Note that the folder must already exist. Check with your web host (or use a full browsing capable FTP client) to confirm the correct path if you are not sure. Typically, it may be something like "public_html/search" or "www/search".

Configuring – Languages

Zoom provides options to cater to searching sites of different languages and encodings.

Encoding and character sets

Change settings regarding foreign language support and character encoding. Firstly, if your website uses Unicode UTF-8, you must enable the "Use Unicode" option. Otherwise, specify the encoding (also known as charset) used. "windows-1252" is the most common option for English, French, German, and a number of other Latin based languages.

International searching options

Enable accent insensitivity

This will map all occurrences of accented characters to their non-accented equivalent (eg. ó, ò, ô, õ will all be treated as "o"). With this enabled, a user can enter the search word "cliché" and it will find all occurrences of the word on your website spelt as either "cliché" or "cliche". You can now specifically enable or disable this feature for accents (ó, ò, ô, õ, etc.), umlauts (ä, ë, ï, ö, ü), and ligatures (å, ø, æ).

Support single-case languages (eg. asian languages)

This should only be used if you are using a language where there is no case-difference and problems can occur when the script or indexer attempts to convert case (such as some east Asian languages).

Substring matches for all searches

This will make the script consider search words which occurs within another word to be considered a match. Eg. a search for the word "hot" will match "hotcake", "shotgun", etc.

Please see the foreign language support section of the PDF Users Guide for more information.

Search page language

You can modify the text that appears on the search page and search results, by customizing the Zoom Language Files (.ZLANG files). Almost every bit of text on the search page can be modified or translated, including "Search results for..." and "x results found", etc. This allows you to translate the search page to the language of your choice, without having to modify the search script.

Zoom also comes with a few pre-translated language files which both serve as examples, and allow you to create French or German search pages straight out of the box, by selecting it from the dropdown menu.

The ZLANG file format is simply a line-by-line entry, beginning with a STR key word (such as STR_FORM_SEARCHFOR), followed by an equals sign, and the text to be used. Note that an entry can not span multiple lines (disable Word Wrap in Notepad). Also, the STR key words are predefined and can not be modified.

Use the "English.zlang" default file as your reference for text which can be translated, and the overall format of the file.

NOTE: All ZLANG files must be in Unicode encoding. You can save a file in Unicode with Notepad, by clicking on the File menu, and selecting "Save As...". Locate the 'Encoding" dropdown menu at the bottom of the Save As file window. Select "Unicode" here, and save your file.

IMPORTANT

The default ZLANG files that come with Zoom ("English.zlang", "French.zlang", etc.) will be overwritten when you re-install the software. This is necessary because there may be new lines of text required in future updates of Zoom. However, you will be able to find an automatic backup of your existing ZLANG files in a folder named "lang-backup", under your Zoom program folder.

Another approach is to create new files with different filenames when customizing, e.g. "MyFrench.zlang". This way, they will not be overwritten when you re-install the software.

Configuring – Status Log

You can configure the type of messages that are displayed in the main status log window during an indexing session here. This allows you to configure Zoom to only display error messages or any information as necessary.

Note that these settings only apply when the "Verbose" button on the main window is not depressed.

Click the "Log all messages" button to quickly check all options. Click the "Reset to default" button to restore default log settings.

Scanned: Messages relating to which files have been scanned successfully.

Skipped: Messages about files which have been skipped or not scanned.

Queued: Messages about which links have been queued for downloading

Initialization: Messages regarding initialization of Zoom

Thread Info: Messages regarding thread activity (in spider mode only)

Downloading: Messages regarding files being downloaded.

Uploading: Messages regarding files being downloaded.

File I/O: Messages regarding writing and reading of files such as the index data files.

Information: Messages regarding miscellaneous information.

Error: Messages regarding critical errors whilst indexing.

Warning: Messages regarding minor problems or errors with indexing.

Plugin: Messages regarding the use of plugins or processing of external binary file formats such as PDF, DOC, WPD, etc.

Summary: End of indexing summary.

Configuring – Advanced

Custom script source path

If you wish to use a copy of the search script (search.asp, search.php or search.js) that is located in a different folder besides the one in the Zoom installation directory, then you can change the path here. This is only necessary if you are making modifications to the search script source code, and you wish to maintain multiple versions to be used with different sites.

Search logging

This enables search word logging so that searches made on your website will be recorded into a serverside log file. Note that you will have to specify the filename and path of the location of the log file on the server-side relative to the location which you will place the search script. You will also have to be able to change server-side file permissions to allow *write access* to the log file you specify. We recommend this option to advanced users familiar with server side file permissions only.

This option is not available when the JavaScript platform is selected. This is due to the fact that JavaScript is a client-side script, so logging of user searches is not possible.

Embedding script

This option allows you to specify the URL used to link back to the search script. By default, it should be left blank.

However, if you wish to embed the search script in your own server-side scripts (eg. you have included search.php into your own PHP file, or included the CGI within a PHP script, etc.) you can then specify the filename of your embedded script here so that it will link back to the preferred file (instead of always referring back to the original search.php file).

PDF plugin options

You can select the conversion process used by the PDF plugin here. Due to the nature of PDF files, certain page layouts can cause issues with indexing, such as pages in columns. If you find that certain words in PDF files are not returning results when searched, or if the context description looks like it has combined sentences from different columns, then switch to the "raw formatting" option provided here. Otherwise, it is best to use the default option.

Configuration file

Optional Configuration file (zoom.zcfg)

The settings of the indexer, including the starting directory, URL, output directory, and configuration settings can be saved as a configuration file. This provides you with the convenience of simply loading the configuration file the next time you wish to re-index your website. By default, the application loads the 'zoom.zcfg' configuration file in its current directory, but this can be changed by the optional command line parameters mentioned in the following section.

If you develop multiple websites, you can also keep a separate configuration file for each website and load up the one you wish to re-index.

Optional Command line Parameters

The Zoom Indexer accepts command line parameters if you wish to use it with an automated scheduling program or just to create a convenient shortcut. The usage definition is:

ZoomIndexer [options] <config-file>

Options:

-5	Auto-run in <i>spider mode</i> , start indexing as soon as the application loads. This option requires that a config-file be specified also.
-0	Auto-run in <i>offline mode</i> , and start indexing as soon as the application loads. This option requires that a config-file be specified also.
-r	Auto-run <i>Statistics Report Generator</i> . This will automatically create the search statistics page based on the report options saved in the config file specified.
<config-file></config-file>	Configuration file to load on start up.

Note that in either of the auto-run modes, the indexer will exit immediately after indexing has completed and an output log file is saved to "zoomindexer.log" in the running directory.

Categories

This feature allows you to specify categories within your search engine, allowing your users to select from a dropdown to restrict searches to certain sections of your site. Each search result will also be tagged to indicate the category that they belong to. Searching with the "All" category will return results from all categories.

To enable categories support, go to the Configurations dialog described in the previous section, and click on the "Categories management" button. This will bring up the following dialog, where you can add, remove, and edit the categories.

A category definition has three fields:

- 1. Name: The name of the category must be unique. These category names will be listed in the dropdown box of the search page.
- 2. Pattern: The pattern is used to determine what pages belong to the category. It is matched against a page's full path or URL. Should the pattern text appear anywhere within a page's URL, the page will be filed under that category. Note that this includes the base URL or domain name of each page, so that you can index multiple domains, and have each defined as a separate category.

For example, a pattern of "test" will collect the following pages: <u>http://www.mysite.com/test.html</u> <u>http://www.mysite.com/test/index.html</u> http://www.test.com/ ...etc.

Hint: You can use patterns such as ".pdf" to create categories based on file extensions.

3. Description: This is a short description of the category that will only be used internally in the indexer for your own convenience in the future. It is not a required field.

You can use the up/down buttons to the right of the list to re-arrange the order of the categories.

The order of the categories determines two things:

- The order of which they will appear in the dropdown list of the search page.
- The process of selecting which category a page belongs to. This means that if a page is deemed to qualify for the first category, it will not qualify for any other categories below it in the list.

Note: all pages that do not qualify for any of the categories in the list will belong to a "catch-all" default category. You can specify the name of this category in the "Default category name" field.

Once you have enabled, and defined your categories, the next time you index your site will produce additional index data.

In PHP/ASP/CGI mode, two additional index files will be generated: zoom_cats.zdat and zoom_catpages.zdat. Upload these files along with your other index files (zoom_dictionary.zdat, etc.), and the search script will automatically recognise the categories, and enable category support. To disable category support, you will have to **remove these two files** from your web server.

In JavaScript mode, the additional data will be appended to the zoom_index.js data file. No additional files or settings are necessary, just put it in the same directory as the search page and script, and load it up in your browser as usual.

Scheduler

This feature allows you to setup Zoom to automatically index your site, upload the files, or generate statistical reports on a regular basis.

You will first need to make sure that your current indexing configurations are correct, and that you have these settings saved in a (.zcfg) configuration file. Make sure that the saved settings in this file will successfully index your site to your liking.

You can then click on the scheduler button on the toolbar, or from the "Tools" menu, select "Schedule automatic indexing".

In the Zoom Scheduler window, you can add, modify, and remove scheduled tasks for various configuration files. Note that you can only schedule a task for the currently selected config file, but you can modify or remove existing scheduled tasks for other configuration files.

Add task

Click on "Add", and select the task option of either running in Spider mode, Offline mode, or to generate reports. Note here that you would need a separate task to index and generate report, they can not be achieved with one scheduled task.

Note also that you MUST specify a username and password for the user account that will be running the scheduled task. On Windows XP, you can leave this empty, and it will use the currently logged in user. However, on most other versions of Windows, you will require both Username and Passwords to be filled in – and you must use an account that has a password specified.

You can also enable the "wake computer to start task" option if required.

You can then click on the "Set schedule" button and create a schedule for the task (by clicking on "New"). A large variety of schedule combinations can be configured here, with options for Daily, Weekly, Monthly, etc. You can also click on the Advanced button for even more flexible scheduling options.

Change schedule

To modify the schedule for a task, simply select the task from the list and click on "Change schedule".

Remove task

To stop or remove a scheduled task, select the task from the list and click on "Remove task".

Plugins

To search and index non-text files such as DOC, PDF, XLS, and PPT files, you will need to install plugins for these file formats. Plugins are only supported in the Standard and Professional editions of Zoom.

You can download plugins from <u>http://www.wrensoft.com/zoom/plugins.html</u>. You will also find a list of the file formats supported here.

Installing a plug-in

Download the plug-in required from the Zoom website. Locate the directory in which the Zoom Indexer application is installed (ZoomIndexer.exe). By default installation, this should be in the "C:\Program Files\ Zoom Search Engine\" directory. Within this directory, there should be a "plugins" directory. Simply unzip the contents of the file downloaded from the website to this folder, and restart Zoom Indexer. When you start Zoom Indexer, the log window will now indicate that it has found the plug-in and that the file extension it supports is now enabled.

Using the plug-ins

Once the plug-in is installed, all you have to do is open the Configure dialog window, and add the file extension supported to the extension list. For example, ".doc" for Word document support. Click OK and start indexing.

Upgrading a plug-in

As new versions of the plug-ins become available on the website, simply copy them into the same "plugins" directory, overwriting the old plug-in.

Using custom description (.desc) files

As many external binary documents do not contain useful title and description information, Zoom allows you to specify custom Meta information for any plug-in supported files. The option can be enabled in the indexer configuration dialog (see Indexer Configuration for more information), and once it is enabled, the indexer will attempt to look for .desc files for any plug-in supported file extensions.

For example, if you have a file called "mydocument.doc", you can create a text file called "mydocument.doc.desc" in the same directory with the following contents:

```
<title>This is my document custom title</title>
<meta name="description" content="This is my document's custom description">
```

Zoom will then index the words found within "mydocument.doc", but use the title and description information found in "mydocument.doc.desc" – so that you will see your custom title and description in your search results. You can also specify extra keywords with the ZOOMWORDS or KEYWORDS fields (refer to the Users Guide PDF and the section on *Optimising and configuring your search engine – Manually add words to the index*).

Search Platforms (PHP, ASP, JavaScript, or CGI)

The index data generated by the Zoom Indexer must be in the right format for the search script you intend to be using. The Zoom Search Engine package comes with three search scripts: PHP, ASP, JavaScript, and CGI.

If you are unsure which script platform you should use, refer to chapter 1.8 of our Users Guide (available for download from our website at <u>http://www.wrensoft.com/zoom)</u>.

The server-side versions (**PHP** or **ASP** or **CGI**) provide the best performance as it operates on your own web server and is the recommended choice for an online website. You should always use this option when you have a web server with support for PHP, ASP or CGI.

The client-side version (**JavaScript**) lets you run the search engine from a CD-ROM or DVD distribution and has no additional requirements besides being viewed by browsers with JavaScript enabled. You should use this only for offline distributions where a web server may not be available, or where network traffic and latency is not an issue.

The **CGI** option is a server-side alternative to the PHP and ASP options, catering for the largest sites, which requires even faster searching than the scripting options above provide. Consider this option if the PHP or ASP version does not provide adequate performance in handling the size of your site.

You can also find out more about the scripting languages and versions available by clicking on one of the following links:

What is PHP?

What is ASP?

What is JavaScript?

What is CGI?

More information on using the different search scripts available can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).Go back to the <u>Help overview</u>

What is PHP?

PHP (PHP: Hypertext Preprocessor) is a software package installed on web servers to provide scripting capabilities. It is commonly found on web servers running Apache (as it is now built-in by default), but it can be installed on IIS servers also. Since March 2004, there are a reported 15,528,732 website domains which use PHP (source: Netcraft Survey).

You can check with your web host to find out if PHP is available on your hosting account. Alternatively, you can upload a text file containing the following (as test.php):

```
<?php
phpinfo();
?>
```

When you open this file up from your browser (http://www.mysite.com/test.php), it should provide detailed information on the PHP installed on the server. If not, this means you do not have PHP support configured on your server and you should consult your web host or use one of the other platform options available.

More information on PHP can be found at http://www.php.net/.

If you are unsure which script platform you should use, refer to chapter 1.8 of our Users Guide (available for download from our website at <u>http://www.wrensoft.com/zoom)</u>.

More information on using the different search scripts available can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).

Go back to the <u>Help overview</u>

What is ASP?

ASP (Active Server Pages) is the Microsoft equivalent to PHP, and comes packaged with most default IIS (Internet Information Services) web servers. Chances are if you have a Windows based web server, you will have ASP available to you. It can also be available on some non-Windows Apache servers, but this is less common.

Check with your web host to find out if ASP is available on your hosting account.

Note that ASP is not the same as ASP.NET, which is a new Microsoft server-side platform and they are not compatible.

If you are unsure which script platform you should use, refer to chapter 1.8 of our Users Guide (available for download from our website at <u>http://www.wrensoft.com/zoom</u>.

More information on using the different search scripts available can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).

Go back to the Help overview

What is CGI?

CGI (Common Gateway Interface) is a method of running programs on a server over the web. This is different to PHP and ASP in that it does not have to load and interpret a "script", and is not limited by the technical capabilities of a scripting platform. In fact, the scripting engines for PHP and ASP are CGI applications themselves. As such, CGI provides a way to run web applications requiring maximum performance and efficiency, and you will find it used on most enterprise-scale sites such as popular sites like eBay, Google, and Yahoo.

Due to the less restrictive nature of CGI applications, some web hosts (especially those offering cheaper packages) do not provide CGI support for security reasons. In addition to this, setting up and installing CGI applications can be more complex, especially if you have never installed one before.

For information on installing and uploading the Zoom CGI, refer to chapter 5.4 of the Users Guide, Additional notes for uploading CGI.

More information on using the different search scripts available can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).

Go back to the Help overview

What is JavaScript?

JavaScript is a scripting language that allows a web page to tell the browser what to do. It is a client-side language, meaning that it is interpreted and processed on the computer viewing the web site - not on the server. This means that it is usually very limited in what it can do, but can be convenient since it does not rely on any special requirements on the server-side, and can also run off a CD with no web server present.

Please note that JavaScript is not the same as Java, and they bear only marginal similarities in syntax. Java is a powerful, general-purpose programming language and can run outside of a web browser but have other disadvantages with speed and system requirements.

Zoom provides a JavaScript version of the search script to cater for users who wish to run their search engines on a CD-ROM, or who have a very restricted hosting environment with no server-side scripting capability. However, due to the technical constraints of JavaScript on most browsers, we are unable to provide the full number of features (such as context descriptions) that are available with the other platform options. It is also unsuitable for large volumes of data, and sites of over 1000 pages or with a large number of unique words, will not run on some browsers.

There is an alternative to running Zoom on CD/DVDs, and that is to include a specialised mini-web server on your CD, so that you can run the CGI or PHP version off the CD. More information is available on our website (under the support page).

More information on using the different search scripts available can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).

Go back to the Help overview

Search statistics

The statistics tool analyses the search logs generated with the <u>Search Logging</u> feature and produces statistic reports featuring pie charts, and graphs, offering insight into the searching patterns of your site's visitors. These reports are generated as HTML web pages and GIF images, allowing you to put them online and make your search engine statistics available to your visitors.

First, you *must* have enabled the <u>Search Logging</u> option (not available in the JavaScript version) to generate the search log files you will need. Once this feature is enabled and running successfully on your website, Zoom will log all searches made to the log file you specified.

You can then download this log file, and specify the location of this file in the search statistics window. You can also simply enter the online URL by which this log file is accessible, and it will download the file for you.

Once you have opened a valid Zoom search log file, you can select one of the many following reports choices to generate. Each will be a new section in the report.

• Top 10 Search phrases

This provides a pie chart breakdown of the top 10 search phrases made on your website. This gives you a good indication of what people are looking for on your website.

• Top 10 "No result" phrases

This provides a pie chart breakdown of the top 10 search phrases on your website which found "no results". This is useful for determining what your visitors are searching for on your website but not managing to find. You can use this information to provide content that better cater your website to your visitors (or determine what meta keywords to add to allow your users to find what they are looking for).

Searches per day (over a 1 to 31 day period)
 This provides a bar or line graph representing the number of searches that are made on your website per day.

Searches per week (over a 1 to 12 day period) This provides a bar or line graph representing the number of searches made on your website per week.

- Searches per month (over a 1 to 6 month period)
 This provides a bar or line graph representing the number of searches made on your website per month.
- List the top x searched words (sorted by popularity) This option appends a HTML table of search words (sorted by popularity). You can specify the number of searched words you would like to list here from 10, to 50,000.

Note: Your report options will only be remembered after you have successfully generated a report. If it failed whilst trying to generate the report, and you click "Close" on the "Search Statistics" window, you will lose your report options. Once Zoom has successfully generated the report, you should save it along with your indexing configuration by clicking on the "File" menu, and selecting "Save configuration".

Web Report

The statistics report generated is a HTML web page with GIF images. These files are located in the "statistics" folder where Zoom is installed (unless configured otherwise, see "Options" below). You can simply upload the contents of this folder to your web server to publish your website's search statistics.

Options

By clicking on the Options button in the main Statistics window, you can specify a different folder to save the reports to.

You can also check the "Append date and time to filenames" option, which will add a number based on the date and time that the report was generated, to the filename of all report files. This ensures that new reports generated do not overwrite existing ones in the same directory.

What to do after indexing

How do I customize the look of the search page?

How do I put the search engine on my website? (See below)

How do I put the search engine on my CD or DVD? (See below)

Required Files

Once you have indexed your website, you will need to upload or copy the files created by the Indexer to your web-server or CD-ROM/DVD distribution.

The list of files required are given in the window that pops up after indexing. These files are created in the Output directory you specified. Note that they are all the files you will need on your website for your search engine to run.

Upload to your website

You will need to upload the listed Required Files to your web-server. Zoom will now prompt you to upload the files for you. You can also upload the files required using your preferred FTP client. Note that you must upload the files in *binary mode*.

Also, make sure that all the files listed are uploaded to the same directory, and to remove or overwrite any files on your web-server from previous indexing sessions.

If you are using the CGI version, you may have to upload to a cgi-bin directory (on Apache servers, check with your web host). You will also have to specify execute permission on the "search.cgi" file and public read permissions on all .zdat files and the search_template.html file.

CD-ROM distribution

If you are putting your search engine on a CD-ROM or DVD (JavaScript platform only), you should copy the Required Files listed to where your CD-ROM content is.

Note that, if you are using relative paths for your base URL, it will be important where you copy these files to make sure that the search result links will still work. For instance, if you specified a base URL of "./" then your search files should be placed on the same directory level as where your CD-ROM content is.

More information on customizing and optimizing your search page can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).

How do I customize the look of my search page?

You can completely alter the appearance of the search page, from changing the fonts and layout to including headers, footers or site navigation menus.

First, you will need to have indexed your site successfully. If so, a default search template file should have been created in the output directory (alongside the index files). This is a HTML file that can be edited like any other ordinary web page. You can open this file for editing from the Indexer, by clicking on the "Templates" menu, and selecting "Customize search page appearance".

You can modify this page as you see fit, however, you must retain the following line in place of where you want the search function to be displayed:

<!--ZOOMSEARCH-->

This is different for the JavaScript version, which has the following line instead:

<script language="JavaScript">ZoomSearch();</script>

If you wish to go back to the default search template, you can simply delete or rename your search template file, and re-index your site. A new "default template" will be created in the output directory at the end of indexing, when it finds that the search template is missing.

To change the fonts, colours, and appearance of your search results (everything after the search form), edit the CSS styles defined at the top of the template HTML file. These styles are defined as below.

Style class	Description
.highlight	The appearance of the highlighted search words in the results, when highlighting is enabled
.searchheading	The appearance of the "Search results for" heading
.summary	The appearance of the summary information for your search results ("129 results found on 2 pages")
.results	The appearance of the search result listing
.category	The appearance of the category tag next to the result link, e.g. "[News article]"
.description	The appearance of the meta description or page description for a search result.
.context	The appearance of the contextual description
.infoline	The appearance of the small information line for each search result ("Terms matched", "Score", and "URL")
.result_title	The appearance of the title line (result number, title, link to result, and category). Note that the link itself will be based on the CSS defined for hypertext links.

You can also modify the options as to which item shows up in the search results (see <u>Configuring Results</u> <u>Layout</u>).

You can also define your own search form or let the script generate the form (see Configuring Search

<u>Page</u>).

More information on customizing and optimizing your search page can be found in the User's Guide (available in PDF format from our website <u>http://www.wrensoft.com/zoom</u>).

Contacting WrenSoft

On the Web

Visit WrenSoft on the web, for more information on Zoom, the User's Guide PDF and purchasing information:

http://www.wrensoft.com

Support & Frequently Asked Questions (FAQ)

A comprehensive list of common support questions (and answers) is available at:

http://www.wrensoft.com/zoom/support.html

E-Mail

For technical support questions, suggestions, or comments, email us at:

zoom@wrensoft.com

Purchasing information

If you wish to use Zoom for a larger website, or you are using it in a commercial project, require professional support, custom features, or simply like the software and wish to support its development, please consider purchasing the licenses available.

Zoom Standard supports larger websites, indexing up to 100 pages of content, while *Zoom Professional* supports unlimited number of pages (restricted only by the amount of memory on your indexing computer). Both versions also provide plug-in support enabling searching and indexing of Word DOC and Acrobat PDF files. And as a licensed user, we will also provide you with free technical support via email.

Please visit http://www.wrensoft.com/sales.html for more information

System Requirements

- Pentium or better PC with Windows 95, 98, ME, NT4, 2000, or XP.
- For Windows 95 and Windows NT, Internet Explorer 5.0 or higher must be installed.
- 64 megabytes of RAM, and 1 megabyte of free disk space.
- For very large web sites containing several 1000's of pages, more RAM is required
- For spider mode, you will require an established connection to the Internet.

Warning with editing HTML with MS Word

There are known issues with using MS Word to edit web pages (HTML files) and as such we recommend against modifying the search template file with the use of MS Word. It can often render the search template useless by stripping out (or misplacing) the HTML comments required by Zoom.

In addition to this, MS Word also appends a large amount of internal meta information, which confuses browsers and consequently affect the way they display the page. One bug even causes any HTML page saved with MS Word to not display at all in Internet Explorer.

For these reasons and more, we strongly suggest you use a different application for editing HTML and in particular, the search template for Zoom. Most dedicated web editing applications are fine – eg. FrontPage, Dreamweaver, and NetObjects Fusion. But even if you have no other application for editing HTML files, we would suggest you use Notepad.

If you do actually have another webpage editor installed, but it is not set as the default editor for HTML files, then you can fix this permanently by changing the File Associations in Windows. You can do this from the Control Panel, under "Folder Options", and selecting the "File types" tab. Select the extension for "HTML" and click on the "Advanced" button. Now select the "Edit" action from the list and click on the "Edit..." button. Here you can specify the application to open this file with. Refer to Windows Help for more information.

Unfortunately due to the way MS Office works, we can not distinguish between whether FrontPage or Word is the currently selected default editor for HTML. So, you may be getting this warning message despite the fact that you are using FrontPage (which is perfectly acceptable and supported). In this case, just click on the "Do not check for Word again" option and continue.

Wizard step – Output folder and limit

Zoom will create a set of index files, which will be required for your search function. Specify a folder here for Zoom to create these files in. This can be a temporary folder of your convenience, or a folder under your local copy of your website.

For Professional users, you can also specify a limit for the maximum number of pages to index. More limits can be changed in the Configuration window, under the "Limits" tab.

Start indexing immediately

This is the default option and will attempt to index your site as soon as you click Finish for the wizard.

Do not start indexing

Select this option if you have more configuration changes or settings that you would like to make, and would like to start indexing at your own convenience.

Wizard step – File types

Here you can select the file types you wish to include in your search functions. Note that you can specify exact file extensions and more options in the Configuration window under the "Scan options" tab.

Also, some file formats are only accessible with plugins for the Standard and Professional editions. For more information on Plugins and these file extensions, please refer to our <u>Plugins</u> page.

Zoom Search Engine Wizard

This wizard is to provide new users with a step-by-step guide to setting up their very own search function for their website or CD/DVD-ROM.

Follow the steps carefully, and select the option required. You can click on the Help button for more information at any point.

If you do not wish to use the Wizard, click "Cancel". You can always start up the Wizard at a later stage by clicking on the "magic wand" icon in the toolbar, or by clicking on the "Tools" menu and selecting "Run Wizard".

You can also check the "Do not show Wizard on startup" option if you do not want to see this window when you next start Zoom.