

Sphider-plus Manual

Content

1. Introduction.....	3
2. Version and legal info.....	3
3. Installation of Sphider-plus version 2.0.....	4
4. Installation of Sphider-plus version 1.0 - 1.9.....	6
5. Settings and customizing.....	8
6. Indexing options.....	10
7. Using the indexer from command line.....	11
8. Keeping pages, words and files from being indexed.....	12
8.1 robots.txt.....	12
8.2 Must include / must not include string list.....	12
8.3 Ignoring links.....	12
8.4 Canonical <link> tag.....	12
8.5 Ignoring parts of a page.....	13
8.6 Ignored words.....	13
8.7 Use of Whitelist.....	13
8.8 Use of Blacklist	14
8.9 Ignored files.....	14
9. UTF-8 Support and 'Preferred Charset'.....	15
10. Search modes.....	17
10.1 Search with wildcards *	17
10.2 Strict search !.....	17
10.3 Tolerant search.....	17
10.4 Link search site:.....	18
10.5 Media search.....	18
11. Chronological order for result listing.....	18
12. PDF converter for Linux/UNIX systems.....	20
13. Clean resources during index / re-index.....	20
14. Enable real-time output of logging data	21
15. Error messages and Debug mode.....	21
16. Delete secondary characters.....	22
17. Media search for images, audio streams and videos.....	23
17.1 Media indexing.....	23
17.2 Not supported media content	24
17.3 Search for media content.....	24
17.4 Statistics for media content.....	25
18. RSS and Atom feeds.....	26
19. Result cache for text and media queries.....	26

20. Multiple database support.....	27
20.1 Overview.....	27
20.2 Definition and configuration.....	27
20.3 Activate / Disable databases.....	28
20.4 Backup & Restore of databases.....	28
20.5 Copy & Move.....	29
20.6 Enhancing functionality of multiple database support.....	29
21. FAQs.....	31
21.1 UTF-8 support does not work.....	31
21.2 Can't search for long words.....	31
21.3 Can't search for words with non-Latin characters.....	31
21.4 How to bypass the Admin log in.....	31
21.5 Links are not followed during Re-index, only main URL is indexed (option 1).....	32
21.6 Links are not followed during Re-index, only main URL is indexed (option 2).....	32
21.7 How to integrate Sphider's search field into existing pages.....	32
21.8 Error message: "Warning: set_time_limit() . . . ".....	33
21.9 Error message: "Unable to flush table 'addurl' ".....	33
21.10 Error message: " Access denied; you need the RELOAD privilege. . . ".....	33
21.11 Fatal error: "Allowed memory size of xxx bytes exhausted (tried to allocate yyy bytes)".....	33
21.12 PDF documents are not indexed.....	34
21.13 PHP security info is not presented in Admin Statistics.....	34
21.14 What kind of input validation is performed?.....	34
21.15 How to protect Database management against Admin access?.....	35
22. Change log.....	36
22.1 Version 1.0 - 1.9.....	36
22.2 Version 2.0.....	52

Last update: Wednesday, May 27, 2009

1. Introduction

Sphider-plus is a search engine based on the original Sphider scripts created by Ando Saabas (www.sphider.eu).

In front of original Sphider additional modules, functions, template designs and debugging have been performed. For details about all changes, please notice the chapter [Change Log](#)

Sphider-plus requires PHP 5 with an installed GD library and MySQL database .

The names of Sphider-plus folders and scripts are often the same like those of original Sphider. But the scripts are not interchangeable between Sphider and Sphider-plus.

Some messages have been added in the language files. You are invited to translate your native language and then to share the files with the community. Also mods, improvements and of course bug fixes are very welcome for future releases of Sphider-plus.

2. Version and legal info

Name: Sphider-plus
Version: 2.0
Created: May 27, 2009

Based on original Sphider version 1.3.4 released 29.04.2008 by Ando Saabas <http://www.sphider.eu>

This program is licensed under the GNU GPL v.3 by Rolf Kellner [Tec] [tec\(a\)t\)sphider-plus.eu](mailto:tec(a)t)sphider-plus.eu)

Original Sphider GNU GPL licence by Ando Saabas [ando\(a\)t\)cs.ioc.ee](mailto:ando(a)t)cs.ioc.ee)

Updates and support for Sphider-plus are available at:

<http://www.sphider-plus.eu>
[tec\(a\)t\)sphider-plus.eu](mailto:tec(a)t)sphider-plus.eu)

If you like Sphider-plus and want to promote further development, your donation at PAYPAL account

tec@sphider-plus.eu

is highly appreciated. Thank you very much.

3. Installation of Sphider-plus version 2.0

Because of the multiple database support this release requires a fresh installation of all scripts and a blank MySQL database created with UTF8_bin collation. An update from former Sphider-plus versions or an upgrade from original Sphider is not foreseen.

In order to get Sphider-plus running, perform the following steps:

1. Unzip the downloaded file, and copy all folders and files to the server, for example to:

`C:\programs\xampp\htdocs\public\sphider-plus\`

Even blank folders will be required later on during index and search procedures. So, also all blank folders and files need to be transferred to the server.

2. Create at minimum one database in MySQL to hold Sphider-plus data tables. Collation of the database must be UTF8_bin

3. Open the file `.../admin/auth.php` and personalize the two variables: 'Username' and 'Password'

As per default download they are set to:

```
$admin = "admin";  
$admin_pw = "admin";
```

These two variables are used as login authorization for the Admin interface.

4. Open the file `.../admin/auth_db.php` and personalize the two variables: 'Username' and 'Password'

As per default download they are set to:

```
$db_admin = "admin";  
$db_admin_pw = "admin";
```

These two variables are used as login authorization for the Database management interface. This is a submenu of the Admin interface.

5. Open the Admin interface with your browser by addressing the Admin with something like:

`http://localhost/public/sphider/admin/admin.php`

After login with Admin 'Username' and 'Password', the Admin interface will be presented. After first login, there will be several warning messages, because no database is allocated to Sphider-plus.

6. Open the Submenu 'Database' and select 'Configure'. Now you will need to login for the Database management with the authorization as defined in step 4 of this installation instruction.

7. Entering the first time into this section, there will be several warning messages. At minimum one database has to be defined by:

- Name of database
- Username
- Password
- Database host
- Prefix for Tables

Pressing the 'Save' button will assign Sphider-plus to these database definitions. Never the less, the warning message **'Tables are not installed for database x'** will remain in the Database settings overview.

The '**Install all tables for database x**' is an independent procedure, which has to be invoked by the Admin after the database has been allocated.

If the database is allocated and the tables are installed, the message '**Database x settings are okay.**' are displayed in the settings overview; showing separately the situation for each of the five databases.

If the application should work with only one database, the settings for the non-required databases may remain blank. A corresponding message will be displayed:

Mysql server for database 2 is not available!

Trying to reconnect to database 2 . . .

Cannot connect to this database.

Never mind if you don't need it.

Installation of multiple databases is described in chapter [Multiple database support](#)

8. Next step to get Sphider-plus to work will be the activation of the database. There are three settings available in the 'Activate / Disable' section:

- Select active database for Admin
- Select active database for 'Search' user
- Select active database for 'Suggest URL' user

Each setting allows activating of one database. So, if multiple databases are configured, an independent use of databases is enabled for 'Admin', 'Search' User and 'Suggest URL' user.

After activating the databases for the different tasks, database configuration is finished. The currently activated database is displayed for the Admin in 'Sites' table like:

Database 1 - Displaying URLs 1 - 10 from 25

9. Turn back to the standard Admin interface by selecting 'Sites' as one of the available menu selections. Again a warning is presented, because up to now no URL was entered, which could be indexed. Site URL could be entered by selecting one of the three possibilities:

- Add site
- Import URL list
- Index

10. Sphider-plus is ready to index the first site now, using the default settings as delivered by download. In order to individualize the settings, the submenu 'Settings' will offer more than 100 items to be defined.

4. Installation of Sphider-plus version 1.0 - 1.9

There are three ways to install Sphider-plus:

- Install from scratch
- Upgrade from original Sphider to Sphider-plus
- Update to new Sphider-plus releases

In order to get correct function of Sphider-plus, please follow the instructions as described below.

4.1.1 Install from scratch

1. Unzip the files, and copy them to the server, for example to:

C:\programs\xampp\htdocs\public\sphider-plus\

2. Create a database in MySQL to hold Sphider-plus data. Collation of the database must be UTF8_bin

a) at command prompt type (to log into MySQL):

mysql -u <your username> -p

Enter your password when prompted.

b) in MySQL, type:

CREATE DATABASE sphider-plus_db;

Of course you can use some other name for database instead of sphider-plus_db.

c) Use exit to exit MySQL.

For more information on how to create a database and give/get the necessary permissions, check www.mysql.com

3. In ../sphider-plus/settings/ directory, edit the file database.php and change:

\$database

\$mysql_user

\$mysql_password

\$mysql_host

to your personal requirements to correct values (if you don't know what \$mysql_host should be, it should probably stay as it is - 'localhost').

4. Open the file ../admin/install_all.php in your browser. This script will create the tables necessary for Sphider-plus to operate.

(http://localhost/public/sphider-plus/admin/install_all.php)

5. In admin directory, edit the file auth.php to change the

- administrator user name

and

- password

(Default values are 'admin' and 'admin').

6. Open admin.php in your browser and start indexing

(<http://localhost/public/sphider-plus/admin/admin.php>)

Installing from scratch your site table is empty. You will be asked to add any URL to be indexed.
You should enter something like:
`http://www.abc.de`

7. `search.php` is the default search script
(`http://localhost/public/sphider-plus/search.php`)

4.1.2 Upgrade from original Sphider to Sphider-plus

1. Create a backup of your current Sphider database and all original Sphider files.
2. Unpack the downloaded files, and copy them into a new folder for your Sphider-plus installation, for example to:
`C:\programs\xampp\htdocs\public\sphider-plus\`
3. In `.../sphider-plus/settings/` directory, edit the file `database.php` and change:
`$database`
`$mysql_user`
`$mysql_password`
`$mysql_host`
to those values you used up to now for the database together with original Sphider.
4. In `admin` directory, edit the file `auth.php` to change the
- administrator user name
and
- password
(default values are 'admin' and 'admin').
5. In front of original Sphider, Sphider-plus needs some additional tables in your database. They will be created by calling the file `.../admin/install_sphider-plus.php` in your browser:
(`http://localhost/public/sphider-plus/admin/install_sphider-plus.php`)
6. Open `admin.php` in your browser and define the settings for your specific application.
(`http://localhost/public/sphider-plus/admin/admin.php`)
7. `search.php` is the default search script
(`http://localhost/public/sphider-plus/search.php`)

Additional info:

When upgrading from original Sphider, after defining the Admin settings, it is absolutely necessary to run an 'Erase & Re-index' procedure.
Also you should keep in mind that for full Sphider-plus support your database needs to be created with utf-8 collation.

4.1.3 Updating to new Sphider-plus releases

If you already use Sphider-plus and want to update to the newest release, you may copy only the new scripts over the existing you used hitherto. In order to find out, which scripts are involved for the different releases, please notice the chapter "[Change log](#)". Some releases of Spider-plus also require additional database tables or just some new rows/columns for an existing table. The corresponding change log informs about necessary procedures.

But Sphider-plus takes care about its database. If you press a 'Save settings' button in Admin settings and Sphider-plus recognizes a database inconsistency, a message will be displayed. Also you will be asked to run the according installation script. Something like:

Please run the .../admin/install_bestclick.php file.

5. Settings and customizing

If you want to change settings, behavior and design of Sphider-plus, you can do so by means of the Admin interface. There is a wide range of settings foreseen in Sphider-plus. Separated into different submenus like:

- Sites
 - Add Site
 - Index only the new
 - Re-index all
 - Erase & Re-index individual
 - Approve and banned domains manager
- Categories
 - Add, edit, delete
 - Create new subcategory
- Index
 - Basic indexing options
 - Advanced options
- Clean
 - Clean keywords not associated with any link
 - Clean links not associated with any site
 - Clean Category table not associated with any site
 - Clean Media links
 - Clear Temp table
 - Clear Search log
 - Clear 'Most Popular Page Links' log
 - Clear 'Most Popular Media Links' log
 - Clear Spider log, separate and bulk delete
 - Clear Thumbnail images, separate and bulk delete
 - Clear Text cache
 - Clear Media cache

- Settings
 - General definitions
 - Index Log settings
 - Spider options
 - Search settings
 - Cache definition, activity and size
 - Order of Result listing
 - Suggest options
 - Page Indexing Weights
- Database
 - Configure up to 5 databases
 - Activate separate for 'Admin', 'Search' user and 'Suggest URL user'
 - Backup / Restore
 - Copy / Move
 - Optimize
- Templates
 - Three different template designs are prepared for Sphider-plus, which you may be selected in submenu 'Settings'. If the layout does not fit the design of your site (which is normal), you may create new template designs and modify the appropriate file .../templates/My_template/thisstyle.css
- Statistics
 - Top keywords (Top 50 with hit counter)
 - All indexed thumbnails with ID3 and EXIF info
 - Largest pages offering link addr. and file size.
 - Most Popular Searches for text links offering:
 - Link addr., total clicks, last clicked, last query (Top 50)
 - Most Popular Searches for media links offering:
 - Link addr., total clicks, last clicked, last query (Top 50)
 - Most Popular Links (click counter)
 - Search log offering:
 - Query, Results, Queried at, Time taken, User IP, Country, Host name (Latest100)
 - Spidering log offering:
 - File-name, index date and delete option
 - Server info offering:
 - Server software, environment, MySQL, PDF-converter, image functions, php.ini file
 - PHP integration, PHP security info. Each item holding lists of details.

All text links, media links and thumbnails are active linked.

6. Indexing options

As part of the Admin Site settings you may select the following options:

Full: Indexing continues until there are no further (permitted) links to follow.

To depth:

Indexes to a given depth, where depth means how many "clicks" away the page can be from the starting page. Depth 0 means that only the starting page is indexed, depth 1 indexes the starting page and all the pages linked from it etc.

Re-index:

By selecting this mod, indexing is forced even if the page already has been indexed. Re-index only detects changes of the pages to be re-indexed. Modifications in Admin settings are not recognized.

Erase & Re-index:

By selecting this mod, indexing is forced even if the page already has been indexed.

Additionally this mod will

Clear Sphider-plus database before the re-index process. It will leave the following untouched:

- Categories
- Query log
- Sites and all options: spider-depth, last indexed, can leave domain, title, description, Url Must include, Url must Not include.

If settings have been modified in Admin section this mod should be selected to update the database.

Spider can leave domain:

By default, Sphider never leaves a given domain, so that links from domain.com pointing to domain2.com are not followed. By checking this option Sphider can leave the domain, however in this case its highly advisable to define proper must include / must not include string lists to prevent the spider from going too far. This option must be activated, if a htaccess. file is used for redirect directives.

Must include / must not include: See below for an [explanation](#).

7. Using the indexer from command line

(Matter of re-definition and re-coding, currently without guarantee for proper function)

It is possible to spider web pages from the command line, using the syntax:

```
php spider.php <options>
```

where <options> are:

-all	Reindex everything in the database
-u <url>	Set the url to index
-f	Set indexing depth to full (unlimited depth)
-d <num>	Set indexing depth to <num>
-l	Allow spider to leave the initial domain
-r	Set spider to reindex a site
-m <string>	Set the string(s) that an url must include (use \n as a delimiter between multiple strings)
-n <string>	Set the string(s) that an url must not include (use \n as a delimiter between multiple strings)

For example, for spidering and indexing <http://www.domain.com/test.html> to depth 2, use:

```
php spider.php -u http://www.domain.com/test.html -d 2
```

If you want to reindex the same url, use:

```
php spider.php -u http://www.domain.com/test.html -r
```

8. Keeping pages, words and files from being indexed

8.1 *robots.txt*

The most common way to prevent pages from being indexed is using the robots.txt standard, by either putting a robots.txt file into the root directory of the server, or adding the necessary meta tags into the page headers.

8.2 *Must include / must not include string list*

A powerful option Sphider-plus supports is defining a 'Must include / Must not include' string list for a site (to be found in Sites / Options / Edit). Any Url containing a string in the 'Url must Not include' list is ignored. Any Url that does not contain any string in the 'Url Must include' list is likewise ignored.

All strings in the string list should be separated by a newline (Enter). For example, to prevent a forum in your site from being indexed, you might add `www.yoursite.com/forum` to the 'Url must Not include' list. This means that all Urls containing the string will be ignored and wont be indexed. Using Perl style regular expressions instead of literal strings is also supported. Every string starting with a '*' in front is considered as a regular expression, so that `*[a]+'` denotes a string with one or more a in it.

8.3 *Ignoring links*

Sphider-plus respect `rel="nofollow"` attribute in `<a href . . . >` tags, so for example the link `foo.html` in ``

will be ignored.

Also if the nofollow flag is set in the header of a site, this link will not been followed.

8.4 *Canonical <link> tag*

As defined by Google, Microsoft and Yahoo! in February 2009, also Sphider-plus will follow the instruction of a `rel="canonical"` link. You may simply add this `<link>` tag to specify your preferred page version:

```
<link rel="canonical" href="http://www.example.com/product.php?item=swedish-fish" />
```

inside the `<head>` section of all the duplicate content URLs:

<http://www.example.com/product.php?item=swedish-fish&category=gummy-candy>

<http://www.example.com/product.php?item=swedish-fish&trackingid=1234&sessionid=5678>

and Sphider-plus will understand that the duplicates all refer to the canonical URL:

<http://www.example.com/product.php?item=swedish-fish>.

The duplicate pages will be ignored and not indexed. Sphider-plus takes the `rel="canonical"` as a directive, not a hint. The canonical link may also be a relative path, but is not allowed to refer to a different domain. Unfortunately the creation of canonical link tags needs to be done manually. So special care has to be taken that other directives like robots.txt or `rel="nofollow"` will not prevent the crawling of the canonical origin.

8.5 Ignoring parts of a page

Sphider-plus includes an option to exclude parts of pages from being indexed. This can for example be used to prevent search result flooding when certain keywords appear on certain part in most pages (like a header, footer or a menu).

Any part of a page between

`<!--sphider_noindex-->` and `<!--/sphider_noindex-->`

tags is not indexed, however links in it are followed.

8.6 Ignored words

Beginning with version 1.7, Sphider-plus offers the capability to prepare language specific common files. Common words that are not to be indexed can be placed into individual files. The names of this files must start with 'common_' and end with the suffix '.txt', like "common_eng.txt ". The files must be placed into the folder `.../include/common/`.

The common word files should not be used, if 'phrase search' is the standard type of search. Sphider-plus will become problems to find complete phrases. Therefore, in Admin / Settings/ Spider settings, the use of common word files may be activated / deactivated by the checkbox:

Use 'commonlist' for words to be ignored during index / re-index?

Take notice, that the 'Ignored words' function is case sensitive. So, if you intend to use UTF-8 support with distinct results for upper- and lowercase queries, you need to include both words (Sphider & sphider) into the common files.

If 'Convert all to UTF-8' is selected in Admin settings the 'common_xyz.txt' files are also converted. If 'Enable distinct results for upper- and lower-case queries' is not selected in Admin settings, the words placed in common_xyz.txt files are converted to lower case characters, so they will match independent of their spelling in the .txt file. These two items are performed always when the script is started, so that this transformation is valid only until the Admin settings are changed again.

In order to reduce indexing time, unused language files may be deleted.

8.7 Use of Whitelist

Sphider-plus offers the capability to control the index / re-index procedure by a list of words called 'whitelist'. Only if the content of the page holds at minimum one word of the whitelist, it will be indexed / re-indexed. The list is placed in the file `.../include/common/whitelist.txt`

In Admin / Settings/ Spider settings, the use of the whitelist may be activated / deactivated by the checkbox:

Use whitelist in order to enable index / re-index only those pages
that include any of the words in whitelist?

Take notice, that this function is not case sensitive. So, you only need to include one spelling into the whitelist.txt file.

If 'Convert all to UTF-8' is selected in Admin settings the words placed in 'whitelist.txt' file are also converted. This translation is performed always when the script is started, so that this transformation is valid only until the Admin settings are changed again.

8.8 Use of Blacklist

Sphider-plus offers the capability to control the index / re-index procedure by a list of words called 'blacklist'. If the content of the page contains one word of the blacklist, it will not be indexed / re-indexed. The list is placed in the file `.../include/common/blacklist.txt`

In Admin / Settings/ Spider settings, the use of the blacklist may be activated / deactivated by the checkbox:

Use blacklist to prevent index / re-index of pages that contain any of the words in blacklist?

A second setting in the same settings section enables the rejection of queries that contain a word of the blacklist. Even if the evil word is only part of the query. If the checkbox

Use blacklist to delete queries that contain any of the words in blacklist?

is activated, the complete query is deleted and a blank search is performed. That ensures that also the table 'Search log' remains clean.

Take notice, that the 'Use of Blacklist' functions are not case sensitive. So, you only need to include one spelling into the `blacklist.txt` file.

If 'Convert all to UTF-8' is selected in Admin settings the words placed in 'blacklist.txt' file are also converted. This translation is performed always when the script is started, so that this transformation is valid only until the Admin settings are changed again.

8.9 Ignored files

The list of file types that are not checked for indexing are places in `.../include/common/ext.txt`. This file holds all file suffixes for those type of files that are to be ignored during index / re-index procedure.

The 'ext.txt' file is independent of the media files to be indexed. All file types not to be followed for text indexing must be placed in 'ext.txt'. To be seen as a blacklist for file suffixes.

While

image.txt
audio.txt
video.txt

are whitelists that include suffixed for files to be indexed according to the type of media.

9. UTF-8 Support and 'Preferred Charset'

Starting with version 1.2, Sphider-plus provides Unicode assistance. The following 63 charset are supported and will be converted into UTF-8 Unicode:

WINDOWS

windows-1250 - Central Europe
windows-1251 - Cyrillic
windows-1252 - Latin I
windows-1253 - Greek
windows-1254 - Turkish
windows-1255 - Hebrew
windows-1256 - Arabic
windows-1257 - Baltic
windows-1258 - Viet Nam
cp874 - Thai - this file is also for DOS

DOS

cp437 - Latin US
cp737 - Greek
cp775 - BaltRim
cp850 - Latin1
cp852 - Latin2
cp855 - Cyrillic
cp857 - Turkish
cp860 - Portuguese
cp861 - Iceland
cp862 - Hebrew
cp863 - Canada
cp864 - Arabic
cp865 - Nordic
cp866 - Cyrylic Russian (this is the one,
used in IE "Cyrillic (DOS)")
cp869 - Greek2

MAC (Apple)

x-mac-cyrillic
x-mac-greek
x-mac-icelandic
x-mac-ce
x-mac-roman

ISO

iso-8859-1
iso-8859-2
iso-8859-3
iso-8859-4
iso-8859-5
iso-8859-6
iso-8859-7
iso-8859-8
iso-8859-9
iso-8859-10
iso-8859-11
iso-8859-12
iso-8859-13
iso-8859-14
iso-8859-15
iso-8859-16

MISCELLANEOUS

gsm0338 (ETSI GSM 03.38)
cp037
cp424
cp500
cp856
cp875
cp1006
cp1026
koi8-r (Cyrillic)
koi8-u (Cyrillic Ukrainian)
nextstep
us-ascii
us-ascii-quotes

DSP implementation for NeXT

stdenc
symbol
zdingbat

And specially for old Polish programs
mazovia

It is only a small checkbox in Admin settings. But in consequence of your selection, the impact will be dramatically.

First of all: if you select this option, the complete text and all keywords to be indexed, will have to be translated into Unicode. Consequence is an increase of time for indexing.

As also suggested by Yiannes [pikos], I integrated three steps to realize this procedure:

1. Detect charset of site, page or file that's content has to be translated.
This information is normally presented as part of the HTML header.
If not available, or for files without header like .doc, .rtf, .pdf, .xls and .ppt files, the 'Preferred charset' (as defined in Admin settings) will be used to translate the file into Unicode. In other words: you can't convert DOCs, PDFs etc. that are coded in 'foreign' charset. Only those with your personal charset will be converted correctly.
2. By means of the PHP function 'iconv()' content and keywords will be converted into UTF-8.
This step is successful, if the required charset (for the content to be translated) is part of your local PHP installation. In order to find out which charset are available in your installation, notice the files in server folder:
.... .../apache/bin/iconv/
Depending of the installation you will find about 200 charset files that iconv() is able to translate into UTF-8
3. If the PHP function fails, finally the class 'ConvertCharset' is invoked. This class, originally designed by Mikolaj Jedrzejak, enables translation for a lot of charset. But it takes more time than the compiled PHP function 'iconv()'.

As result of this translation, you are also enabled to search for words that contain non-Latin characters.

Some additional remarks:

It is not enough to select the checkbox 'Convert all into UTF-8 charset'. You are obliged to do a fresh index or 'Erase & Re-index'. Also if you de-select this option, you must update your database again. Otherwise content of database and search query will not match, as different character coding is used.

Even if you didn't notice immediately: with release date of Sphider-plus version 1.2 all files in folder .../languages/ have been converted to Unicode. (By the way, I'm still hoping for your translations.)

In order to enable translation of all entities into UTF-8, the Unicode requires also upper case characters. So, if you enable the UTF-8 option, your query 'html' will not deliver results for sites and files that contain the string 'HTML'. Both keywords are stored separately. This behavior is in contrast to former versions of Sphider and Sphider-plus and should be remembered.

But I also improved the 'Did you mean' option. Let's assume there are no sites with 'html' but some with 'HTML'. If your query is 'html', Sphider-plus will not capitulate without results. You will be asked 'Did you mean HTML?'

Starting with version 1.6 Sphider-plus offers the additional option:

'Enable distinct results for upper- and lowercase queries'

If enabled in Admin settings, everything remains as described above. But if this checkbox is unchecked, result listing will deliver all results independent of the query input. HTML, html or even hTmL will produce the same (all) results.

The checkbox for this option is placed with full intention in section 'Spider settings', as activating and also deactivating requires always an 'Erase & Re-index' procedure.

10. Search modes

Beside the original Sphider search queries like:

- Search for a single word
- AND and OR search
- Search for a phrase

Sphider-plus offers 5 additional modes to enter queries:

- Search with wildcard
- Strict search
- Tolerant search
- Link search
- Media search

Wildcard, strict and tolerant search modes are available only for single query word input.

10.1 Search with wildcards *

This mod enhances the Sphider-plus capabilities to search also for parts of a word. The mod is invoked by entering a * as wildcard for the unknown part of the search query.

Wildcards could be used like:

- *searchme
- *searchall*
- *search*more*

Depending on Sphiders keyword table, a lot more results may appear. In order not to confuse the user, the printout of relevance (weight/hits) is suppressed in result listing.

10.2 Strict search !

This variant is invoked by entering a ! as first character of the search query. If you search for '!plus' only results for the word 'plus' will be presented in the result pages. No results for words that contain 'spider-plus' or 'spiderplustec' will be shown. This is the reverse function of 'Search for part of a word by means of * wildcards'. Strict search only indicates results in the text part of the indexed pages. If utf-8 support is enabled and 'Distinct results for upper- and lowercase queries' is also activated, 'Strict search' will respond case sensitive.

10.3 Tolerant search

This mod enables a tolerant search for Sphider. Selectable in Search-form like AND, OR and Phrase Search a new item "Tolerant Search" is added.

If this item is selected, query input "perdida" will also deliver results for all sites that contain the word "pérdida". Inverse function is also implemented: "pérdida" input will deliver all results for "perdida".

If enabled, this mod equalizes search input for e=é=è=ê and all the other vowels like: ä=a=à=â , ü=u , o=ö etc. The upper-case letters like Ä=A are also taken into account. Tolerant search overwrites the 'Distinct results for upper- and lowercase queries' setting and will mark all results.

Natively developed to deliver most possible results for queries with entities and accents and also to simplify user input, this mod also delivers results that are "like" the query input. So something as the "Did you mean" facility is already integral part of this search method.

10.4 Link search site:

Invoked by starting the query input with ' site: ', the user is enabled to search for all pages of a domain. It is not necessary to enter the full domain address. For example if you enter 'site:sphider-plus.eu' you will get a list of all pages that belong to the domain <http://www.sphider-plus.eu>

If the search query is part of more than one domain address in Sphiders site table, a list of these domains will be presented as intermediate result. If you then click on the desired domain of this list, all links (pages) of this domain will be presented as final result listing.

10.5 Media search

Media search is invoked by an additional checkbox in the Search Form. Media will be found individually for:

- Images
- Audio
- Video

Entering 'media:' (without quotes) will present all media stored in the database. For more details please notice the chapter [Media Search](#)

11. Chronological order for result listing

Sphider-plus offers 5 methods how to sort the results:

- By relevance (weight/hits)
- Main URLs (domains) on top
- By URL names and then weight
- Only top 2 per URL (like Google)
- Most Popular Links on top

The according selection is placed in Admin settings section 'Page index weight'. This was done, because the method **'Main URLs (domains) on top'** changes the weight of words in all involved pages. The normal relationship (weight) as defined for:

- Word in web page Title tag
- Word in the Domain name
- Word in the Path name
- Word in web page Keywords tag
- Word in full text

remains valid and is calculated as before. But additionally all words in main URLs are weighted with an additional factor. In other words:

The weight of each word found in a page that is accepted as main URL, will become an increased weight. A multiplier performs this. This multiplier may be reduced or increased by modifying 'Multiplier for words in main URLs'. So you may influence the weight of main URLs in front of all other pages.

Per default this multiplier is set to 5. In order to keep the results of the main URL on top, it might become necessary to increase this factor slightly. An according input field called 'Multiplier for words in main URLs (domains)' is also added to the same section of Admin settings. This multiplier is valid only for the method 'Main URLs (domains) on top'.

In order to be accepted as 'Main URL', the address must fulfill the following conditions:

1. The domain may include at maximum 2 dots. So valid are:
http://www.abc.de and http://forum.abc.de
But sub domains like http://www.forum.abc.de will be ignored.
2. As file path are only accepted: index.php, index.html, index.htm or a blank path.
If additional file names should become valid, open the file ../admin/spiderfuncs.php and
search for: `$act_path = str_replace('index.htm', '', $act_path);`
Beyond this row include a row like above, but holding the additional file name.

The current selection of 'order for result listing' is visible for the users as additional headline on all result pages. If not desired, this option can be de-selected in Admin setting: 'Show mode of chronological order for result listing as headline'.

In order not to confuse the user, for the 3 methods

- By URL names and then weight
- Only top 2 per URL (like Google)
- Most Popular Links on top

the output of relevance (weight/hits) is suppressed in result listing.

Additional reminder: As weight of each keyword is already defined during index and re-index, any modification of weight or 'Order for result listing' requires an 'Erase & Re-index'.

For the '**Most Popular Links on top**' method, Sphider-plus uses the before learned link acceptance. If a user leaves the result listing by clicking on any of the offered links, Sphider-plus will memorize this decision. The user is temporary redirected to the script ../include/click_counter.php, which stores the users link decision, last query, time and date before leading the user to the real destination.

This link specific 'best click' counter is used as teach-in to define the chronological order of result listing. In order to prevent promoted clicks on a specific link, there is a delay timer before the next user click will be accepted. To be set in Admin /Settings/ Index Log Settings, the setting defines the idle-time in seconds.

If there are more results as rated links, the rest of the result listing will be presented by relevance (weight), using the weighting of the last index / re-index.

As the 'Most Popular Links On Top' item overwrites all other order of result listing, it might be selected without re-index procedure.

For the method of result ordering '**By relevance (weight/hits)**' the weight is calculated as described above in this chapter. Situation changes, if in Admin settings the item 'Instead of weighting %, show count of query hits in full text' is activated. Now only the hits in full text are used to calculate the order of result listing. Keyword hits in URL name, path, title tag etc. are not taken into consideration.

The weighting of:

- Word in web page Title tag
- Word in the Domain name
- Word in the Path name
- Word in web page Keywords tag
- Word in full text

may be influenced individually for personal preferences.

12. PDF converter for Linux/UNIX systems

Starting with version 1.5 Sphider-plus includes one PDF converter for Windows systems and another converter for LINUX/UNIX systems. The Windows converter is ready for use, but the Linux version needs your attention. So, before using the Linux converter you need to do the following steps:

1. Identify the physical path of your web site. if available, Admin / Statistics / Server / PDF-converter should present this info. Otherwise your hoster should provide this information anywhere.
2. Take this full path as above and use simple slashes (not double backslashes) and include your individual path to the file .../converter/pdftotext
First line of that file holds `#!/bin/sh` nothing more.
Second line of that file begins with a slash and ends with the minus sign.
A third line is not allowed.
3. Set permissions of both pdftotext and pdftotext.script to 755 or 777 (whatever needed to run correctly).
4. Set permissions of the converter dir to 777! Otherwise indexing fails because of pdftotext is unable to write a temp file needed!
5. Adapt the variable `$pdftotext_path` in file .../settings/conf.php to your personal needs:
`$pdftotext_path = '/PATH/TO/YOUR/WEB/DOWN/TO/converter/pdftotext';`
Do this modification with closed Admin browser window.

Perform the above modifications with closed Admin browser window.

The DOC, RTF, PTT and XLS converter are not available for LINUX/UNIX systems.

Warning: When editing the file 'pdftotext', additional characters, line feeds, blank rows, or what ever are not allowed to be added. The content of this files must remain **pure**. Otherwise the server will be unable to find the PDF converter. Please also take notice of the FAQ chapter: [PDF documents are not indexed](#)

13. Clean resources during index / re-index.

In order to prevent performance problems and memory overflow for large amount of URLs, Sphider-plus may clean unused resources during index / re-index. Selectable in Admin settings, this item periodical will:

- Free memory that is allocated to unused MySQL recourses.
- Unset PHP variables, which are no longer required.

As this clearing work is done several times during index / re-index of every URL, additional capacity is required. Consequently overall indexing time will increase. So this item should be selected only for huge amount of URLs. Depending on

- Memory size allocated to PHP
- Total number of URLs
- Number of internal and external links
- Size of text to be indexed for each page
- CPU clock rate
- System RAM

there will be an individual limit when to enable this feature. Following the discussion on the Sphider forum this feature should be activated only if > 100 sites are to be indexed, or when Sphider-plus dies a silent death during index prodcedure, not indexing any more sites.

Please take notice of the FAQ chapter:

Error message: "[Unable to flush table 'addurl'](#) "

and

Error message: " [Access denied; you need the RELOAD privilege...](#) "

14. Enable real-time output of logging data

Up to version 1.5 of Sphider-plus, during index / re-index there was no printout available because:

- Several servers, especially on Win32, buffer the output from the script until it terminates before transmitting the results to the browser.
- Server modules for Apache do buffering of their own that will cause flush() to not result in data being sent immediately to the client.
- Browser may buffer its input before displaying it. Netscape, for example, buffers text until it receives an end-of-line or the beginning of a tag, and it won't render tables until the </table> tag of the outermost table is seen.
- Some versions of Microsoft Internet Explorer only start to display the page after they have received 256 bytes of output.

As progress was not presented during index / re-index procedure, waiting for results became a pain in the neck.

Selectable in Admin setting together with the update interval (1 - 10 seconds), AJAX technology was the approach to realize this feature.

Pressing one of the 'Start index / re-index' buttons, three additional scripts are involved.

(onclick="window.open('real_log.php')"))

.../admin/real_log.php By opening a new browser window / tab, this script takes over to display latest logging data. Requesting fresh data from the JavaScript file 'real_ping.js', all new logging data will always be placed into <div id='realLogContainer' />. So, better not to press the 'Reload' button of your browser. The current <div /> might be already empty.

.../admin/real_ping.js Script that transfers requests from HTML client to PHP server script and vice versa. Handling refresh for real-time logging during index and re-index procedure by means of asynchronous requests (AJAX) to the server.

.../admin/real_get.php This script delivers 'refresh rate' and latest 'logging data', requested from the JavaScript file 'real_ping.js'. Also performs the reset of the 'real_log' table in Sphiders database.

Latest logging data is delivered by the .../admin/messages.php script that, besides writing into the normal log file, feeds the table 'real_log' in Sphiders database. This is the buffer for latest logging data.

Prerequisites are the enabled 'Log spidering dates' and 'Log file format = HTML'. When activating the real-time output, both pre-conditions are automatically selected.

15. Error messages and Debug mode

Starting with version 1.7, Sphider-plus offers the capability to enable / disable the output of MySQL error messages as well as PHP error messages. To be activated in Admin / Settings / General Settings, this capability should only be used for debug purpose. It is recommended to disable the output of these messages for production systems, as they could reveal sensitive information.

Selection of the 'Debug mode' is implemented in Admin settings. If the 'Debug mode' is enabled, for all pages that are indexed the found links and keywords are presented in Log-file output and also in Log-file real time output. It has to take into consideration, that only the new links and keywords found on the respective page will be presented. Links and keywords already stored in Sphider-plus database (because they were already detected on a former page) will not be presented again.

The 'Debug mode' adds a comma and a blank to each keyword. So, debug output will be something like:

New keywords found here:

abc, defg, hijklm, nop, . . .

As Spider-plus also indexes special characters like commas and dots, keywords like **defg**, and **hijklm**. will be presented like:

New keywords found here:

abc, defg,, hijklm., nop, . . .

The Debug mode only modifies the Log-Files. Spider-plus database remains unaffected and will hold the same values as indexing without Debug mode. In other words, activating / deactivating this mode has no effect on the later search results.

For activated Debug mode, also the output of MySQL and PHP error messages is activated. Debug mode overwrites the according setting. When deselecting a debug session, also the error messages must be disabled manually.

If 'Debug mode' is enabled, also the cache activity is presented above the Result listing in the form of status messages.

16. Delete secondary characters

This feature was implemented in order to kill unimportant (secondary) characters at the end of words and also as leading characters of words.

If activated in Admin / Settings / Spider Settings, the following characters in front of words are deleted:

" (

Also, if at the end of words, these characters are deleted:

) "). , . : ? !

If placed at the end of words that contains only digits, the dots are not deleted (e.g. 27.). So the search for
27. November 2008
remains available.

For personal requirements the following two rows in .../admin/spiderfuncs.php may be edited.

```
$file = preg_replace('/', |[^0-9]\. |! |? |" |: |\\ |\\), |\\)./', " ", $file); // kill characters at the end of words
$file = preg_replace('/ "| \\(|, " ', $file); // kill special characters in front of words
```

Warning: This option should be used with special care and not be activated for non ISO-8859 charsets. Some special characters as part of the word ending might be erased by accidental.

17. Media search for images, audio streams and videos

17.1 Media indexing

Index of media files is enabled by separated Admin settings for:

- Images
- Audio streams
- Videos

Three separate files in subfolder `.../include/common/` that are named

`image.txt`
`audio.txt`
`video.tx`

hold a list of associated file suffixes. Only media files with the corresponding suffix will be taken into account during index / re-index procedure. These three files may be edited for personal purpose.

In order to be indexed, for images additionally the minimum width and height (H x V pixel) may be defined in Admin settings. Image size will be observed for the following image types:

`.bmp .gif .j2c .j2k .jp2 .jpc .jpeg .jpeg2000 .jpg .jpx .png .swc .tif .tiff .wbmp`

Admin settings also allow selecting whether embed and nested media files should be indexed. This was implemented, as some server hide their media files as embedded objects.

Another Admin setting is used to enable indexing of external media content. When linked by the currently indexed page, also external hosted media files will be indexed. This setting is independent from the Sites / Advanced Option setting 'Sphider can leave domain', which is used for text links only.

Depending of the installed GD-library, during index / re-index procedure Sphider-plus will create thumbnails for the following image types:

`gif, png, jpg, ipeg, jif, jpe, gd, gd2 and wbmp`

Details about the currently installed GD-library (as part of the PHP environment) and the supported image formats are available at:

Admin / Statistics / Server Info / Image funcs.

Thumbnails will be created as 'gif' or 'png' files. To be selected in Admin settings, the gif files do have a lower quality, but will reduce the required memory for about 50%. Re-sampling the original images, size of thumbnail is defined to a maximum of 160 x 100 pixel. In result listing these stored thumbnails are used as preview.

As far as available the Meta data ID3 and for images EXIF information is indexed and herewith become searchable.

In order to create thumbnails and to index ID3 and EXIF information, it is necessary to download the media file. For pages with multiple media content, the time for index /re-index procedure may increase dramatically.

As ID3 information is not available for all audio and video files, the minimum play time in order to be indexed was not yet implemented.

In order to save memory resources, Sphider-plus does not store the media content. Only the links, thumbnails and Meta information are stored.

The limit in Admin settings " Max. links to be followed for each Site" is not taken into account for media links. Only page links are counted and the limitation is valid only for page links.

17.2 Not supported media content

The following examples demonstrate the currently existing limitations for media data that will not be indexed:

1. If inserted in documents like pdf, doc, ppt, etc.

2. If inserted in Java or applets like:

```
<P><OBJECT classid="java:program.start"></OBJECT>
```

and also direct applet implementations like:

```
<APPLET code="Bubbles.class" width="500" height="500">  
Java applet that draws animated bubbles.  
</APPLET>
```

3. Image maps that are server-side or client-side included like:

```
<P><A href="http://www.acme.com/cgi-bin/competition">  
  <IMG src="game.gif" ismap alt="target"></A>
```

17.3 Search for media content

The search mode is enabled by the checkbox

'Beside text results also show media results in result page'

in Admin / Settings / Search Settings

Once activated, the result listing for each keyword match will be separated into the 4 sections:

- text results
- image results
- audio results
- video results

Each section is marked with an according thumbnail. Result listing will present only those sections that contain results.

Each section will present result number, media title and the page address (link) at which the media was found. The text section will show the results as previous with highlighted keywords and surrounding text.

The image result section additionally presents a thumbnail, the image size (H x V pixel) and a link to EXIF information for each found image. Clicking on the thumbnails will open the original image in a new window / tab.

Video and audio results are presented with title, play time and a link to ID3 information. Media content will be opened with the belonging software by clicking on the media title.

As the media sections are presented separatley for each keyword match, an additionl link called 'All media' is shown. Clicking here will force Sphider-plus to present all media results of the corresponding page (link). In order to return to the standard search modus, the section thumbnails could be clicked.

The search function at first will look for text results (keyword match) and receive the according pages (links). Afterwards media files are searched for the pages defined by the text results. So, only those media results that also generate text results will be presented in result listing.

To get all media results (independent of the text results) another search mode is available:

If in Admin / Settings / Search Settings the checkbox

'Advanced search? (Shows 'AND/OR/PHRASE/TOLERANT/MEDIA' etc.)'

is activated, the Search Form will present the additional checkbox

'Search only Media'

If this checkbox is activated, only media results will be presented in result listing, while possible text results will be ignored.

Media search follows the rules of pre-defined categories. If 'Search only in category xyz' is selected in Search form, media results will be presented only as found in the particular category.

Search input for media queries is always interpreted as tolerant. So the query 'logo' will present results e.g. for the image 'sphider-logo.gif', while the input 'gif' will show all available gif files.

The query 'media:' (without quotes) forces Sphider-plus to search for all media stored in its database. If used together with a category selection, all media content of the particular category will be presented.

If in Search Form the checkbox 'Search only Media' is activated, also the suggest framework will present only media suggestions; taking into account also the eventually pre-selected limitation for category search.

An additional Admin setting in section 'Suggest options' allows selection whether suggestions should be taken also from EXIF info and ID3 tags. Never the less suggested keywords will always be the title of the media file.

For media search the Admin setting 'Enable distinct results for upper- and lower-case queries' is also taken into account.

17.4 Statistics for media content

In Admin / Statistics the following tables are available:

'Most Popular Media' presenting:

- Thumbnail
- Details like 'Title' and 'Found at'
- Total clicks
- Last clicked
- Query

'Indexed Image Thumbnails' presenting:

- Thumbnail 150 x 100 pixel
- Image details like title, filename size of original image, link- and thumb-id
- Option to delete the thumbnail

In order to open the media file, all tables contain active links.

Media results are also stored in 'Search log', and are presented like the keyword results with:

- Query
- Result count
- Queried at
- Time taken
- User IP
- Users country code
- Users host name

18. RSS and Atom feeds

To be activated in Admin / Settings / section 'Spider settings', content of RSS (v.0.93 - v.2.0) and Atom (v.1.0) feeds will be indexed / re-indexed. The following content is indexed and herewith becomes searchable after indexing:

- Channel/Feed: Title and Description.
- Item: Titel, Description, Guid, Author, Category, Publication date and time.

RSS and Atom feeds are treated as normal text pages. The suggest framework will offer keyword proposals. Also pre-selection of categories is taken into account. Feed links are treated like the standard page links, so that the limit in Admin settings "Max. links to be followed for each Site" is influenced also by feed links (they count).

19. Result cache for text and media queries

To be activated in Admin settings, section 'Search Settings', the cache will store the results of the 'Most Popular Queries'. Before connecting to the database, each query will request the cache for results. If available, results are presented extremely fast. On the other hand each query, necessary to get results from the database, will automatically store its result into the cache.

Individual cache results are stored following the different Search selections (AND, OR, Phrase, Tolerant). Also individualized cache results are stored for each category and all-sites search requests.

Text and media queries cooperate with different caches. Size of each cache is definable in Admin settings [MByte]. On overflow of a cache, the least important result is deleted from the cache, while 'Most Popular Queries' is updated with each search input.

If in Admin settings the 'Debug mode' is enabled, cache activity is presented above the Result listing in the form of status messages. Text cache and media cache could be manually cleaned in Admin 'Clean' section, also offering the count of files in each cache and the consumed memory space separately for each cache. Another selectable cache setting allows automatic cache reset, performed on 'Erase & Re-index' procedures.

Another Admin setting is called:

'Define **max. number of results** (links) per query stored in cache'

The separate input fields for text and media cache allow limitation of results found for a query. Usually a search engine user will not follow 9999 results found for a query. To limit the number of results will speed up first search in database, reduce required cache size and will also speed up result presentation when fetching the results from the cache.

Definition of required **cache size**, that is also to be defined separately for both caches, depends on personal preferences. There is a conflict between two opposed requirements: the cache should hold as much as possible 'Most Popular Queries' but not consume too many resources by controlling hundreds of files in a big memory. For a first assumption, size per result should be defined to 2 Kbyte. Multiplied with the matches in database (e.g. found in 20 pages), each result requires approximately 40 Kbytes of RAM. So, a cache of 2 MByte could hold the results for 40 to 50 'Most Popular Queries'. After some time of usage, it might be helpful to observe the information given in 'Clear' section of Admin. Count of result files in cache and consumed memory space are presented. Depending on personal preferences, consumed result size and count of query hits in x pages, it might be necessary to adapt the size for text and media cache.

20. Multiple database support

20.1 Overview

Starting with version 2.0, Sphider-plus offers the capability to cooperate with multiple databases. Currently prepared to work with up to five databases, the development was done under the following aims:

Independent allocation of different databases for the tasks:

- Admin
- Search user
- Suggest URL user

This offers the capability to assign the 'Search' user to database1 and let him use the search engine. Meanwhile the 'Admin' may re-index database2. Also 'add new sites' and index them into database2 is performed by the Admin without disturbing the 'Search' user. Also backup, restore and copy functions could be done by the Admin without influence on the availability of the search engine. Later on the Admin may switch the 'Search' user to the updated database, or copy the fresh database content into the 'Search' user database.

As Sphider-plus has to survive also in Shared Hosting applications there are some limitations for multiple database support:

- It is not possible to cooperate with a cluster of databases.
- Master/Slave Replications are not supported, because the MySQL configuration file my.cnf is not accessible.
- Sharding by scaling data-tables is not supported.
- Dynamical allocating as a pro-shared assignment is not possible.

Sphider-plus Admin interface offers the management of multiple databases. There are different menus in section 'Database' as described below.

20.2 Definition and configuration

Sphider-plus version 2.0 (and following) does not require the install_all.php script any longer. Database assigning and table installation is integrated into the Admin interface.

The menu for database definition and configuration is protected by an additional login. Independent from the Admin login, a username and password is required to enter into this section. Username and password are defined in the file .../admin/auth_db.php. As per default download, username and password are both set to 'admin'.

Entering the first time into this section, there will be several warning messages. At minimum one database has to be defined by:

Name of database
Username
Password
Database host
Prefix for Tables

Pressing the 'Save' button will assign Sphider-plus to these database definitions. Never the less, the warning message **'Tables are not installed for database x'** will remain in the Database settings overview.

The **'Install all tables for database x'** is an independent procedure, which has to be invoked by the Admin after the database has been allocated. Chapter [Enhancing functionality of multiple database support](#) will describe the reason for these two independent steps.

If the database is allocated and the tables are installed, the message **' Database x settings are okay.'** are displayed in the settings overview; showing separately the situation for each of the five databases.

If the application should work with only one or two databases, the settings for the non-required databases may remain blank. A corresponding message will be displayed:

Mysql server for database 3 is not available!

Trying to reconnect to database 3 . . .

Cannot connect to this database.

Never mind if you don't need it.

So the Admin may assign up to five databases, as required for the application. Assigning of another (the next) database will be possible only, if the settings for the previous database are okay and the tables are installed. Further database setting fields are suppressed.

20.3 Activate / Disable databases

Next step to get multiple databases to work will be the activation of the databases. This section of the Database Management will present only those databases, which are correctly configured, assigned and do have a set of installed tables as described in chapter [Definition and configuration](#).

There are three settings available in the 'Activate / Disable' section:

- Select active database for Admin
- Select active database for 'Search' user
- Select active database for 'Suggest URL' user

Each setting allows activating of one database. So independent use of databases is enabled for 'Admin', 'Search' User and 'Suggest URL' user.

After activating the databases for the different tasks, multiple database support is ready to use. The currently activated database is displayed for the Admin in 'Sites' table like:

Database 1 - Displaying URLs 1 - 10 from 25

If 'Debug' mode is activated in Admin settings, also the result listing will inform the user about the actual situation:

Results from database 2

20.4 Backup & Restore of databases

This section of the Database Management will present only those databases, which are correctly configured, assigned and do have a set of installed tables as described in chapter [Definition and configuration](#).

This section enables the Admin to create backups from the current situation of a selectable database. Vice versa the backup files may be restored into the database.

Backup files are compatible to phpMyAdmin structure and contain the table prefix and date + time of creation as part of the file names. Backup files are stored in subfolders (.../admin/backup/dbx), separated for each database.

Restore of backup files is only possible into that database, which had been used before to create the backup files. Current content of the database tables (those with the same table prefix) will be destroyed by the restore procedure.

20.5 Copy & Move

This section of the Database Management will present only those databases, which are correctly configured, assigned and do have a set of installed tables as described in chapter [Definition and configuration](#).

This section allows to copy or to move the content from one database to another. By selecting:

- Source database
- Destination database
- and
- Define Copy or Move utility

it is possible to copy / move the content from one database to any other database. Beside the table content, both utilities inevitably will also copy the table suffix (of the source db) into the destination database. If tables with the same prefix already exist in the destination database, the content of these tables will be overwritten. Beside the table content also the corresponding thumbnails will be copied.

In contrast to the 'Copy' utility, the 'Move' function additionally will clear the source database and delete the corresponding (source) thumbnails.

20.6 Enhancing functionality of multiple database support

1. 'Backup & Restore' as well as the 'Copy / Move' function will always work with all tables of a selected database. In contrast to these global actions, the 'Import / Export URL list' function is only acting with the currently (for the Admin) activated table prefix. This allows a selective import and export of only those URLs, used for the activated tables as defined by the prefix. The name of the exported URL list contains the (source) database number, the table prefix and the date of creation. Crossover usage of URL lists is enabling to import any URL list (created from database x) into database y

2. When configuring databases, it is strongly recommended to create and use prefixes for the tables. Table prefixes are the key for creating new sets of tables in each database. As described in chapter [Definition and configuration](#), the tables need to be installed separately; after the configuration of the database was saved. After these settings are finished and the database is assigned, Admin may use this database and index sites into the database tables with the given table prefix.

It is evident that one database could be configured with several table prefixes. That is the key for additional 'virtual' databases. By configuring the given database with a new table prefix, Admin is able to install another set of tables into the same database. This set of tables (with the new prefix), may be used to index another set of sites into the same database. This is performed without destroying the content of the prior used tables.

3. The above mentioned allows to add quasi-additional databases without really creating new databases. It was also mentioned before that Sphider-plus has to survive in 'Shared Hosting' applications. Consequently Admin may assign one database to the 'Search' user.

But there is a feature integrated into Sphider-plus to bypass this restriction. Assuming that result listing should be offered in two (or even more) versions. For example in English and another language. One result listing for global users, the other for registered users. One info result, one shopping result listing etc.

To enable such a feature, the search form of Sphider-plus contains two hidden variables called 'db' and 'prefix':

```
<input type="hidden" name="db" value="0" />
<input type="hidden" name="prefix" value="0" />
```

As long as the values are set to '0', the search script will use the settings as defined in Admin settings: "Select database for 'Search' user". This standard setting may be used for the first search form, offering the results of the first database (which e.g. holds the English results). But for a second search form, the value for 'db' may be set for another database (1-5) that holds the results of the second language. The value setting of the second search form will (temporary) overwrite the Admin settings for its own result listing.

The same procedure may be used for the 'prefix' variable. Entering here the prefix name of a set of tables (holding special results) will temporary overwrite the standard table prefix as defined in Admin settings. Combining different values for 'db' and 'prefix' enables unlimited individualization of result listing for specific search forms.

This implementation could be interpreted as a super category feature. Not requiring the selection of a category, or even a sub-category, by the 'Search' user. Not predicating that the normal category function would be lost by use of multi database support and its extended features.

21. FAQs

21.1 UTF-8 support does not work.

In order to enable correct utf-8 support, charset of the search page and also of the result page must be set to utf-8. So, if you integrate (embed) Sphider-plus into your site design, the html header of your files should include something like:

```
<meta http-equiv="content-type" content="text/html; charset=utf-8">
```

21.2 Can't search for long words.

Current length of words to be indexed is set to 30 character. This might be not enough for several applications. In order to increase this value to 255, the database table \$mysql_table_prefix 'keywords' must be manually modified. Set keyword varchar(255) not null

Also in .../admin/spiderfuncs.php the row

```
if (strlen($word)<= 30) {
```

must be modified to

```
if (strlen($word)<= 255) {
```

Please remember to do an 'Erase & Re-index' after you made this modification.

Starting with version 1.4, the database installation and spider scripts were changed to the new values.

Example:

The Gregorian word

□□□□□□□□□□□□□□□□(13 letters)

is represented in Sphider-plus database as:

áf>áfœáf~áf~áf•áfœáf"áfšáfáf•áfáfœáf~ (39 letter string)

21.3 Can't search for words with non-Latin characters.

in Admin settings the two checkboxes:

- Convert all into UTF-8 charset
- Enable distinct results for upper- and lower-case queries

must be selected. Even if the site to be indexed is UTF-8 coded. In order to get the query input UTF-8 coded and also to present the results correctly, Sphider-plus needs these two checkboxes to be activated. If the site to be indexed is not UTF-8 coded, the limitation for words to be queried is defined by the 63 charsets currently available to be transferred to UTF-8. For details, please take notice of the chapter:

[UTF-8 Support and 'Preferred Charset'](#)

21.4 How to bypass the Admin log in.

For Intranet applications and during debugging it might be more comfortable to bypass the Admin authorization. There are two possibilities:

Option 1. This version still shows the **Log In** page as warning that you now enter into the Admin section, but you just have to click on the Login button.

```
In .../admin/auth.php set
    $admin = "";
    $admin_pw = "";
```

Option 2. This version removes the **Log In** page totally:

```
Rename the file .../admin/auth.php into auth_backup.php
Rename the file .../admin/auth_bypass.php into auth.php
```

21.5 Links are not followed during Re-index, only main URL is indexed (option 1).

It is not a bug, it is a feature. If 'Follow sitemap.xml' is activated in Admin settings, links will only be followed if:

- 'last modified' date in sitemap.xml is newer than Sphiders 'last indexed' date.
- New link that is not yet known in Sphiders link table.

The main URL will always be indexed, because status and content of the sitemap file is required for further decision what necessarily has to be indexed. Because only relevant pages will be indexed, this approach significantly reduces the time required for index and re-index.

21.6 Links are not followed during Re-index, only main URL is indexed (option 2).

If you use a .htaccess file on your server in order to redirect requests, or to 'produce' seo friendly link names, you must enable the checkbox 'Spider can leave domain' in Admin/Sites/Options/Edit/. Otherwise Spider will not follow the redirect directive of your .htaccess. file.

21.7 How to integrate Sphider's search field into existing pages.

Add the following code at the according position into the HTML code of your page and personalize the path_to_sphider-plus address relativ to the HTML code:

```
<form action="/path_to_sphider-plus/search.php" method="get">
<table border="2" width="150" cellpadding="0" cellspacing="2">
<tr>
<td align="center"><input type="text" name="query" size="30" value="" /></td>
<td align="center"><input type="submit" value="Search" />
<input type="hidden" name="search" value="1" /></td>
</tr>
</table>
</form>
```

This simple example does not support all facilities of Sphider-plus. It is foreseen only as first step into your personal adaption. For example if you add

```
<input type="hidden" name="mark" value="markyellow" />
```

the found keywords will be marked yellow.

More details and examples how to integrate Sphider-plus into existing pages may be found on the Sphider forum. For example at:

<http://www.sphider.eu/forum/read.php?2,4505>

21.8 Error message: "Warning: set_time_limit() . . . "

Sphider does not work if the server is in 'safe' mode. That server setting must be disabled in the PHP initialisation file (e.g.: .../apache/bin/php.ini).

```
safe_mode = Off
```

The current state is shown in Admin / Statistics / Server Info / php.ini file key: safe_mode

Before modifying this value, stop your server and afterwards restart the server again.

21.9 Error message: "Unable to flush table 'addurl' "

Sphider has not enough privileges to close the tables of your database. Sphider needs the privilege 'Reload' to perform the flush instruction (MySQL-Manual chapter 13.5.5.2). Please check your database installation, grant enough privileges to Sphider and shut down other scripts that could use the Sphider database.

If you don't succeed with these fundamentals because you use a shared hosting server, open the file
.../admin/db_backup.php
and delete the row
mysql_query("FLUSH TABLE \$row[0]") or die("Unable to flush table \$row[0].");

Also open the file
.../admin/spiderfuncs.php
and delete the row
mysql_query("FLUSH QUERY CACHE");

Please keep in mind that by deleting these rows you will loose parts of the 'Optimize database' and 'Clean resources during index/re-index' functions.

21.10 Error message: " Access denied; you need the RELOAD privilege. . . "

The same problem as error message: "Unable to flush table 'addurl' " This time your server sends the error message. Sphider has not enough privileges to flush the tables of your database. Sphider needs the privilege 'Reload' to perform the mysql flush instruction. For more details see chapter above.

21.11 Fatal error: "Allowed memory size of xxx bytes exhausted (tried to allocate yyy bytes)"

This is a limitation of your server that does not allow PHP to allocate enough memory. In order to prevent this error message, increase the memory size in the PHP initialisation file (e.g.: .../apache/bin/php.ini)

```
memory_limit = 64M
```

The currently allocated memory size is shown in Admin / Statistics / Server Info / php.ini file
key: memory_limit

Before modifying this value, stop your server and afterwards restart the server again.

21.12 PDF documents are not indexed

If you are sure that physical path to the converter is correct (see: Admin / Statistics / Server-Info / PDF-converter), but your PDF documents are not converted, there might be another (final?) approach. Technical support for your hosting service may tell that you could run scripts from any directory, but it looks like that is not true for all providers. Meanwhile there are some according user reports.

Move the 2 scripts

pdftotext

and

pdftotext.script

to a directory called 'cgi-local' or something similar that your provider offers for cgi, set the proper permissions, change the \$pdftotext_path in all involved scripts to the new destination and then run the index / re-index procedure.

21.13 PHP security info is not presented in Admin Statistics

Unfortunately not all servers are supporting this feature. They take their security settings as a secret. A 'blank' admin is the typical response. As consequence, this feature per default is disabled. In order to get the security info, perform the following steps:

In ../admin/admin_header search for the row:

```
// require_once('PhpSecInfo/PhpSecInfo.php');
```

Uncomment that row by deleting the //

Also in ../admin/admin.php search for the row:

```
// phpsecinfo();
```

Uncomment that row by deleting the //

21.14 What kind of input validation is performed?

The following protections are implemented:

- Prevent SQL-injections
- Prevent XSS-attacks
- Prevent Shell-executes
- Suppress JavaScript executions
- Suppress Tag inclusions
- Prevent Directory Traversal attacks
- Delete input if query contains any word of (editable) blacklist
- Prevent buffer overflow errors.
- Suppress JavaScript execution and tag inclusions masked as XSS attacks.
- Prevent C-function 'format-string' vulnerability.

21.15 How to protect Database management against Admin access?

As per default, the submenu 'Configuration' is already protected by a separate username and password. This protection could be extended to the complete Database management by uncomment the row:

```
//include "auth_db.php";
```

in the following scripts:

```
.../admin/db_activate.php  
.../admin/db_common.php  
.../admin/db_copy.php  
.../admin/db_main.php
```

22. Change log

22.1 Version 1.0 - 1.9

22.1.1 Version 1.0

Release date: February 15, 2008

Based on the original Sphider v.1.3.4.a by Ando Saabas, the following items are modified:

Define min. relevance level (weight %) for results to be presented at result pages.

To be defined in Admin settings.

Enable user suggestion for new Url to become part of Sphider-plus database (addurl by user).

- To be activated in Admin settings, the user is enabled to suggest sites.
- If enabled, a link at the footer of the result page leads to the suggestion form.
- The user will have to fulfil 'Url', 'Title', 'Description' and 'Dispatchers e-mail account'.
- Checked for valid input, DNS availability and MX-RR validation of dispatchers account.
Suggested Url will be stored in the Sphider-plus database until Admin decision.
- Suggested sites are presented in Admin submenu 'Approve sites' so that the admin may decide to
 - accept
 - reject
 - bann
- Result of decision will be mailed to the dispatcher (if selected in Admin settings).
- Included is also the submenu 'Banned domains' to refuse all sites not welcome for this search-engine.

Create a sitemap during index/re-index.

- Compatible with <http://www.sitemaps.org/schemas/sitemap/0.9>
this module automatically creates a sitemap.xml file.
- In Admin settings the folder name for the sitemaps can be defined.
- The xml files will be individually named like 'sitemap_www.abc.de.xml'
- When running a 'Re-index', 'Re-index all' or 'Erase & Re-index'
existing sitemaps will be overwritten with the actual data set.

For index/re-index follow sitemap.xml (to be activated in Admin settings).

If available Sphider-plus will use the sitemap to follow all links of that domain.

This increases significant the speed for index and re-index.

The mod will also force Sphider-plus to re-index only links that are:

- New and not yet known in Sphiders link table
and
- Links whose 'last modified' date is newer than Sphider's 'last indexed' date.

Search for part of a word by means of * wildcards.

This mod enhances the Sphider-plus capabilities to search also for parts of a word.

Invoke this mod by entering a * as first character of your search query.

You may use * wildcards like:

*searchme
searchall
*search*more*

Search !strictly for the search query.

Invoke this variant by entering a ! as first character of your search query.

If you search for '!plus' only results for the word 'plus' will be presented in the result pages.

No results for words that contain 'spider-plus' or 'spiderplustec' will be shown.

This is the reverse function of 'Search for part of a word by means of * wildcards'

Search for all pages of a site.

This utility searches for all that pages, which belong to a domain.

Initialize your search query with 'site:' followed by the domain you want to check.

Also parts of domain names like 'site:www.abc.de' or 'site:abc.de' are valid search queries.

The mod searches for all links in Sphider's link-table but not in the stored keywords.

The search output has the same look and feel as usual in Sphider-plus search results.

Enabled search for dates like 2009-05-27, 27/05/2009 or 27.05.2009

Enabled suggestion also for search queries that containing upper case characters.

Automatically adapt Sphider's dialog to user language.

This mod detects the language of visitors client and selects the according language from Sphider's language folder. If not available, Sphider will use the language as defined in Admin settings.

Auto-detection may be enabled by checkbox in Admin settings

Show 'Most popular searches' table at the bottom of result pages.

Selectable in Admin settings, the most popular queries are presented on the bottom of each result page.

Count of rows for 'Most popular searches' is also to be defined in Admin settings.

Warning message if search string is only found in Url or <title> tag.

If the search string will be found only in title or Url, but not in the HTML body or meta tags, there is no short description for that Url with no possibility to highlight the search string.

A warning message will be displayed instead: "Search string was found only in page title or Url."

This mod is Admin selectable.

Index only new sites.

Additional item in Admin Sites submenu for bulk indexing of all the new sites that were added since last index/re-index.

Erase & Re-index.

Additional item in Admin Sites submenu that will clear the database and perform a re-index.

Clear database done before the re-index will leave the following untouched:

- Categories
- Query log
- Sites and all options: spider-depth, last indexed, can leave domain, title, description, url must include, url must not include.

Limit max. link count to be indexed for each Url.

In Admin settings the count of links to be followed per Url is selectable.

Will be followed by:

- Index
- Index only the new

Perform a link-check instead of re-index.

Selectable in Admin settings, a fast running link-check can be performed.

Unreachable links are automatically deleted from Sphiders database.

Define max. length of title presented in result pages.

An additional input field in Admin "Search Settings" is presented for Admin determination.

Dynamic adaptation of <title> and <h1> tags.

In order to create an individual title for the result pages,

a new input field in Admin settings 'Search Settings' is presented.

Additionally the result page <title> in HTML-header is provided with

- User defined title
- Category (if selected)
- Search query
- Page number of results

New Admin Sites Option menu design with additional utilities.

- Based on the XHTML valid Admin by Peter__LT

3 new template designs selectable in Admin settings

- Based on preparatory work by Peter__LT
- The template folder contain only those files that are responsible for the design

Additional Admin Sites submenu: List all pages that belong to the selected site.

To be found in Sites / Options / Pages a list is shown with:

- Page Url
- Last indexed date
- Page size

Validate all user input for security acceptance.

All entries are checked

- Delete quotes
- Place backslash in front of special characters
- Shell commands, XSS attacks and SQL injections are blocked

Additional .htaccess security file.

Prepared for:

- Prevent listing of folder content (files)
- Redirect client queries to search.php
- Prevent delivery of internal files

Sort Admin's Site table in alphabetic order.

Selectable by checkbox, the table is presented in alphabetic order or by index date.

Export all current Url's from Admin section.

A file 'url.txt' will be created with all existing Url's in folder .../admin/urls/

Import url.txt file from folder .../admin/urls/

The content of file 'url.txt' will be copied into Spider-plus database.

Existing Url's will be lost and overwritten.

Following rules are valid for the url.txt file:

- Url's must be in format: url|spider-depth|category
like:
http://www.abc.de
http://www.abc.de|2
http://www.abc.de|-1|Info
http://www.abc.de|3|Funny things
- Rows must be separated with 'LF'
- Url, spider-depth and category must be separated by. "|"
- If you don't specify spider-depth it is automatically set to '-1'.
- Also category is optional. If not specified the new site will be stored without category.
- Not specifying spider-depth but category requires: url||category-name

Delete Spider log.

Spider log files now can be deleted separately or as bulk delete.

Added in the submenus:

- Admin / Clean / Clean Spider log
- Admin / Statistics / Spider logs

Search in categories: Four bugs fixed.

The following items are modified for proper function of Spider's Category search:

- The 'Search' button now also sends the variable \$catid to the search script.
- Selecting 'Next' or the other page selections (on bottom of the result page), now transfers also the variable \$catid and \$category to the search script.
- The check boxes 'Search only in category . . .' and 'All sites' are no longer pre-selected.
So, once selected 'Search only in category . . .', you may now select search result page 2, 3, 'Next' and 'Previous' together with the category search.
- If 'Search only in category . . .' is selected, an additional headline is presented.
So the user is informed about the actual situation.

Database Backup and Restore. Bug fix by re-writing the complete Database Management.
Before backup, the 'Optimize Database' function is automatically performed.
Separated folders for each backup task.
Backups now are stored in individual files for each table.
Backup utility selectable for: 'Structure only' or 'structure plus data'
Unlimited file size for restore function is ensured.
Backup files compatible to phpMyAdmin.

Optimize Database. Bug fix by re-writing the complete Database Management.

Links that do not contain page name are now correctly followed (Bug fix by BenRosey)
Original Sphider does not except links like link text
Thanks to the bug fix of BenRosey Sphider-plus follows correctly.

Links that do not contain slash at the end of the Url are now correctly followed (Bug fix).
Original Sphider does not except links like: http://www.abc.de
Sphider-plus adds the required slash automatically like: http://www.abc.de/

Correct template selection for different css files in different template folders (Bug fix).

22.1.2 Version 1.0.a

Build up with Sphider v.1.3.4.b

Bug fixed in function validate_email

Involved files that have been modified / added for this release:
.../include/commonfuncs.php

22.1.3 Version 1.1

Build up with Sphider v.1.3.4.b

Included converters for indexing PDF, DOC, RTF, XLS and PPT files.
To be activated individually in Admin settings
Warning message during index process when deactivated file was found

Captcha protection for Submission Form '*Suggest a new Site*'.
Use of Captcha to be activated in Admin settings

Automatically adapt Sphider's dialog to user language.
Improved version by ^demon

Bug fixed in language depending user dialog.

Bug fixed in function check_robot_txt.

Involved files that have been modified / added for this release:
.../addurl.php
.../search.php
.../converter/ all files

.../converter/charsets/ all files
.../admin/configset.php
.../admin/ext.txt
.../admin/messages.php
.../admin/spiderfuncs.php
.../include/captcha.TTF
.../include/make_captcha.php
.../languages/ all files
.../settings/conf.php

22.1.4 Version 1.2

Build up with Sphider v.1.3.4.b

UTF-8 support for (nearly) all charsets.

- Selectable in Admin settings the translation into UTF-8 charset can be enabled.
- Index and search functionality for Unicode.
- Please notice the important information and details to be found in chapter: UTF-8 Support and 'Preferred Charset'

Individual preferred charset.

- Charset for result page can be defined in Admin settings.
- This option will be overwritten by the UTF-8 option.

Use of '*Default results per page*' (10, 20, 30, 50) also for Sites table in Admin section.

Use of '*Default results per page*' (10, 20, 30, 50) also for Link search (site:).

Included PHP version check before admin.php could be used.

Translated Danish language file.

- Thanks to Brian Jorgensen

Media files excluded from index/re-index procedure.

- Enlarged file list in .../admin/ext.txt

Improvements and bug fixes in:

- 'Admin settings' dialog
- 'Did you mean' option
- !strict search
- Converter for non-HTML files
- Site search (site:)
- Addurl suggest form

Involved files that have been modified / added for this release:

.../addurl.php
.../search.php
.../admin/admin.php
.../admin/admin_header.php
.../admin/auth.php
.../admin/configset.php
.../admin/db_main.php
.../admin/spider.php
.../admin/spiderfuncs.php
.../admin/ext.txt
.../converter/ConvertCharset.class.php
.../converter/charsets/ all files
.../include/searchfuncs.php

.../include/search_links.php
.../include/js_suggest/suggest.php
.../language/ all files
.../settings/conf.php

22.1.5 Version 1.3

Release date: March 31, 2008
Build up with Sphider v.1.3.4.b

Tolerant search

- Selectable in search-box like AND/OR/Phrase and as new item: *'Tolerant search'*
- Presents results that are 'like' the query as an integrated *'Did you mean'*
- Presents search results for queries with e=é=è=ê, ä=a, Ü=U etc.
- Results are independent whether the user enters e or é or ê in the search query

Clear Category table

- Additional item in Admin section *'Database & Log Cleaning Options'*
- Deletes all categories not associated with any valid site

Fixed charset to UTF-8 for User Suggestion Form (addurl).

Involved files that have been modified / added for this release:

.../addurl.php
.../search.php
.../admin/admin.php
.../admin/spider.php
.../include/searchfuncs.php
.../languages/ all files

22.1.6 Version 1.3.a

Build up with Sphider v.1.3.4.b

Individual *'Erase & Re-index'* function for single sites.

- Additional item in Admin sites submenu *'Manage Site Indexing Options'*
- *'Erase & Re-index'* functionality for selected site

Translated Spanish and Dutch language files.

- Thanks to Willy

Involved files that have been modified / added for this release:

.../admin/admin.php
.../languages/es-language.php
.../languages/nl-language.php

22.1.7 Version 1.4

Release date: May 28, 2008
Build up with Sphider v.1.3.4

In Admin settings the method of chronological order for result listing can be defined.

Results ordered by:

- Relevance (weight)
- Main URLs (domains) on top
- First URL names and then weight
- Only top 2 per URL

The mode of chronological order for result listing is shown as additional headline on top of the result pages.
To be activated in Admin settings.

Select method of highlighting for found keywords in result listing.

If 'Advanced search' is activated, the user may select:

- bold text
- marked yellow
- marked green
- marked blue

The default highlighting can be defined in Admin settings.

If in Admin settings the option '*Index words in Domain Name and URL path*' is activated, found keywords now are highlighted also in result listing (row URL).

If in users browser JavaScript is disabled, a warning message is displayed on top of the search form that full functionality of Sphider-plus will not be available (required for the suggest framework).

Improved printout for '*Show sites in category*'. If in Admins '*Site options*' the content for 'title' was not included, now title and short-description will be fetched from the HTML header (of the indexed sites). If also this information is not available, a warning message will be displayed.

Enable index and re-index for pages with duplicate content.

Additional item in Admin settings:

- If selected, pages with content that was already indexed by another page will also be indexed/re-indexed. A warning message together with the URL that also holds the duplicate content will be presented in spider log output.
- If not selected, the link (page) will be ignored. Never the less the message and URL info will be presented.

Improved function '*If available follow sitemap.xml*' in order to prevent '*Page is duplicate*' messages.

Improved printout if PDF files cause indexing problems.

If '*Follow sitemap.xml*' is activated and a valid sitemap was found, the log output

Links found: 0 - New links: 0

is no longer shown. Because all links are delivered from the sitemap file and new links are not searched during index / re-index.

An eventually non-existing log folder will be created automatically during index / re-index process. So, the message '*Logging option is set, but cannot open a file for logging.*' will be prevented.

If in Admin browser JavaScript is disabled, a warning message is displayed on top of Admin page that full functionality of Sphider-plus administration will not be available (required for warning messages).

Updated Romanian language file by CyBerNet.

Corrected Spanish language file by Willy.

Bug fixed in index / re-index function that caused problems to index words which consist only of upper case characters.

Bug fixed in index / re-index function that caused problems to index words containing the ' à ' character.

Some small improvements for result printout.

Length of words to be indexed is increased to 255 characters per word. For the required modification in Sphider-plus database, please notice the additional FAQ information ([Can't search for long words](#)) in the readme.pdf document. If not required, this item must not be installed. Functionality of Sphider-plus does not depend on this modification.

Involved files that have been modified / added for this release:

```
.../search.php
.../admin/admin_header.php
.../admin/auth.php
.../admin/configset.php
.../admin/install_all.php
.../admin/messages.php
.../admin/spider.php
.../admin/spiderfuncs.php
.../include/searchfuncs.php
.../include/search_links.php
.../languages/ all files
.../settings/conf.php
.../templates/all_folders/thisstyle.css
```

22.1.8 Version 1.5

Release date: July 14, 2008
Build up with Sphider v.1.3.4

Improved Suggest Framework. Now suggestions are presented also for queries with accented letters.

Enable real-time output of logging data. Selectable in Admin setting together with the update interval (1 - 10 seconds).

In order to prevent performance problems and memory overflow for large amount of URLs, Sphider-plus may clean resources during index / re-index. Selectable in Admin settings, this item periodical will:

- Free memory that is allocated to unused MySQL recourses.
- Unset PHP variables, which are no longer required.

Define max. length of URL presented in result pages.

An additional input field in Admin "Search Settings" is presented for Admin determination.

For 'Maximum length of page title displayed in search results' the title now will be broken at the end of the word exceeding the defined length. Not inside a word at the character count limit defined in Admin setting.

PDF converter for Linux Operating System included.

Needs to be individualized according to readme.pdf documentation, chapter ['PDF converter for Linux server'](#). Thanks to rasc.

Additional item in Admin section: Server Info

To be found in submenu 'Statistics', important information are presented for:

- Server
- Environment
- MySQL

- PDF converter
- php.ini file
- PHP integration

Enlarged Admin interface if database is empty.

Improved printout for database connection problems. Now MySQL error message is included.

Improved printout if text converter could not extract words from PDF, DOC, XLS etc. files.

Improved printout for Database Backup Management.

Modified installation script. Thanks to Flemp.

Font file renamed to captcha.tff (former: captcha.TTF). Thanks to ethix.

All style sheets now are centralized in ../templates/all_folders/thisstyle.css
Consequently the file ../include/js_suggest/SuggestFramework.css is no longer required.

Function 'create sitemap()' improved for XML conformity and moved from script ../admin/spider.php to script ../admin/spiderfuncs.php.

Bug fixed in 'Phrase Search' if UTF-8 support is not selected.

Bug fixed in highlighting of found keywords on result page.

Some small bug fixed for mysql queries.

Involved files that have been modified / added for this release:

```

../search.php
../admin/admin.php
../admin/admin_footer.php
../admin/configset.php
../admin/db_backup.php
../admin/db_main.php
../admin/install_all.php
../admin/install_reallog.php
../admin/install_sphider-plus.php
../admin/messages.php
../admin/real_get.php
../admin/real_log.php
../admin/real_ping.js
../admin/spider.php
../admin/spiderfuncs.php
../converter/pdftotext
../converter/pdftotext.script
../include/captcha.tff
../include/commonfuncs.php
../include/searchfuncs.php
../include/js_suggest/suggest.php
../settings/conf.php
../settings/database.php
../templates/all_folders/navdown.jpg
../templates/all_folders/thisstyle.css

```

Attention: Starting with version 1.5, Sphider-plus supports real-time output of logging info during index / re-index procedure. This item requires an additional table for the database. If you update from a former version of Sphider-plus, please run the ../admin/install_reallog.php script. If you upgrade from original Sphider or install from scratch, you don't need to run this script. Its features are also included in the other installation scripts.

22.1.9 Version 1.6

Release date: September 06, 2008
Build up with Sphider: v.1.3.4

Additional item in Admin settings to select:

- Instead of weighting %, show count of query hits in full text.
Selecting this item will also influence the order of result listing. Now only the number of keyword hits in full text will define the position of a page in result listing.

Additional item in Admin settings to select the chronological order of result listing:

- 'Most Popular Links ' on top.
Activating this item, Sphider-plus will present the result listing in order of before learned link attractivity. Defined as those links with the best user acceptance (clicks).

Additional items in Statistics overview:

- Queries total
- Link clicks total

Additional item in Admin / Statistics:

- Most Popular Links.
Presenting the quantity of clicks individual for each link with date and time of last click.
Also the latest query before clicking that link is presented.

Additional item in Admin / Clean:

- Clear 'Most Popular Links' log.

Additional item for re-index procedure:

- Temporary ignore 'robots.txt'.

If utf-8 support is activated, result listing now is independent for queries with upper- or lowercase letters. Or alternatively, if selected in Admin settings, distinct results for case sensitive queries could be performed.

Improved utf-8 support for non-Latin characters.

Improved suggest framework for utf-8 support. Now offering suggestions

- for phrases
- for accented letters
- for non-Latin characters

Known issue: Well working for Firefox and Opera browser, for non-Latin characters IE is not cooperative. Need to rewrite the Suggest Framework completely for a browser independent presentation of the suggestions.

Improved search functionality for queries with accent letters without selecting the utf-8 support.

Phrase search improved, so that common words and too short (min_word_length) words could be used as part of the query phrase and are no longer marked as ignored.

Improved functionality for 'Most popular searches'. Now also

- Advanced search settings
- Categories
- Mode of highlighting
- Results per page

will be taken into account when clicking a 'Most popular searches' suggestion.

Bug fixed that seduced Sphider to follow links that are placed in HTML comments.

Bug fixed that created a wrong weighting calculation for keywords placed

- behind a word that did not match 'min_word_length'
- behind a 'common' word
- first found in full text

Bug fixed in 'Strict search' that caused invalid highlighting in result listing.

Involved files that have been modified / added for this release:

```
.../search.php
.../admin/admin.php
.../admin/configset.php
.../admin/install_all.php
.../admin/install_bestclick.php
.../admin/install_sphider-plus.php
.../admin/spider.php
.../admin/spiderfuncs.php
.../include/click_counter.php
.../include/commonfuncs.php
.../include/searchfuncs.php
.../include/js_suggest/suggest.php
.../include/js_suggest/SuggestFramework.js
.../languages/ all files
.../settings/conf.php
```

Attention: Starting with version 1.6, Sphider-plus supports logging of 'Most popular links'. This item requires additional rows in 'links' table of the database. If you update from a former version of Sphider-plus, please run the .../admin/install_bestclick.php script. If you upgrade from original Sphider or install from scratch, you don't need to run this script. Its features are also included in the other installation scripts.

22.1.10 Version 1.7

Release date: November 20, 2008
Build up with Sphider: v.1.3.4

New item in Admin / Settings / General Settings:

- Enable Debug mode.
If selected, during index / re-index procedure the following information will be presented individual for each page:
 - New links found here
 - New keywords found here
- For more details, please notice chapter [Error messages and Debug mode](#)

New item in Admin / Settings / General Settings:

- Enable / Disable MySQL and PHP error messages.
It is recommended to disable the output of these messages for production systems, as they could reveal sensitive information.
For more details, please notice chapter [Error messages and Debug mode](#)

New item in Admin / Statistics / Server Info:

- PHP security Info.
Some basic info about current server configuration, presenting the security information status of the PHP environment.

Completely rewritten Suggest framework. Based on 'script.aculo.us' and 'prototype' scripts, now suggestions for non-Latin symbol and accent characters are also presented in IE browser. Additional items in Admin settings:

- Define minimum count of query letters in order to get a suggestion.
- Show / Hide the amount of found keywords in suggestion table.

New capability to prepare language specific common files.

If multilingual sites, or sites with different languages, are to be indexed, this feature improves overview. Common words to be ignored during index / re-index procedure can be placed in individual files. The common word files should not be used, if 'phrase search' is the standard type of search. Sphider-plus will become problems to find complete phrases. Therefore, in Admin settings the use of the common word files may be activated / deactivated by a checkbox. For more details, please notice chapter [Ignored words](#)

New feature: Use a blacklist.

If the content of a page to be indexed / re-indexed contains one word of the blacklist, it will not be indexed / re-indexed. To be activated / deactivated in Admin settings
For more details, please notice chapter [Use of Blacklist](#)

New feature: Use a whitelist.

The content of a page to be indexed / re-indexed must contain at minimum one word of the whitelist to be indexed / re-indexed. To be activated / deactivated in Admin settings
For more details, please notice chapter [Use of Whitelist](#)

New feature: If available, show multiple hits of search result (per page) in result listing.
To be defined (1 - 9) in Admin / Search Settings.

Improved URL import / export function:

- The names of URL files now are including date and timestamp of export procedure.
- This enables the Admin to import selected URL files.
- Also a file individual delete function was included.
- Delimiter in URL file changed from "," to "|". As suggested by Ranbir.

Improved Admin / Settings section:

- Included directory with links to the different Setting blocks.

New item in Admin / Settings section:

- Backup current configuration settings. Individual files are created with date and timestamp.
- Restore configuration settings from former created backup file.
- Individual delete of backup files.
- Delete protected backup file that holds the default settings.

New item in Admin / Settings / Spider settings:

- Use a unique name (sitemap.xml) for all created sitemap files.
Could be selected, if only one single Site is to be indexed.
To be used in conjunction with selecting the destination folder for the sitemap files.
../ is the root folder of the Spider-plus installation.

If the charset of a page to be indexed / re-indexed is not detectable, the home charset as defined in Admin settings is used.

Improved search function for non-Latin symbols.

Search function enabled for queries containing an apostrophe.

Bug fixed in index/re-index procedure that prevented indexing of last word in full text that should be stored as new keyword.

Improved storage of keywords in index/re-index procedure.

Updated Romanian language file, thanks to CyBerNet.

Some file types added to exclusion list in order not to be indexed / re-indexed. Thanks to clubmaster3.

Improved Admin Log-in for Microsoft IIS. Thanks to bobyn.

Involved files that have been modified / added for this release:

```
.../addurl.php
.../search.php
.../admin/admin.php
.../admin/admin_header.php
.../admin/auth.php
.../admin/configset.php
.../admin/confirm.js
.../admin/dbase.js      (file no longer required)
.../admin/db_backup.php
.../admin/db_main.php
.../admin/ext.txt
.../admin/messages.php
.../admin/real_get.php
.../admin/real_log.php
.../admin/spider.php
.../admin/spiderfuncs.php
.../admin/url_manage.php
.../admin/phpSecInfo/ (all files)
.../converter/ConvertCharset.Class.php
.../include/categoryfuncs.php
.../include/commonfuncs.php
.../include/searchfuncs.php
.../include/search_links.php
.../include/suggest.php
.../include/ajax/      (all files)
.../include/common/    (all files)
.../include/js_suggest/ (folder no longer required)
.../languages/ro-language.php
```


.../settings/conf.php
.../templates/all folder/thisstyle.css

22.1.11 Version 1.7a

Release date: November 27, 2008

New item in Admin / Settings / Spider Settings

Delete special characters like dots, commas, quotes, exclamation and question marks etc. as part of words.

If activated, only the 'pure' words are indexed. Secondary characters before and at the end of words are deleted.

For more details, please notice chapter [Delete secondary characters](#)

Improved behaviour if charset of page to be indexed can't be detected.

Bug fixed that prevented correct link to search result.

Additional translation table to convert upper to lower case characters for Cyrillic charset.

Updated Russian language file, thanks to vipraskrutka.

22.1.12 Version 1.8

Release date:	February 26, 2009
Build up with Sphider:	v.1.3.4
Sphider-plus vs. original Sphider:	124 items worked out

In front of Sphider-plus version 1.7a the following items have been added / modified:

New feature: Search for media content. If activated in Admin settings, media files like

- Images
- Audios
- Videos

will be indexed and become searchable. Result listing is separated into 4 sections: found text, found images, found audio streams and found videos. Thumbnails are presented for the image results. All media results are linked to the source, so that the files could be opened with the appropriate media player.

As also ID3 and EXIF data is indexed, it is possible not only to query for a media title, part of a title or suffix, but also to search for e.g. all songs of a specific author, or for all images done with 'f/2.0' or perhaps flash setting 'red-eye'.

For more details, please notice chapter [Media Search](#)

New feature: Index RSS and Atom feeds.

If activated in Admin settings, RSS (v.0.93 - v.2.0) and Atom (v.1.0) feeds are indexed and the content becomes searchable. For more details, please notice chapter [RSS and Atom feeds](#)

New feature: Result cache for text and media queries. If activated in Admin settings, this item offers:

- Extremely reduced response time for queries already cached.
- Controller to keep the 'Most Popular Queries' always in cache.
- Separate caches for text and media results, configurable in Admin settings.
- Automatic cleaning of caches during 'Erase & Re-index' procedure.
- If Debug mode is enabled, activity/status of cache is presented in result listing.

For more details, please notice chapter [Result cache for text and media queries](#)

Enlarged Admin statistics. In table 'Search Log' the following items are additionally presented:

- User IP
- Users country code
- Users hostname.

New item in Admin Settings (Section: Index Log Settings):

Suppress browser output of logging data during index / re-index.

This item will speed up index / re-index procedure and prevent browser overflow on huge amount of sites to be indexed.

If activated, this setting also disables the real-time output of logging data.

New feature: Use the blacklist to reject queries.

If the query input contains a word of the blacklist, the complete query will be deleted.

To be activated in Admin settings.

For more details, please notice chapter [Use of Blacklist](#)

If 'Convert all to UTF-8' is activated, the files

common_xyz.txt

whitelist.txt

blacklist.txt

are also converted. This is performed always when the script is started, so that this transformation is valid only for the current session.

If 'Enable distinct results for upper- and lower-case queries' is not selected in Admin settings, the words placed in

common_xyz.txt

whitelist.txt

blacklist.txt

are converted to lower case characters, so they will match independent of their spelling in the .txt files. This is performed always when the script is started, so that this adaptation is valid only for the current session.

New feature in Admin statistics. In table 'Image functions', details about the installed GD library as part of the PHP environment will be presented.

New feature in Admin 'Clean' section:

Clean text and media cache (separate items). Additionally count of results in cache and currently used memory space are presented.

The status of last search request (done 'in category xyz only' or in 'All sites') is cached for next query input.

Improved Log output if file mode is set to 'text'.

Additional common file for French language. Thanks to Florian Vugier.

Updated French language file. Thanks to Manuel Pardo, Florian Vugier and Marie-Cécile.

Updated Portuguese language file. Thanks to Júnio Branco.

Involved files that have been modified / added for this release:

- .../addurl.php
- .../php.ini
- .../search.php

- .../admin/admin.php
- .../admin/admin_header.php
- .../admin/configset.php
- .../admin/db_main.php
- .../admin/Geolp.dat
- .../admin/geoip.php
- .../admin/index_media.php
- .../admin/install_all.php
- .../admin/install_sphider-plus.php
- .../admin/install_v.1.8.php
- .../admin/messages.php
- .../admin/php.ini
- .../admin/spider.php
- .../admin/spiderfuncs.php
- .../admin/thumbs/ (new empty folder)

- .../converter/rss2html.php
- .../converter/rss.html
- .../converter/rss_parser.php

- .../include/commonfuncs.php
- .../include/searchfuncs.php
- .../include/search_media.php
- .../include/media_counter.php
- .../include/search_links.php
- .../include/search_media.php
- .../include/common/audio.txt
- .../include/common/image.txt
- .../include/common/suffix.txt
- .../include/common/video.txt
- .../include/images/ all files
- .../include/mediacache/ (new empty folder)
- .../include/textcache/ (new empty folder)

- .../languages/ all files

Attention: This release requires additional database tables and additional table rows in already existing tables. If you update from a former version of Sphider-plus, please run the .../admin/install_v.1.8.php script. If you upgrade from original Sphider or install from scratch, you don't need to run this script. Its features are also included in the other installation scripts.

22.1.13 Version 1.9

Release date: not published; only internal developing version.

22.2 Version 2.0

Release date:	May 27, 2009
Build up with Sphider:	v.1.3.4
Sphider-plus vs. original Sphider:	141 items worked out

In front of Sphider-plus version 1.9 the following items have been added / modified:

Multiple database support for up to 5 independent databases (expandable).

Individual activation of one database for:

- Admin
- Search user
- Suggest URL

For more details, please notice chapter [Multiple database support](#)

Independent configuration and activation for each database is integrated into the Admin interface.

Additional password protected access permission for database configuration, independent from Admin login.

Integrated availability check for all databases and their release relevant table structure.

Individual for each database:

- Backup and restore
- Copy / Move from each database to each other database

32 MByte query cache for MySQL database.

- To be activated in Admin settings.
- Status of cache is observable in Admin / Statistics / Server-Info / MySQL.
(Cache might not work for 'Shared Hosting' applications)

Obey the<link> tag specification:

rel="canonical"

If defined in page header of a website, the crawler will be redirected to the canonical link and Sphider-plus will understand that the duplicates all refer to the canonical URL.

For more details, please notice chapter [Canonical <link> tag](#)

Index websites that are created with ASP.NET

Definition for path to PDF converter integrated into Admin Settings interface.

Additionally the default setting - as required for the Operating System environment - is suggested.

If path to PDF converter is invalid and converter is not accessible, an error message (in Admin Settings dialog) is created.

Additional Admin setting to enable optionally indexing of external hosted media content.

Improved index procedure of media files, by avoiding indexing of duplicate media content.

Improved image indexing by reducing the required download time.

Improved index / re-index procedure to avoid 'MySQL server has gone away' messages.

prototype.js framework adapted to cooperate with XHTML valid parameter handling.

XHTML1.0 output for

- Admin interface
- Search form and Result listing
- Suggest URL form

Improved vulnerability check of User input and Admin log-in:

- Prevent buffer overflow errors.
- Suppress JavaScript execution and tag inclusions masked as XSS attacks.
- Prevent C-function 'format-string' vulnerability.

URL Suggestion Form includes character counter for remaining input in 'title' and 'description' field.

For 'Search with wildcards' now the complete word is highlighted in result listing. Not only the query part of the found keyword.

Phrase search is enabled now also for title tag, not only for full text.

Improved suggest framework: for search in categories, the suggestions now will be presented with respect to the pre-selected category.

Additional Admin setting in section 'Suggest Options':

For 'Media search' get suggestions also from EXIF info and ID3 tags

Files for database setting and script configuration are protected now against direct client access by pre-defining a named constant.

Updated Swedish language file. Thanks to Holger Gremminger.

Bug fixed in 'Search for suggestions in query log', which prevented to disable this option

Bug fixed that caused multiple listing of the same result, when
"Define maximum count of result hits per page, displayed in
search results (if multiple occurrence is available on a page)"
was activated.

Involved files that have been modified / added for this release:

Nearly all scripts.

Attention: This release requires a fresh installation of all scripts and a blank MySQL database. An update from former Sphider-plus versions or an upgrade from original Sphider is not foreseen.
For more details, please notice the chapter [Installation of Sphider-plus version 2.0](#)