

## NAME

aspseek.conf – index configuration file

## SYNOPSIS

**/usr/local/aspseek/etc/aspseek.conf**

## DESCRIPTION

**aspseek.conf** is a configuration file for **index(1)**. It completely defines all the aspects of ASPseek indexing process - what to index and how to do it.

The following can be defined:

### General

**DBAddr DBType**:*[[User[:Pass]@]Host[:Port]]/DBName/*

Defines SQL server connection parameters.

**DBType** is SQL server type, it can be *mysql* or *oracle8* for now.

**User** is a SQL server's user to connect as.

**Pass** is a **User**'s password. If this field is omitted, no password is used.

**Host** is a host name or IP address of host to connect to. If you are running SQL server on the same machine, use *localhost*.

**Port** is a port number on which SQL server is listening at. Default is the same as default port of used SQL server.

**DBName** is a name of the database used.

**DBLibDir** */some/dir*

Adds */some/dir* to list of directories to search for database backend library (*libdbname-version.so*). Default library search path is */usr/local/aspseek/lib*. Several such options can be used, each adding one more directory to the list. Last added directory is used first; compiled in path is last.

**DataDir** */some/dir*

Sets directory in which delta files and files with information about words, subsets, spaces will be stored. Default is */usr/local/aspseek/var*.

**DebugLevel** *none | error | warning | info | debug*

Sets the level of debugging. If set to *none*, nothing will be logged. If set to *debug*, you will get a bunch of messages. Default value is *info*.

**Include file**

Includes the contents of *file* at this point, so you can specify some parameters in that included *file*. File name is relative to ASPseek etc directory (*/usr/local/aspseek/etc*).

### Parameters that affects memory usage vs. performance

These parameters can be tuned to achieve the better performance on boxes with enough memory. They can also be used to reduce the amount of memory used by **index(1)**.

**DeltaBufferSize** *kilobytes*

Size of buffer for each of 100 delta files, in kilobytes. Setting of low value for this parameter can result in big fragmentation of delta files. Value of this parameter affects used memory. If default value is used, then 50 Mb of memory is used for buffers. Default value is *512*.

**UrlBufferSize** *kilobytes*

Size of read and write buffer allocated during inverted index merging for **ind** files, in kilobytes. Value of this parameter affects used memory during inverted index merging. Default value is **DeltaBufferSize** \* 8.

**WordCacheSize** *number*

Maximum word count in the word cache. Word cache is used to reduce database load for converting word to its word ID. Default value is *50000*.

**HrefCacheSize** *number*

Maximum URL count in the href cache. Href cache is used to reduce database load for converting URL of outgoing hyperlink to its URL ID. Default value is *10000*.

**NextDocLimit** *number*

Maximum number of URLs loaded from database at each request. Default value is *1000*.

This option is used only if URLs to be indexed are ordered by next index time; otherwise, if **-o** option to **index(1)** is used, all URLs for current hop is taken at once.

### Database format parameters

These parameters set different modes of storing indexed information. Note that database format is different if you change these options, so the same value **must** be set in **searchd.conf(5)** file, and you **must not** change the values on a non-empty database.

#### **HiByteFirst** *yes | no*

Sets the byte ordering used in field **wordurl[1].word** (only in Unicode version). Default is *no*.

#### **IncrementalCitations** *yes | no*

Sets whether to build citation index, ranks of pages and lastmod incrementally. If value of this parameter is set to *yes*, then calculating of citations, ranks of pages and lastmod file will require less memory and take less time on large databases. So it is very handy if you want to index large number of URLs and have relatively small amount of memory. Default is *yes*.

#### **CompactStorage** *yes | no*

Sets the storage mode of reverse index. In compact storage mode, file/BLOB is not created for each word. Instead, information about all words is stored in 300 files. In this mode, updating of reverse index is generally much faster and requires a bit less memory than in the old mode. Default is *yes*.

#### **UtfStorage** *yes | no*

This parameter has sense only in Unicode version and only for MySQL back-end. In UTF8 storage mode fields **wordurl[1].word** are stored in UTF8 encoding. This mode can reduce sizes of data and index files for **wordurl** table. To convert existing Unicode database to this mode, run **index -b**. Default value is *no*.

### Bandwidth control

#### **MaxBandwidth** *bytes [starttime [endtime]]*

Sets maximum used bandwidth for incoming traffic to *bytes* per second for the specified period of time of day. Arguments *starttime* and *endtime* are in seconds from midnight (0:00). If *endtime* is omitted, then it is implied to be the end of the day (86400). If both *starttime* and *endtime* are omitted, then the limit is for the whole day. You can use several **MaxBandwidth** commands. Note that if *endtime* is less than *starttime*, **index(1)** will handle it correctly, setting two intervals from *starttime* to midnight and from midnight to *endtime*. By default bandwidth is not limited.

### Indexing

#### **Server** *URL*

Add *URL* as an URL to start indexing from. You can specify many **Server** commands, and set the different options for different sites - see below. Note that if *URL* contains path, the whole site will be indexed nevertheless, so to limit indexing to some subdirectory of site use **Disallow** parameter described below.

### Global indexing parameters

Each of the below parameters can be specified only once in configuration file and takes the global effect for the whole **index(1)** session.

#### **MaxDocSize** *bytes*

Sets the maximum document size in bytes, so if document size is bigger than *bytes*, only the first *bytes* of the document will be processed. Default value is *1048576* bytes (1Mb).

#### **HTTPHeader** *header*

Add *header* to headers that **index(1)** sends in HTTP request. You should not use *If-Modified-Since* or *Accept-Charset* headers here, as **index(1)** sends it anyway. Header *User-Agent: aspseek/1.2.10* is sent too, but you may override it.

#### **Clones** *yes | no*

Sets whether to enable clones eliminating. Clone is a document which is absolutely the same as another document. If this set to *yes*, clone is not parsed/stored in the database, instead word information for original document is used. Default value is *yes*.

**MinWordLength** *number*

Sets the minimum length of word to be stored in the database, so words shorter than *number* is not stored. Default value is *1*.

**MaxWordLength** *number*

Sets the maximum length of word to be stored in the database, so words longer than *number* is not stored. Default value is *32*. Note that you can't set the value higher than *32*.

**DeleteNoServer** *yes | no*

Sets whether to delete URLs which have no correspondent "Server" commands. Default value is *yes*.

**AddressExpiry** *time*

Sets expiration time for "DNS name -> IP" entry in address cache. After entry is expired, resolver will make DNS lookup again. Argument *time* can be set in seconds, or the same way as in **Period** command below. Default value is 1 hour.

**Indexing scope**

These parameters can be used to limit the scope of indexing. **index(1)** will compare all URLs against all **CheckOnly**, **CheckOnlyNoMatch**, **Allow**, **AllowNoMatch**, **Disallow** and **DisallowNoMatch** directives in the order specified in configuration file, so order is important. Note that by default everything is allowed.

Some directives below use POSIX regular expressions (regex) for flexibility. For description of what regex is, see **regex(7)**, **grep(1)**, **awk(1)**.

**FollowOutside** *yes | no*

Sets whether **index(1)** should index outside sites defined in **Server** directives. Default is *no*. If you set it to *yes*, be sure to limit the scope of indexing in some other way (for example, with **MaxHops**).

**CheckOnly** *regex [regex...]*

Use **HEAD** request instead of **GET** for URLs matching *regex*. So, such URLs will not be downloaded, just information about them will be stored in **urlword** table.

**CheckOnlyNoMatch** *regex [regex...]*

Use **HEAD** request instead of **GET** for URLs **not** matching *regex*. So, such URLs will not be downloaded, just information about them will be stored in **urlword** table.

**Allow** *regex [regex...]*

Allows to index URLs matching *regex*.

**AllowNoMatch** *regex [regex...]*

Allows to index URLs **not** matching *regex*.

**Disallow** *regex [regex...]***DisallowNoMatch** *regex [regex...]*

Disallows to index URLs **not** matching *regex*.

**Countries** *file*

Loads countries IP information from *file*. File consists of lines in the form "sss.sss.sss.sss - eee.eee.eee.eee cc", where *sss.sss.sss.sss* is starting IP address, *eee.eee.eee.eee* is ending IP address, and *cc* is a country code (like *ru*, *de*, etc.). Note that value of ending address should be more than starting address.

**AllowCountries** *cc1 [cc2...]*

Specifies to index only sites from countries specified by *cc1*, *cc2*, etc. Should be used together with the **Countries**.

**Indexing parameters - local**

Each of the below parameters can be specified many times in configuration file, applies to all **Server** parameters below it, and valid till next parameter with the same name, or till the end of configuration file.

**Period** *time*

Sets the re-index period to *time*. Value can be set just in seconds, or using a special characters right after the number (no spaces allowed): **s** for seconds, **M** for minutes, **h** for hours, **d** for

days, *m* for months and *y* for years. You can combine several values together, for example string *1m12d* means "one month and twelve days". You can also specify negative numbers, say *1m-10d* stands for "one month minus ten days". Default value is *7d*.

**Tag** *number*

Use this field to "tag" several **Servers** with value *number*, which can later be used with option **-t** *number* of **index(1)** command. Note that if you want to group several sites together for searching purposes, you should use "spaces" or "subsets" features of ASPseek, not tag.

**MaxHops** *number*

Sets the maximum hops ("mouse clicks") from URL specified by **Server** command, so documents that are "deeper" will not be indexed. Default value is *256*.

**IncrementHopsOnRedirect** *yes | no*

Sets whether **index(1)** should increment hops value when HTTP redirect is encountered. Applies only to redirects generated by "**Location:**" HTTP headers. Setting this option to *no* allows a greater number of documents to be indexed for sites that redirect frequently (e.g. for cookie testing, typically on each page). Default value is *yes*.

**RedirectLoopLimit** *number*

Allow no more than *number* of contiguous redirects. This option is especially useful if you set **IncrementHopsOnRedirect** to *no*, because **index(1)** can fall in an endless redirect loop. Limiting the number of redirects prevents **index** from such redirect loops. Default value is *8*.

**MaxDocsPerServer** *number*

Sets that no more than *number* of documents will be indexed from one site during one run of **index(1)**. Default value is *-1*, which means no limits.

**MaxDocsAtOnce** *number*

Sets the maximum number of pages to be downloaded from the same host before switching to the next host. Large values are believed to increase indexing performance when number of indexed sites is large. Default value is *1*.

**ReadTimeOut** *time*

Sets the maximum timeout to *time* for downloading a document from site. Argument can be expressed in seconds, or in the same form as in **Period** command above. Default value is 90 seconds.

**Robots** *yes | no*

Sets whether the robot exclusion standard (`robots.txt` file and `META NAME="robots"`) will be honored. Default is *yes*.

**DeleteBad** *yes | no*

Sets whether to delete bad (not found, forbidden etc.) URLs from the database. Default value is *no*.

**Index** *yes | no*

Sets whether to store words into database. Default value is *yes*.

**Follow** *yes | no*

Sets whether to store links found into database. Default value is *yes*.

**Charset** *charset*

Usable to set charset for the servers that do not return it. Argument should be known charset name (see below for charset configuration). Alternatively, you can use charset guesser feature of **index(1)**.

**Replace** [*regexp* [*replacement*]]

This parameter allows to replace URL matching *regexp* by *replacement*, or by empty string if *replacement* is not specified. This is useful for sites with dynamic contents where the same information can be obtained by many different URLs. **Replace** without arguments disables any replacements for subsequent **Server** commands.

As in **sed(1)** command *s*, the *replacement* can contain  $\backslash N$  (*N* being a number from 1 to 9, inclusive) references, which refer to the portion of the match which is contained between *N*th  $\backslash($  and its matching  $\backslash)$ . To include a literal  $\backslash$ , precede it with another  $\backslash$ .

**MinDelay** *time*

Sets minimum time between finishing of access to server and beginning of next access to the server. This is useful if site owner blames you for "bombing" his site with your **index(1)** queries. Argument *time* can be set in seconds, or in the same way as described in **Period** command above. Default value is 0.

**Proxy** [*host.com[:port]*]

Use proxy rather than direct connection. You can also index FTP sites via proxy. If *port* is not specified, default is 3128 (sqiud). **Proxy** without arguments disables proxy.

**External converters**

**index(1)** has an ability to deal with document types other than **text/plain** and **text/html**. It does so with the help of an external programs or scripts, which can convert from some format to **text/plain** (or **text/html**), so you are able to index .ps, .pdf etc.

**Converter** *from/type to/type*[:**charset=cset**] *command line*

Specifies that for converting documents with MIME-type *from/type* to MIME-type *to/type* the command specified by *command line* will be used. Argument *from/type* can be any type returned by Web server. Argument *to/type* can be either *text/plain* or *text/html*. If you add **;charset=cset** string after *to/type*, **index** will know that resulting document has a charset *cset*, otherwise it is assumed to be **us-ascii**.

In the *command line* you usually specify program or script to run, together with its options. Program is expected to read from stdin and write the converted document to stdout.

If your program can't deal with stdin/stdout streams, you should use **\$in** and **\$out** strings in **command line**, and they will be substituted with two file names in /tmp directory. **index(1)** will create files with unique names, write the document downloaded to the first file (referenced as **\$in**), run the */bin/prog*, read the second file (referenced as **\$out**) into memory, and then delete both files.

You can also use **\$url** in *command line*, it will be substituted with the actual URL of downloaded document. You can use it in your own scripts to distinguish between a different document variations, or to be able to write one script for many different MIME-types.

Please note that **index(1)** relies on a **Content-Type** header returned by a Web server. Some Web-servers are misconfigured and give wrong info (for example, return header **Content-Type: audio/x-pn-realaudio-plugin** for .rpm files).

Examples:

```
Converter app/ps text/plain; charset=iso8859-1 ps2ascii
# ps2ascii can't deal with PDF files from stdin
Converter application/pdf text/plain ps2ascii $in $out
```

**Charset configuration for non-Unicode version**

Charset configuration for non-Unicode version is usually stored in file /usr/local/aspseek/etc/charsets.conf. Charset files for non-Unicode version can be found in /usr/local/aspseek/etc/charsets directory. Langmap files can be found in /usr/local/aspseek/etc/langmap directory.

**CharsetTable** *charset lang file* [*lmfile*]

Loads the table for *charset* of language *lang* from *file*. Optionally load langmap file *lmfile*, which is used for charset guesser.

**CharsetAlias** *charset alias1* [*alias2...*]

Defines *alias1*, *alias2*, ... as aliases (alternative names) for *charset*. This is needed because in many cases there is no "one true name" for the charset - different web servers and page authors use different names.

**LocalCharset** *charset*

Sets the local charset for ASPseek, so all data in the database is assumed to be in that charset.

**Charset configuration for Unicode version**

Charset configuration for Unicode version is usually stored in file `/usr/local/aspseek/etc/ucharset.conf`. Charset files for Unicode version can be found in `/usr/local/aspseek/etc/tables` directory.

**CharsetTableU1** *charset lang file [lmfile]*

Loads the Unicode mapping for *charset* of language *lang* from *file*. Optionally load langmap file *lmfile*, which is used for charset guesser.

**CharsetTableU2** *charset lang file [lmfile]*

Loads the Unicode mapping for multibyte *charset* of language *lang* from *file*. Optionally load langmap file *lmfile*, which is used for charset guesser.

**Dictionary2** *lang file [charset]*

Loads dictionary for *lang* from *file*. If *charset* is not specified, it is assumed that the file is in Unicode. Dictionary is used for tokenizing of text in Chinese, Japanese and Korean languages.

**Stopwords**

Stopwords configuration is usually stored in the file `/usr/local/aspseek/etc/stopwords.conf`. Stopword files for different languages can be found in `/usr/local/aspseek/etc/stopwords` directory.

**StopwordFile** *lang file [charset]*

Loads stopwords for language *lang* from *file*. If *charset* is not specified, file contents is assumed to be in **LocalCharset**, otherwise it is in *charset*.

**FILES**

```
/usr/local/aspseek/etc/aspseek.conf  
/usr/local/aspseek/etc/charsets.conf  
/usr/local/aspseek/etc/ucharset.conf  
/usr/local/aspseek/etc/stopwords.conf
```

**BUGS**

Many parameters are the same in `searchd.conf` and in `aspseek.conf(5)`.

**SEE ALSO**

`index(1)`, `aspseek.conf(5)`, `regex(7)`, <http://www.robotstxt.org/wc/robots.html>.

**AUTHORS**

Copyright (C) 2000, 2001, 2002 by SWsoft.  
Man page by Kir Kolyshkin <kir@asplinux.ru>