Sphider-plus manual

Content

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

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.

In front of original Sphider several 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.

Sphider-plus offers a wide range of customizing the index and search procedures. By means of an Admin backend, all settings are presented. As stated above, this search engine uses some PHP libraries and extensions. When opening the Setting interface, the existence off these libraries are tested by software, and in case that a library is not part of the server environment, the according option is not presented in the Settings interface. For example, if the 'rar' extension is not available, it will not be possible to index RAR archives and the belonging check-box will not be presented in 'Spider Settings'. In order to check the availability of all required libraries and extensions, the Debug mode will present the corresponding messages.

Sphider-plus does not contain a JavaScript engine. Consequently all content created in real-time, while loading the page, will not be indexed.

Indexing with a search engine like Sphider-plus is very problematic on a 'Shared Hosting' server. Indexing huge amount of links might be interrupted, because the granted time slice can be finished before index procedure is finished. Especially if you intend to index not only text, but also media content like images, as well as audio and video streams. Sphider-plus tries 3 times to reconnect to the database. But if the server canceled the script, it will become necessary to manually invoke again the index procedure to continue. Sphider-plus will remember the last indexed link and continue the suspended process.

Some special functions like e.g. 'cyclical indexing' in any case will fail on 'Shared Hosting' server.

2.   Version and legal info

Name: Sphider-plus
Version: 3.2016b
Created: 2016-03-22

Based on original Sphider version 1.3.5 released 2009-12-13 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
Original Sphider GNU GPL license by Ando Saabas   ando(a t)cs.ioc.ee

Updates and support for Sphider-plus are available at: http://www.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 version 3.0 – 3.2016b

3.1   Preconditions

Sphider-plus requires *PHP 5.3 – 7.0* with installed *GD, mbstring, PECL* and *zlib* libraries. Additionally, if RAR compressed files should be indexed, the rar extension is required. Also a MySQL (version 5.3+) database must be available.
The .DOC, .RTF, and the .PTT converter supplied together with the Sphider-plus scripts are not available for LINUX/UNIX systems.

The following settings should be adjusted on the server
PHP safe_mode:	Off
Webserver mod_rewrite:	On
Display_errors:	Off
allow_url_fopen	On
memory_limit	128M

Additional note: The .htaccess files supplied with Sphider-plus may cause problems on some servers and might need to be disabled by renaming them.

3.2   New installation

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

1. Unzip the downloaded file, and upload all folders and files to the server, for example to:
        C:\programs\xampp\htdocs\public\sphider-plus\
Admin scripts need the privilege to write to the admin sub-folders (all sub-folders of .../admin/).

2. Create at minimum one database as part of the MySQL server, which will hold the Sphider-plus data tables. Collation of the database must be defined to   **utf8_bin** already during creation. In order to get full UNICODE support, the collation must be set to **utf8mb4_bin**, which is only available for MySQL servers version 5.5.3+

Creation of this database needs to be done outside of the Sphider-plus scripts, before entering into the admin backend the first time. For example with a tool like phpMyAdmin, PLESK, or something similar.

During database creation you already define

        - Name of database
        - User name
        - Password
        - Database host

which will be required later on in step 5 of the installation process. It is obligatory to define passwords for the database. Otherwise Sphider-plus will not accept the database. Additionally take care that your password should not contain special characters, because some older MySQL versions do not process them. Sometimes the 'Database host' is also called 'Server' or 'Database server', like in the tool phpMyAdmin.

3. Open the Admin interface with your browser by addressing the Admin with something like:
        http://localhost/public/sphider/admin/admin.php

At first you will be asked to enter a user-name and password. Use '**admin**' as user name and also as password.

After log-in, the Admin interface will be presented. On first log-in, there will be several warning messages, because no database is allocated to Sphider-plus. By means of a self test performed each time the Admin is called, also write permission to sub folders, converter availability, etc. are checked. Eventually it might become necessary to follow some warning messages like:

   chmod 777 the folder …/admin/tmp/

4. Open the sub-menu 'Database' and select 'Configure'. Now you will need to log-in for the database management. Use '**admin**' as user name and password. Afterwards, only the menu selection will be presented again. Select 'Configure' again to enter.

5. Entering the first time into this section, there will be again several warning messages.
At minimum one database needs to be declared by entering
   Name of database
   User name
   Password
   Database host
   Prefix for tables
as defined before in step2 externally, when you created the database. Additionally you need to add a free selectable name as table_prefix. This table prefix is obligatory. A blank character as prefix is not enough.

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. Now you should get the first (green) message like

  Database 1 settings are okay.
  Tables are not installed for database 1

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 first database is allocated and the tables are installed, the message ' **Database x settings are okay**.' is displayed in the 'Config' menu overview; showing separately the situation for each of the five databases. If the Sphider-plus application should work with only one database, the settings for the non-required databases 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](#)

6. If only one database should be activated, this step 6 could be skipped.
Next step to get Sphider-plus to work will be the activation of the database. There are three settings available in the 'Activate / Disable' menu:
   - 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.

7. 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. A site URL could be entered by selecting one of the three possibilities:
- Add site
- Import URL list
- Index

At the bottom of 'Sites' view a database overview is presented like:

Database 1 with table prefix 'search1_' contains:

2 sites **+** 9 page links **+** 0 categories **+** 417 keywords + 13 media links

8 As final step of the installation, modify user-name and password for Admin log-in, as well as for the database 'Configuration' menu. In order to do so, open the menu 'Settings' and find the sub menu 'Authentication for admin and database configuration'. First, you will be asked for the currently valid values, which are all set to '**admin**' as per default download of the scripts. Afterwards you may define new values for your individual usage.

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

3.3 Updating from version 1 and 2

No longer supported.

3.4 Updating from 3.x to 3.y

3.2014a In order to update to this version please follow the following short instruction:

Close all Admin windows of your Sphider-plus installation and copy the scripts

> …/admin/settings/authentication.php
> …/converter/pdftotext
> …/settings/database.php
> …/templates/your_individual_template_folder

to any folder outside of your Sphider-plus installation.

Now copy all the new Sphider-plus scripts as per download over the existing scripts of your former v.3 version and then restore your saved files back into the new Sphider-plus installation. So you will keep the access to your database, the Admin access authorization, your individual template, etc.

Afterwards **obligatory** enter into the Admin backend => Settings menu and press any of the 'Save' buttons. So you will get access to the new options of this version.

Now you will have to personalize the settings for your individual requirement again.

3.2014b Please follow the instructions as described above for 3.2014a

3.2014c Please follow the instructions as described above for 3.2014a
and   Additionally keep in mind that user-name and password for Admin access
3.2015a as well as for the database configuration, now are stored as hashed values.
and   Modification of the default values (as delivered by script download)
3.2015b need to be done as part of the admin backend.
       For details, please notice step 8 of chapter 3.2

3.2015c Please follow the instructions as described above for 3.2014a
       But in any case use all scripts as per download for your installation,
       because nearly all script were modified for this version.

3.2015d Please follow the instructions as described above for 3.2014a

3.2015e Please follow the instructions as described above for 3.2014a

3.2016a Because database collation and connector had been modified for this version,
       please perform a fresh installation like described in chapter '[New Installation](#)'.

3.2016b Please follow the instructions as described above for 3.2014a

4.   Settings and customizing

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

- Sites
Add Site
Index only the new
Re-index all
Re-index only preferred URLs
Erase & Re-index (available also for individual URLs)
Import/export URL list
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
Clear IDS log file
Clear flood attempt log file
Clear all entries in addurl or banned table
Truncate all tables in database

- Settings
General definitions
Admin options
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 with unlimited number of table sets
Activate separately for 'Admin', 'Search' user and 'Suggest URL user'
Backup / Restore
Copy / Move
Optimize

- Templates

In order to enable customer's integration of Sphider-plus into existing sites, HTML templates are prepared for

Search form
Text result listing
Media result listing
Most popular queries
etc.

Three different designs are offered, which may be selected in sub-menu 'Settings'. If the layout does not fit the design of your site (which is normal), you may create new designs and modify the appropriate file

.../templates/My_template/adminstyle.css
    .../templates/My_template/userstyle.css

- Statistics

- Top keywords (Top 50 with hit counter)
- All indexed thumbnails with ID3 and EXIF info
- Larges pages offering link URL 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)
- Spider 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 presenting lists of details.
- IDS log file offering:
  IP, host, query, impact, involved tags, date and time of intrusion.
- Flood attempts log file offering:
  IP, query, date and time of flood attempt.

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

As stated in Introduction, this search engine uses some PHP libraries and extensions. When opening the Settings interface, the existence of these libraries are tested by software, and in case that a library is not part of the server environment, the according option is not presented in the Settings interface. For example, if the 'rar' extension is not available, it will not be possible to index RAR archives and the belonging check-box will not be presented in 'Spider Settings'. In order to check the availability of all required libraries and extensions, the Debug mode will present the corresponding messages.

5. Indexing

5.1   Various options

As part of the Admin Site settings you may select several options to influence the index procedure.

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 an .htaccess file is used for redirect directives.

Must include / must not include: See below for an explanation.

Multithreaded indexing: See below for an explanation

Follow Sitemap file. See below for details.

Create Sitemap file. See below for details.

Word stemming. See below for details.

Etc.

5.2 Allow other hosts in same domain

This Admin selectable option allows indexing other hosts with the same domain name and it also ignores TLD, SLD and www.

If e.g. calling from http://www.sphider-plus.eu links like:

- http://sphider-plus.eu	without www.
- http://www.info/sphider-plus.eu	additional sub-domain
- http://www.sphider-plus.com	different TDL
- http://www.sphider-plus.tec.eu	additional SLD

will be followed if this option is activated in Admin settings.

There are 2 different options available in Admin setting to cover this feature. The first one is following all links found during index procedure. The second one is only following the links to other hosts, if the found links are redirected.

5.3 Word stemming

Sphider-plus is offering language specific stemming algorithms for 15 languages.

Bulgarian, Chinese, Czech, Dutch, English, Finnish, French, German, Greek, Hungarian, Italian, Portuguese, Russian, Spanish and Swedish

To be activated individually for the language that needs to be indexed. Automatically the according common word list (holding the stop words not to be stored in database) will be activated together with the stemming language. For Chinese, Greek and Russian, additionally the corresponding language support is automatically activated. These additional features remain activated, even if word stemming later on is reset to 'none' and need to be deselected manually.

On the other hand, if activated for indexing, the stemming selection must remain activated, because also the query input must be stemmed. As during the index procedure only the etymons are stored in database, this will create results independent whether the query 'walk', 'walks', walked' or 'walking' is entered (for English stemming).

5.4 Periodical Re-indexing

This mode offers automatically Re-indexing of all sites, or site specific, started periodically at each defined time interval. In Admin backend the time interval is selectable for
        3 hours, 12 hours, 1 day, 1 week or 1 month.
Also the count of periodically performed re-indexing procedures could be defined in Admin backend.

Once started, the periodical Re-index procedures will silently work in the background without creating monitor output, but like in all other indexing modes, writing the index results into log files. Additionally a log file showing the status of the periodical indexer is created, presenting all dates and times when the Re-index procedures were started, as well as the count. This additional log file is available for the Admin in 'Statistics' menu called 'Auto Re-index log file'.

The periodical Re-indexer for all sites could be started and aborted in Admin backend, by selecting the 'Periodical Re-index' sub-menu in 'Sites' view.

Instead for site individual Re-indexing, the periodical Re-indexer could be started and aborted in the "Options" menu of each site.

5.5 Preferred indexing

Each new URL added to the Admin backend, could be supplied with a priority level. This level will be used by the option 'Re-index only preferred sites'. Level 1 will be interpreted as most important, while level 5 should be used for non-preferred sites. In case that new URLs are not manually supplied with any priority, the level will automatically set to '1'.

While invoking the option 'Re-index only preferred sites', the admin may select a suitable level for the next index procedure. Thus, only those URLs, containing the according level, will be re-indexed.

5.6 Multi-threaded indexing

The Admin setting: Define number of threads allowed for index procedures (max. 10)
activates parallel indexing. For multiple site indexing, this option will speed up the procedure significant. If this option is activated, browser output of logging data, as well as real-time output in a second browser window (tab), is suppressed. Never the less all index results will be stored in log files in sub folder …/admin/log/
The names of the log files look like:

        db2_100524-21.47.56_1.html   (log file of first thread)

        db2_100524-21.48.12_2.html   (log file of second thread)

The names are build by the following items:

        - db2   Number of database.

        - 100524   Date (May 24, 2010)

        - 21.47.56   Time when this thread was started (hours.minutes.second).

        - 1   ID-number, which will be incremented by each thread.

                           If multithreaded indexing is not activated, the ID will be set to '0'.

The individual threads will be activated by means of the Admin dialogue. For example, if 'Erase & Re-index' is selected, after the 'Erasing' dialogue, the threads could be started in sequencing order. It is not necessary

to invoke all predefined threads. The dialogue (browser) window will always present the last started thread. It is strongly recommended not to close this window or to use the 'Return' button of the browser. If the thread finished indexing, a 'Ready' message will be shown and the 'Back to admin' button is presented. Never the less, a former started thread still might be busy to index another site. To be seen in Admin 'Sites' view by the 'Unfinished' message at the corresponding site. Refreshing the 'Sites' window offers the successful end for all threads; by replacing the 'Unfinished' message with the date of last index.

During multi-threaded indexing browser options like caching, pre-fetching and turbo modes should be disabled.
Multi-threaded indexing for command line operation is presented below in chapter Multi-threaded indexing

5.7   Create thumbnails as a webshot during index procedure

This Admin selectable option is supplied by the web service http://plsein.in/ The web shots will be presented as thumbnails with the size of 130 X 174 pixels as part of the text result listing, individual for each result. It should be taken into account that this option will slow down the index procedure significantly. As delivered by a web service, this option is not available for Intranet and all 'localhost' applications and requires an individual "secret-code" and "key", which are available on request at *PLS-E.*

In 'Sites' view of the Sphider-plus admin backend you may check ability of this web service. Also required time (seconds) to create one thumbnail is presented. So you may decide, whether you like to use this web service for all pages to be indexed.

5.8   Follow Sitemap file

To be activated in Admin settings, Sphider-plus will use the links found in sitemap.xml or sitemap.xml.gz files. This significantly increases the speed for index and re-index, because the links will not have to be searched in text part of each page.
This option will also force Sphider-plus to re-index only links that are:
- New and not yet known in Sphider's link table
- Links with a 'last modified' date, which is newer than Sphider's 'last indexed' date in database.

Sitemaps are always expected in the root folder of the site to be indexed and must be named
      sitemap.xml   or   sitemap.xml.gz

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.

If <sitemapindex . . > is detected in a sitemap.xml file, and if multiple sitemap files are available, Sphider-plus will process the secondary Sitemaps and extract all links for index / re-index. Also gzip-compressed files (Index Sitemap files as well as the sitemap files) will be processed, independent of their file suffix.
A sitemap index file can only specify sitemaps that are found on the same site as the sitemap index file. For example, http://www.yoursite.com/sitemap_index.xml can include Sitemaps on http://www.yoursite.com but not on http://www.example.com or http://yourhost.yoursite.com. As with Sitemaps, the Sitemap index file must be UTF-8 encoded.

For individual sitemaps with different names and/or sitemaps that are stored in sub-folders, Sphider-plus offers the option of defining their URL and name in 'Add site' menu, as well as in 'Edit site' menu. Never the less, links in these individual sitemaps need to follow the rules as defined at http://www.sitemaps.org/ and are always treated as absolute links and must be from a single host. RSS (Real Simple Syndication) 2.0 or Atom 0.3 or 1.0 feed sitemaps are currently not supported by Sphider-plus. Also extensions of the sitemaps protocol like creating an own name-space <! – namespace extension -- > are not supported by Sphider-plus.

5.9   *Create sitemap file*

Create a sitemap during index/re-index could be activated in Admin settings. This option offers the following features:
   - 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.
   - Additional option: 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.

6. Using the indexer from command line

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

6.1   Overview and options

It is possible to spider web pages from the command line, using the syntax:
        php spider.php <options>

where <options> are:

-all	Re index everything in the database
-eall	Erase database and afterwards re-index all
-new	Index all new URLs in database which had not yet been indexed
-erase	Erase content of database
-erased	Index all meanwhile erased sites
-preferred <level>	Index with respect to preference level
-preall	Set 'Last indexed' date and time to 0000
-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 re-index the same URL, use:
        php spider.php -u http://www.domain.com/test.html -r

6.2   Multi-threaded indexing

For command line operation parallel indexing has no restrictions for the count of threads. Just limited by the server resources. Parallel indexing is enabled for several different methods as described below.

6.3 Index only the new

Index all new URLs in database which had not yet been indexed <-new>
Simply start several threads and add individual IDs to the option parameter like

php spider.php –new1

php spider.php –new2

etc.

The IDs will be added to the name of the corresponding log files like:

db2_100524-21.47.56_ID1.html   (log file of first thread)
db2_100524-21.48.12_ID2.html   (log file of second thread)

IDs could be defined by personal requirements, but the limitations for file names with respect to the OS should be taken into consideration. There is no auto-increment of IDs like for multi indexing that is initialized by the Admin dialogue.
If IDs are not added, it is obligatory to delay the start of each thread for 1 second. The names of the log files are created with a resolution of one second. If started too early, several threads will write into one log file. All unsynchronized, the resulting log file will be unreadable.

6.4 Reindex all

To be invoked by once preparing the database with the command
php spider.php –preall
This will reset all 'Last indexed' tables to '0000', but will not erase the content of all the other tables. So the check whether the content of a page has changed (MD5sum) is still available for a fast re-index procedure.

Once prepared, multi-threaded re-indexing could be invoked by starting several threads and adding individual IDs to the option parameter like:

php spider.php –erased1
php spider.php –erased2
etc.

The IDs will be added to the names of the log files as described above in Index only the new

6.5 Index erased sites

Index all meanwhile erased sites <-erased> will index only those sites that had been individually or bulk erased. Multi-threaded indexing could be invoked by starting several threads and adding individual IDs to the option parameter like:

php spider.php –erased1
php spider.php –erased2
etc.

The IDs will be added to the names of the log files as described above in Index only the new

7. **Keeping pages, words and files from being indexed**

7.1 *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.

This directive could be temporary overwritten site specific for the next index procedure by the advanced option:
> Temporary ignore 'robots.txt'

7.2 *URL 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 new line (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. But only a string starting with a '*' in front is considered to be a regular expression, so that '*/[a]+/' denotes a string with one or more a in it.

In case that all new sites, which are added to Sphider-plus, should contain a 'URL Must include' or 'URL must Not include' rule, the strings could be placed into the files:
> …/include/common/must_include.txt
> …/include/common/must_not_include.txt

While calling 'Add site' in Admin menu, the content of these files are transferred into the corresponding option fields of the new site. They could be edited afterwards by calling the site options.

As per default Sphider-plus download, the two files are empty. So there is no problem to leave the corresponding check-box activated, even if no rule should be used for the new site. On the other hand by deselecting the check-box:
> Use default values as defined in common files must_include.txt and must_not_include.txt

the admin may prevent the transfer of the predefined rules for individual sites.

There is an additional Admin setting in section "General Settings" called:
> Use string list of 'URL Must Not include' also to prevent erasing of involved URLs

If activated, also erasing of the involved sites and pages (links) will be prevented. In order to erase all sites and all pages completely, it might become necessary to uncheck this option.

7.3   *Ignoring links*

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

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

This directive could be temporary overwritten site specific for the next index procedure by the advanced option
> Temporary ignore 'nofollow' directive

7.4   *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.

7.5   *Ignoring parts of a page by <!--sphider_noindex--> . . . . <!--/sphider_noindex-->*

Sphider-plus includes the feature to exclude parts of pages from being indexed. This may 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.

7.6   *Ignoring parts of a page defined by <div id='abc'> or <div class='abc'>*

Ignoring parts of a page by the <!--sphider_noindex--> tags requires direct access to the page, because the tags need to be added (edited) to the page.

A more flexible method, which does not require direct access, is enabled by the Admin setting:
'Use list of div ids or classes to ignore the complete div content during index/re-index'

If enabled in Admin settings, the values as defined in the list-file …/include/common/divs_not.txt will be used to delete the content between <div id='abc'> and </div> . Never the less links inside of the div tags will be followed. The values in the list-file will not be interpreted case sensitive. Values in this common list may end with a wildcard, so that 'menu*' will work for ids like
menu1, menu2, menu_left, etc.

Multiple and nested divs will be attended. Alternately this option could also be used for <div class='abc'>

For even more flexability, the file …/include/common/divs_not.txt may alternately contain a regexp pattern. The regexp needs to be introduced by */ and must be ended with another slash.
Example: */menu[0-5]/

7.7   *Indexing only parts of a page defined by <div id='abc'> or <div class='abc'>*

If enabled in Admin settings, the values as defined in the list-file …/include/common/divs_use.txt will be used to index only the content between <div id='abc'> and </div> . Never the less links outside of the div tags will be followed. The values in the list-file will not be interpreted case sensitive. Values in this common list may end with a wildcard, so that 'menu*' will work for ids like
menu1, menu2, menu_left, etc.

Multiple and nested divs will be attended. This is the contray function to
Ignoring parts of a page by <div id='abc'>
which is controlled by the list file …/include/common/divs_not.txt (see chapter above).

Alternately this option could also be used for <div class='abc'>

For even more flexability, the file …/include/common/divs_use.txt may alternately contain a regexp pattern. The regexp needs to be introduced by */ and must be ended with another slash.
Example: */table[0-9]/

7.8   *Ignoring HTML elements defined by <tagname> . . . . </tagname>*

This option is foreseen to cooperate with the new HTML5 elements like
section, nav, aside, hgroup, article, header, footer, etc

HTML elements are written with a **start** tag, with an **end** tag, with the **content** in between:
<tagname> this content</tagname> .

For more details, please notice the HTML element and tag references at
http://www.w3schools.com/html/html_elements.asp
http://www.w3schools.com/tags/

If enabled in Admin settings, the values as defined in the list-file …/include/common/elements_not.txt will be used to delete the content between <tagname> and </tagname> .

Never the less, links inside of the tags will be followed. Values in this common list are automatically added with a wildcard, so that 'aside' will work for HTML elements like
aside1, aside2, aside_left, etc.

For even more flexability, the file …/include/common/elements_not.txt may alternately contain a regexp pattern. The regexp needs to be introduced by */ and must be ended with another slash.
Example: ***/**nav[0-5]**/**

7.9   *Indexing only HTML elements defined by <tagname> . . . . </tagname>*

This option is foreseen to cooperate with the new HTML5 elements like
section, nav, aside, hgroup, article, header, footer, etc.

For more details, please notice the HTML element and tag references at
http://www.w3schools.com/html/html_elements.asp
http://www.w3schools.com/tags/

If enabled in Admin settings, the values as defined in the list file …/include/common/elements_use.txt will be used to index only the content between <tagname> and </tagname>

Never the less links outside of the element tags will be followed. Values in this common list are automatically added with a wildcard, so that 'article' will work for all HTML elements like
aricle1, article2, article_left, etc.

This is the contrary function to
Ignoring HTML elements defined by <tagname> . . . </tagname>
which is controlled by the list file …/include/common/elements_not.txt (see chapter above).

For even more flexibility, the file …/include/common/elements_use.txt may alternately contain a regexp pattern. The regexp needs to be introduced by */ and must be ended with another slash.
Example: ***/**section[0-9]**/**

7.10 *Ignoring parts of a page defined by <ul class='abc'> . . . . *

Ignoring parts of a page by the <!--sphider_noindex--> tags requires direct access to the page, because the tags need to be added (edited) to the page.

A more flexible method, which does not require direct access, is enabled by the Admin setting:
'Use list of ul classes to ignore the complete ul content during index/re-index'

If enabled in Admin settings, the values as defined in the list-file …/include/common/uls_not.txt will be used to delete the content between <ul class='abc'> and . Never the less links inside of the ul tags will be followed. Values in this common list may end with a wildcard, so that 'menu*' will work for classes like menu1, menu2, menu_left, etc.
Multiple and nested ul tags will be attended.

For even more flexability, the file …/include/common/uls_not.txt may alternately contain a regexp pattern. The regexp needs to be introduced by */ and must be ended with another slash.
Example: */menu[0-5]/

7.11 *Ignoring parts of a page defined by <pre class='abc'> . . . . </pre>*

Ignoring parts of a page by the <!--sphider_noindex--> tags requires direct access to the page, because the tags need to be added (edited) to the page.

A more flexible method, which does not require direct access, is enabled by the Admin setting:
'Use list of pre classes to ignore the complete pre content during index/re-index'

If enabled in Admin settings, the values as defined in the list-file …/include/common/pres_not.txt will be used to delete the content between <pre class='abc'> and </pre> . Never the less links inside of the pre tags will be followed. Values in this common list may end with a wildcard, so that 'menu*' will work for classes like menu1, menu2, menu_left, etc.
Multiple and nested pre tags will be attended.

For even more flexability, the file …/include/common/pres_not.txt may alternately contain a regexp pattern. The regexp needs to be introduced by */ and must be ended with another slash.
Example: */menu[0-5]/

7.12 *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 check-box:
    Use 'commonlist' for words to be ignored during index / re-index?

Take notice, that the 'Ignored words' function for many languages is not case sensitive. So, you only need to include one spelling into the common_xyz.txt file.
Instead the common word list is case sensitive for the following languages:
    - Arabic
    - Chinese
    - Cyrillic

7.13 *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 text of the page contains words of the whitelist, the according page will be indexed / re-indexed. The list is placed in the file .../include/common/whitelist.txt

Text-content is defined by Admin settings by means of what to index: full text, title, keywords etc.
Content of links(URLs) is controlled separately by "Must include / must not include string list"

The use of the whitelist may be activated / deactivated by two different check-boxes in Admin / Settings/ Spider settings:
    - Use whitelist in order to index / re-index only those pages
      that include **ANY** of the words in whitelist
    - Use whitelist in order to index / re-index only those pages
      that include **ALL** the words in whitelist

Take notice, that these functions are not case sensitive. So, you only need to include one spelling into the whitelist.txt file.

Content of whitelist is treated as 'words'. So the word 'kinder' in your whitelist will not accept pages that contain the word 'kindergarten'.

Be aware not to place blank rows into the whitelist. Also the list should end with the last word; not with a line feed or a blank row.
- Each word in list must be in a separate row.
- One word per row.
- No blank rows.
- No blank row at the end of the file.

7.14   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 check-box
   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.

Please keep in mind that 'Use of Blacklist is implemented in a different way than implementation of 'Use of whitelist'. Blacklist is interpreting its content as a string. So, the word 'kinder' in blacklist, will also prevent indexing of a page containing the word 'kindergarten'.

Be aware not to place blank rows into the blacklist. Also the list should end with the last word; not with a line feed or a blank row.
- Each word in list must be in a separate row.
- One word per row.
- No blank rows.
- No blank row at the end of the file.

7.15   Ignored files (by suffix)

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

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

While
   image.txt
   audio.txt
   video.txt
are whitelists that include suffixes for files to be indexed, according to the type of media.

If a redirection caused by JavaScript should not be followed (as defined in Admin setting), the suffix 'js' is automatically added to the list in .../include/common/suffix.txt

7.16   Index only files and documents with defined suffix

The list of file suffixes needs to be placed in .../include/common/docs.txt. This file holds a list of all file suffixes for those files (documents) that are indexed / re-indexed.

If the regarding option is activated in Admin backend, all pages of the site will be searched for links, but only files with suffixes as defined in the docs list will become indexed.

8.   UTF-8 support and 'Preferred charset'

Starting with version 1.2, Sphider-plus provides Unicode assistance and starting with version 2.1, the conversion of text into UTF-8 charset is obligatory. In consequence, the impact will be powerful.
First of all: the complete full text, and all header information like title, keywords and description tags need to be converted into Unicode. Consequence is an increase of time required for indexing.

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

1. Detect the charset of site, page or file.
   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 .ptt files,
   the 'Preferred charset' (as defined in Admin settings) will be used
….to convert the file into Unicode.
   In other words: it is not possible to convert DOCs, PDFs etc. that are coded in 'foreign' charset.
   Only those with the charset as defined in 'Preferred charset' will be converted correctly.
   Also it is not possible to convert a Chinese and a Cyrillic coded PDF document
   at the same time. It is necessary to adapt the 'Preferred charset' before invoking
   the index procedure for the sites holding these documents.

2. By means of the PHP function 'iconv()' all texts will be converted into UTF-8.
   This step is successful, if the required charset (for the content to be converted)
   is part of your local PHP installation. In order to find out which charset are available
   in your installation, please 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 use for converting.

3. If the PHP function fails, finally the class 'ConvertCharset' is invoked. This class,
   originally designed by Mikolaj Jedrzejak, enables converting for a lot of charset.
   But it takes more time than the compiled PHP function 'iconv()'.

As result of the charset conversion, the user is enabled to search also for words with non-Latin characters.

In order to enable converting of all charsets into UTF-8, upper and lower case characters are required. So (normally) the query 'html' will not deliver results for sites and files that contain the string 'HTML'. Both are different keywords and stored separately in the Sphider-plus database.

Starting with version 1.6 Sphider-plus offers the additional option:
    'Enable distinct results for upper- and lower-case queries'
If enabled in Admin settings, everything remains as described above. But if this check box is unchecked, result listing will deliver all results; independent of the query input. **HTML**, **html** or even **hTmL** queries will deliver the same (all) results.

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

The following 63 charset are supported by the ConvertCharset function and will be used to convert text into UTF-8 Unicode:

ISO
iso-8859-1

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

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

DOS
cp437 - Latin US
cp737 - Greek
cp775 - BaltRim
cp850 - Latin1
cp852 - Latin2
cp855 - Cyryllic
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

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

MAC (Apple)
x-mac-cyrillic
x-mac-greek
x-mac-icelandic
x-mac-ce
x-mac-roman

DSP implementation for NeXT
stdenc
symbol
zdingbat

And specially for old Polish programs
mazovia

This list is to be read only as completion to the list of charsets as to be found in the sub folder /iconv/ of your server.

9.   Search modes

Beside the original Sphider search modes like:
- Search for a single word.
- AND and OR search.
- Search for a phrase.

Sphider-plus offers 7 additional modes to enter queries:

- Search with wildcard.
- Strict search.
- Tolerant search.
- Link search.
- Media search
- Search only in one domain.
- Search in suggested categories.

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

9.1   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 Sphider-plus database, a lot more results may appear using this search mode. In order not to confuse the user, the printout of relevance (weight/hits) is suppressed. But all available results of a page are presented in result listing.

9.2   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 like '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 and will respond only on a single word query. Strict search will overwrite the 'Phrase Search' option, which is nullified.

9.3   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 lower case 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.

9.4 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.

9.5 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

9.6 Search only in one domain

This mode is invoked by entering:
      site:www.abcd.de query
and will present only results delivered by the (example) domain http://www.abc.de. Search input does not need a blank between site: and the URL of the domain, but between the URL and the query. In contrast to Link search site: , this mode requires the full URL to be entered into the search form (inclusive http://).

Beside www domains, this mode is usable also for localhost applications. The Admin settings 'Address to localhost document root' is used to define the basic address of the local domains.

9.7 Search in categories

After the Admin assigned the indexed sites to different categories, the user of the search engine may rarefy the result listing to manually selectable categories or follow the suggestions of Sphider-plus, offering alternate categories and subcategories that would also deliver results.
For more details please notice the chapter Search in categories

In order to support international languages, Sphider-plus stores all data UTF-8 coded into the database. But some languages require some special attention. For example the Greek language is supported as following.

In Admin backend there are some special settings, controlling the Greek language support:

- Support Greek language. Offers correct support for upper and lower case Greek letters.

- Convert all kind of Greek accents (upper and lower case) into their basic vowels
   Will present e.g. the same query results, when searching with the letter **α** as well as using:
   ἀ, ἁ, ᾄ, ᾅ, ἄ, ἅ, ᾆ, ᾇ, ὰ, ά, ά, ᾀ, ᾁ, ᾂ, ᾃ, ᾄ, ᾅ, ᾆ, ᾇ, ᾰ, ᾱ, ὰ, ᾳ, ᾴ, ᾶ or ᾷ

- Transliterate queries with Latin characters into their Greek equivalents
Supports the old and new Greek alphabet

By activating the above 3 options, the following behaviour is available:

Query input			Find
with/without accents	Latin transliteration old (Perseus)	new Greek	
κυπρος		kypros	κύπρος
ῥύσασθαι , ρυσασθαι	rusasqai	rusasthai	ῥύσασθαι
βαπτίσματος, βαπτισματος	baptismatos	baptismatos	βαπτίσματος
ἀλλὰ, αλλα	alla	alla	ἀλλὰ
ψυχρω		psuchrw	ψυχρῷ
αμφοτερα	amfotera	amphotera	ἀμφότερα
θέλημά	qelhma	thelhma	θέλημά
δοξα	doca	doxa	δόξα
προσεύχεσθε		proseuchesthe	προσεύχεσθε
απορρήξας	aporrhcas	aporrhxas	ἀπορρήξας
Queries with wildcards are followed:		anthrwp*	ἄνθρωπος άνθρωποι Ἄνθρωπος
		noia	Ὁμόνοια ἔννοια

30

Additional remarks:

Searching for "αλλα" will not present any results for words written in Latin characters.
Instead searching for "alla" will present results for:
      alla, ἀλλα, ἀλλὰ, Αλλα, Άλλα, etc.

Depending on which result was found first in full text, the according text extract would show the first hit. In order to see all available results in result listing (Latin + transliterated Greek), it is suggested to increase the value in Admin setting:
      Define maximum count of result hits per page, displayed in search results

If a "strict" search is invoked (!ἀλλὰ), the conversion of Greek accents into their basic vowels will not be obeyed. Also the "strict" search (!alla) will overwrite the transliteration option, so that αλλα, ἀλλὰ, Αλλα, Άλλα, etc. will not be found for the query "!alla".

If the option " Transliterate queries with Latin characters into their Greek equivalents" is activated, only Greek suggestions will be offered. It is assumed that this option is activated, because Greek results are preferred.

10. Chronological order for result listing

10.1 Sorting text results

Sphider-plus offers 10 methods how to sort the text results:
      - By relevance (weight %)
      - By count of hits in full text
      - By last indexed (date and time)
      - 'Most Popular Links' on top
      - Main URLs (domains) on top
      - By URL names
      - Only top 2 per URL (like Google)
      - Promoted domain on top
      - Pages holding a catchword on top
      - Single result per page (ignoring arguments in URL)

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 could be deselected 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.

For the method **'By URL names'** an additional setting is available called:
      'Define number of results shown per domain in result listing'
Using this option, the result listing will present an output similar to 'Like Google', but the count of links added to the main domain is selectable. Instead **'Like Google'** will always present one additional link beyond the main domain.

31

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 %)'** the weight is calculated. 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

may be influenced individually for personal preferences. The according settings could be performed in section 'Page indexing weights' of the Admin settings.

For the method '**Single result per page**', result listing will ignore arguments in URL. This will present only one result for URLs like:
. . . /pizza-restaurants.html
. . . /pizza-restaurants.html?page=1
. . . /pizza-restaurants.html?page=2

Additionally it is possible to create two different methods for **'promoted sorting'** of the result listing:

1. As part of the Admin settings, in section 'Page Indexing Weights', **domain names** could be entered. All search results belonging to this (these) domain(s) will be placed on top of result listing. Multiple domains entered here, need to be separated with commas

2. All pages containing **catchwords** will be displayed on top of the search result listing. As part of the Admin settings, catchwords, separated by commas, could be entered.

Both methods of promoted sorting can be combined. If domain names and also the catchwords are entered in Admin settings, both conditions must be fulfilled to become a promoted link in result listing.

10.2 Sorting media results

Independent from sorting the text results, 5 different modes of sorting the media results are Admin definable:

- By title (alphabetic)
- By file suffix
- By image size
- By 'Last queried'

- By 'Most popular'

11. PDF converter for Linux/UNIX systems

The included PDF converters are not only usable for Latin text, but also convert non-Latin text like Arabic, Cyrillic, Chinese, Greece and Hebrew coded documents.

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.
   Also in 'Settings' menu of the Admin backend, the usually path is shown as suggestion for the option:
   Path and filename of PDF converter:

2. Open the Admin / Settings / General settings menu and palce the path to the PDF converter into the selection field:
   Path and filename of PDF converter:
   Under normal circumstances the correct path is presented in the comment row below the selection field.
   Be aware that the path here needs to end with 'pdftotext' No blanks in front or behind the path setting.

3. Take the full path as above and use simple slashes (not double backslashes) and include your individual path into the file
   .../converter/pdftotext
   First line of this file holds   #!/bin/sh   nothing more.
   Second line of this file begins with a slash and ends with the minus sign.
   A third line is not allowed.
   Overall the file …/converter.pdftotext should look like
   ```
   #!/bin/sh
   /PATH/TO/YOUR/WEB/DOWN/TO/converter/pdftotext.script $1 $2 $3 \-
   ```
   Be aware that in this file you need to define the path to 'pdftotext.script'

   There are 3 different PDF to text converter supplied with Sphider-plus.
pdftotext.script	This is the standard script, which will work in most environments.
pdftotext32.script	This is converter especially developed for 32 bit Operating Systems.
pdftotext64.script	This is converter especially developed for 64 bit Operating Systems.

   In file
   …/converter.pdftotext
   you may define, which of these 3 converter should be used in your application.

4. Set permissions of both pdftotext and pdftotext.script to 755 or 777
   (whatever needed to run correctly).

5. Set permissions of the converter dir to 777, otherwise indexing fails because of pdftotext.script is unable to write a temp file needed! Also the sub folder …/admin/tmp needs to be writable.

The .DOC, .RTF, and the .PPT 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. Also ensure that the editor will store the file UNIX formatted. Please also take notice of

the FAQ chapter: [PDF documents are not indexed](#)

12. 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 resources.
- 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 prodedure, 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. . .](#) "

13. 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 been 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.

14.   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 / Admin 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.

Debug modes are individual available for the Admin backend, as well as for the 'Search User'.

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 Sphider-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. Sphider-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.

In order to check the availability of all required libraries and extensions Sphider-plus is using, the Debug mode will present the corresponding messages on top of the 'Settings' menu.

If 'Debug mode' is enabled for the 'Search user', the cache activity is presented above the result listing in the form of status messages, as well as automatically performed mode settings for Strict search and search with wildcards.

15. **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.

16. Media search for images, audio streams and videos

16.1 Media indexing

Index of media files is enabled by separated Admin settings for:
- Images
- Audio streams
- Videos

Three separate files in sub folder .../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' files. Re-sampling the original images, size of thumbnail is defined to a maximum of 160 x 100 pixel. In result listing these thumbnails are used as a preview.

As far as available the Meta data ID3 and for images EXIF information is indexed and herewith become searchable. In admin backend, indexing and searching EXIF and ID3 info is separately selectable.

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 playtime 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 in the database.

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.

16.2   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>

16.3   Search for media content

The search mode is enabled by the check-box
>   '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 resuls
>   - 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 separately for each keyword match, an additional 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 check-box
>   'Advanced search? (Shows 'AND/OR/PHRASE/TOLERANT/MEDIA' etc.)'
>   is activated, the Search Form will present the additional checkbox

'Search only Media'
If this check-box 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.

The query input 'logo' will present results e.g. for the image 'sphider-logo.gif', while the input 'gif' will show all available gif files.

Additionally the AND, OR and TOLERANT modes are selectable for media search, while the PHRASE mode will be interpreted as an AND search.

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 check-box '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.

In admin backend, searching for media not only by 'title', but also by EXIF and ID3 info is selectable.

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

Additionally there is an Admin setting called
'If found on different pages, index also duplicate media content'
If activated, all images, audio and video stream will be presented in result listing. Otherwise only the first occurrence (page/link) will be presented.

5 different modes of sorting the media result listing are Admin selectable:
- By title(alphabetic)
- By file suffix
- By image size
- By 'Last queried'
- By 'Most popular'

16.4 *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, file name 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

17. RDF, RSD, RSS and Atom feeds

To be activated in Admin / Settings / section 'Spider settings', the content of the following feeds will be indexed / re-indexed:

RDF(v.1.0)   **RSD** (v.1.0)   **RSS** (v.0.91 / v0.92 / v.2.0)   **Atom** (v.1.0)

Feed content and also the links found as part of the feeds will be followed. Before indexing the feeds, a validation check for a well-formed XML is performed. Corresponding log output is generated to inform the admin.

Depending of the feed content, the following tags are indexed and herewith become searchable:

For **RDF** and **RSS** feeds the following standard tags are processed:

Channeltags:	'title', 'link', 'description', 'language', 'copyright', 'managingEditor', 'webMaster', 'pubDate', 'lastBuildDate', 'category', 'generator', 'rating', 'docs'
Itemtags:	'title', 'link', 'description', 'author', 'category', 'comments', 'enclosure', 'guid', 'pubDate', 'source'
Textinputtags:	'title', 'description', 'name', 'link
Imagetags:	'title', 'url', 'link', 'width', 'height'

Additional remark for RSS feeds: The optional sub-elements of the CATEGORY element
(that identifies a categorization taxonomy)
are currently not supported.

For **RDF** feeds, the following individual tags are additionally processed:

Dublin Core tags:	'dc:', 'sy:', 'prn:'
Personal channel tags:	'publisher', 'rights', 'date'
Personal item tags:	'country', 'coverage', 'contributor', 'date', 'industry', 'language', 'publisher', 'state', 'subject'

For **Atom** feeds the following tags are processed:

Metatags:	'author', 'category', 'contributor', 'title', 'subtitle', 'link', 'id', 'published', 'updated', 'summary', 'rights', 'generator', 'icon','logo'
Entrytags:	'author', 'category', 'contributor', 'title', 'link', 'id', 'published', 'updated', 'summary', 'rights'
Authortags:	'name', 'uri', 'email'
Contributortags:	'name', 'uri', 'email'
Categorytags:	'term', 'scheme', 'label'
Generatortags:	'uri', 'version'

Additional remark for Atom feeds: SOURCE elements are currently not supported

For **RSD** feeds the following tags are processed:

Service tags: 'engineName', 'engineLink', 'homePageLink'

API tags: 'name', 'apiLink', 'blogID'

Settings   'docs', 'notes'

There is an Admin setting to strip CDATA tags. It is called: 'Follow CDATA directives'. A blank checkbox in Admin settings will ignore the CDATA directives in RSS and RDF feeds.

An additional Admin setting enables/disables, whether 'Dublin Core' and other individually marked tags in RDF feeds should be indexed.

Another Admin setting allows defining that the 'preferred' directive in RSD feeds should be followed. If activated in Admin settings, only those API tags with 'preferred = true' will be indexed. If the check-box remains blank, all API tags will be indexed, even if 'preferred = false is encountered.

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).

After indexing the feeds, they are treated like other (HTML) pages. The suggest framework will offer keyword proposals. Also pre-selection of categories is taken into account.

18.   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 K bytes of RAM. So, a cache of 2 M Byte 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.

19. Multiple database support

19.1 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.

With respect to the database, the Sphider-plus scripts create automatically individual settings. These settings might be individualized for each database with respect to the personal requirements.

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.

19.2 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. User-name and password are defined in the file …/admin/auth_db.php. As per default download, user- name 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. Assining 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.

19.3 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 four settings available in the 'Activate / Disable' section:

      - Select active database for 'Admin'
      - Select active database for 'Search User'
      - Select all databases that should deliver search results
      - Select active database for 'Suggest URL User'

These settings enable independent use of different databases for:

      Admin
      Search User
      Suggest URL User

'Select all databases that should deliver search results' offers the additional capability to fetch results from more than one database. In any case the active database for 'Search User' will be activated to fetch results, as this database is defined to be the default user database. Searching for results in several databases is available for text and media search and all search modes, taking into account all eventually pre-defined categories Search results are logged with respect to the database that delivered results. Consequently the table 'Most popular searches' at the bottom of result listing is offering results for the currently allocated databases, so that clicking on any of these most popular searches will again deliver results from one or several of the currently available databases.

If multiple sets of tables are available, because they have been created for a database before, you will be able to activate any of these table sets by selecting the corresponding prefix. The selection will be presented below the 'Store all selections' button for all databases containing more than one table set. The selected prefix will be commonly used for Admin, Search user, suggest framework etc.

Consequently the corresponding settings are activated with respect to the database and the activated set of tables.If the table prefix is modified as described in '[Enhancing functionality of multiple database support](#)', this modification is valid for all databases, which are activated to deliver results. In other words, all databases that are used to deliver results and the prefix is manually modified in …/templates/html/020_search-form.html need to contain table sets with the same prefix names

After activating the databases for the different tasks, multiple database support is ready to use. The currently activated database and the prefix (name) of the actual selected table set (for the Admin) is displayed in 'Sites' table like:

> **Database 1 with table prefix 'search1_'   -   Displaying URLs 1 - 10   from 25**

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

> **Results from database 2, 5**

When 'Store all selections' is activated to complete the database activation procedure, also the text cache and media cache will be cleared.

19.4 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. Vise 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 sub-folders (…/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.

19.5 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.

19.6 *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="$user_db" />`
        `<input type="hidden" name="prefix" value="$user_prefix" />`

As long as these variables are set to '0' (how to alter, see below), the search script will use the settings as defined in the Admin settings:
        "Select database for 'Search' user".
This standard setting may be used for the first search form, offering the results of the first set of tables (which e.g. holds the English results). But for a second search form, the value for 'prefix' may be set (name of prefix) for another set of tables that hold the results of the second language. The setting of the second search form will (temporary) overwrite the Admin settings for its own result listing.

Selection of different sets of tables could be performed in the Database => Activate / Disable menu.
If multiple sets of tables are available, because they have been created for a database before, you will be able to activate any of these table sets by selecting the corresponding prefix. The selection will be presented for all databases containing more than one table set.
The selected prefix will be commonly used for Admin, Search user, suggest framework etc.

Multiple database enhancements are assisted by the fact that Sphider-plus is supporting multiple settings.
Each database and each set of tables contains an individual Admin setting.

The selected set of tables could be overwritten individual for a search form by modifying the variable:

    $user_prefix   = "0";

Here the prefix name of the table set, which should be used instead of the default table set, needs to be entered

The selected database could be overwritten individual for a search form by modifying the variable:

    $user_db   = "0";

Here the number of the database, which should be used instead of the default db, needs to be entered

Both variables are to be found in the script .../search_ini.php

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.

Another useful application for multiple table sets would be the support of several languages. By indexing language specific sites into different sets of tables, the 2 hidden fields in the search form will define, which language is presented in result listing.

Usually the language for the user dialog is defined in the Admin backend and could be automatically adapted to the user's language by an additional Admin setting.
In order to overwrite these Admin settings individually for one search form and the according result listing, there is a variable placed in the script …/search_ini.php

    $user_lng   = "0";

As long as the value is set to "0", the Admin settings will be used. Entering "fr", "it", etc., will force this search form to use the languages French, Italian, etc.

20. Search in categories

20.1 Hierarchical structure

In order to prepare Sphider-plus for category search, the categories need to be defined in Admin 'Categories' menu. Different categories (top level) as well as subcategories (Create new subcategory under . . .) are added here.

Second step to prepare Sphider-plus, is assigning the sites to the different categories and sub-categories. To be found at Admin / Sites / Options / Edit / Category.

Assigning and even change of category affiliation may be done also after the index procedure.

Third step is to activate one or both of the Admin settings:

    - Show category selection in search form.

    - If available, user may select 'More results of category . . . ' at each result in results listing.

The first setting will present all prior defined categories and their subcategories as part of the search form and the user may select one top-level category or a subcategory to limit the search results.

The Admin setting

    "If available, user may select 'More results of category . . . ' at each result in results listing"

will present all additional categories that would also deliver results for the search query. Presented under the link URL, the user may click on any of these suggestions to automatically perform a new search in the other category.

The new result listing will present all results of that category and, if available, will also suggest subcategories that would be able to deliver results for the current query. Again the user may rarefy the result listing. Clicking on any suggested subcategory, will again perform a search for the query; now in the selected subcategory.

As additional headline of the result listing, the user will be informed about the source (category) of the results like:

    Presented results are captured from category: <mark>abc</mark>

In order to return to the standard search without category-selection, the user only needs to activate the checkbox

      'Search in all sites'

as part of the Search form.

20.2 Parallel structure

Starting with version 3.0, Sphider-plus offers the feature to restrict the search results by means of up to 5 categories simultaneously. The titles (names) for the 5 categories could not be defined in the Settings menu of the Admin backend (see section 'General Settings').

If the title for category0 or category1 is defined with any value, the hierarchical category structure is disabled, and the search results will be restricted by the parallel category structure. After defining the names of all categories that should be used to restrict the search results, any of the 'Save' buttons in the 'Settings' menu of the admin backend needs to be activated. Afterwards the category names are available for all Sphider-plus scripts.

Undefined categories, because the regarding title is left blank, will remain uncared by the search algorithm, and also will not be presented in the search form.

The relationship between URL and all involved categories is to be defined in a file, which is imported into Sphider-plus by means of the 'Import / export URL list' feature of the Admin backend.
As per default, the import file needs to be placed into
      …/admin/urls/
as a sub folder of the Sphider-plus installation. If the import file is to be found in another folder, the regarding address needs to be edited in the script
      …/admin/configset.php
Mostly on top of the script, the variable $url_path has to be adapted. After editing the value of this variable, any of the 'Save' buttons in the 'Settings' menu of the Admin backend needs to be activated. Afterwards the variable is available for all Sphider-plus scripts.

The import file could be a pure text file or an EXEL table. The name of the file is user-defined, but must contain at minimum one underscore. Examples:
      2012.12.07_import.txt
      My_friends.xml

If a text file should be imported, each URL and the involved category values is to be presented in a separate row. Each value is to be separated by a delimiter, which is defined to: |

Instead the EXEL file could be used without delimiter and each cell has to hold one value.

The import file could transfer the following values:

$url   **Always start with http://**
$title   **user-defined title, which will be presented in the 'Sites' view**
$short_desc   **Short description**
$index_date   Date of last indexation. For new URLs: 0000-00-00
$spider_depth   Depth of indexation (sub folders)
$required   Required words in link URLs
$disallowed   Forbidden words in link URLs
$can_leave_domain   Crawler may leave the domain during indexation
$db   Active database for the Admin backend
$smap_url   If available, follow the directives of a sitemap.xml file
$authent   Indexation only with identification
$use_prefcharset   Obligatory use the preferred charset as defined I the Admin backend
$cat_sel0   **Value for category0**
$cat_sel1   **Value for category1**
$cat_sel2   **Value for category2**
$cat_sel3   **Value for category3**
$cat_sel4   **Value for category4**
$prior_level   Value for preferred index procedure

Usually only the bold printed variables should be defined. The other variables should only be set, if complete knowledge about their meanings is present. In case that one or more values are not defined, the value 'all' will be used for the regarding URL. Each value may consist of maximum 255 characters. The import file should be UTF-8 coded.

The variable $cat_sel0 could be used to build up a range selection (min. – max. values) in the search form. This might be useful for postal codes, etc. The search form will present 2 selection fields for $cat_sel0, which could be used by the user of the search engine to define the minimum and maximum values. All results matching this range will be presented in result listing. Additional the categories $cat_sel1 - $catsel4 could be used to improve the search results.

21. User suggested sites

Reachable via a link at the bottom of result listing, a form is presented that allows users to suggest URLs to become indexed by the search engine. The user needs to enter:
- URL
- Title
- Short description
- Dispatcher e-mail account.

In order to prevent spam proposals, the form optionally will present a Captcha.

The admin of the search engine may
- approve
- reject
- bann

the suggested sites by means of a menu, presented in Admin backend. A corresponding e-mail is automatically generated and sent to the dispatcher.

All features of 'User suggested sites' are optional and could be defined as part of the Admin settings.

An additional option offers the function:
Suggested sites require authentication tags

If activated, all suggested sites would need an additional meta tag in their header. This authentication tag needs to be written as:
<meta name='Sphider-plus' content='1234'>

The content value (here e.g. 1234) is defined by the administrator of the search engine. As part of the approval form, an additional field needs to be filled in by the admin. So, individual values could be defined for each suggested site. The text of the automatically generated acknowledgement e-mail, sent to the dispatcher, is altered to:
Your suggestion was accepted by the system administrator and will be indexed shortly.
Please add the following tag into the header of the suggested site:
<meta name='Sphider-plus' content='1234>
In order to enable indexing of your site, this tag is mandatory
and is tested periodically by the indexer of Sphider-plus.
We appreciate your help and effort in building this search engine.
This mail was automatically generated by Sphider-plus.

The meta tag needs to be implemented only into the suggested site. It is not necessary to add this tag into all pages of the site. Only the header of the suggested URL will be verified for existence of the tag and correct content value.

The authentication value may be altered by the admin of the search engine later on.
In
Sites view => Site Options => Edit
an additional input field is presented. If the value is left empty, the site will be indexed without verification of the header tag. The dispatcher will not be informed about any modification done by the admin.

The additional input field to enter/modify the authentication value is offered for all sites stored in the database of Sphider-plus, so that an authentication value could be added also subsequently by the admin.

If the tag is missing or contains an invalid authentication value, a corresponding warning message is created during index procedure. The complete site with all their pages will be skipped by the index procedure, but the former content as well as the known links will remain part of the Sphider-plus database. This behaviour offers the capability to reactivate the site by the admin later on.

22. Vulnerability protection

22.1 *Intrusion Detection System (IDS)*

Based on the PHPIDS scripts, the modified IDS is used to protect Sphider-plus against hacking attempts. All user input is observed and analysed. The IDS includes extensive regexp rules to tags like:

CSRF	Cross-site request forgery
DoS	Denial of service
DT	Directory traversal
ID	Information disclosure
LFI	Local file inclusion
RFE	Remote file execution
SQLI	SQL injection
XSS	Cross-site scripting
LDAP Injections	Lightweight directory access protocol injection

The IDS is to be activated in Admin settings (section General Settings) by the check-box:
- Use 'Intrusion Detection System' to prevent input attacks

If there's no attack, the Sphider-plus scripts are executed, but if the IDS recognizes an attack, it prevents the scripts from being executed and displays a warning message to the hacker. This message is always displayed in English, so that hackers from foreign countries will always get a readable info.

An additional Admin setting allows activation of:
- Block further Internet traffic of IP's, which caused intrusion attempts
If activated, and the IDS detects a serious attack, all user input for the 'Search' form as well as for the 'User may suggest a site' form is blocked with respect to the hacker's IP. Independent from the further input. In order to enable cleaning of the IDS log-file, admin authentication remains always accessible.

The IDS is creating a log entry for every intrusion attempt. This log file could be observed in the admin backend. Available as part of the 'Statistics', as well as in the 'Clean' menu. Documented in the log file are
- Date and time
- IP
- Impact value
- Involved tags (XSS, DoS, SQLi, etc)
- Input data of the attack.

If the impact is >= 14 a warning message is created for the hacker, while the IP traffic is blocked if an impact is >= 25

The log-file could be cleaned completely by the Sphider-plus admin. In order to enable a specific IP traffic again, the admin is also enabled to delete individual entries of the log-file.

In order to check the correct functionality of the IDS, you may enter
        qwerty">< '<>
into the search form. This will cause an impact of 23 and consequently 'only' causes a warning message.

In order to seriously try to intrude with an evil string, enter
        %5C%22%3E%3Cscript%28window.name%29%3C%2Fscript%3E","%3D%2522%253E%
Be aware: after entering the above string into the search form, complete IP traffic (which caused the intrusion attempt) will be blocked. Only by means of the Admin backend and deleting the corresponding entry from the IDS log file, will enable IP traffic again.

In case that the IDS is not activated in Admin settings, all user input in any case will still be observed by the function cleaninput (…/include/commonfuncs). This function will also protect against hacks, but does not log the attempts and also does not block the Internet traffic for the Sphider-plus input forms.

The input data of the attacks are not presented in the Admin backend. In order to perform a deeper analyze of the attack, the log-file needs to be opened with a text editor at …/include/IDS/tmp/phpids.txt

22.2 *Prevent queries from Meta search engines and crawler known to be evil*

In order to reduce Internet traffic and server load, there are 2 settings available in Admin backend in section 'Search Settings' called:
- Block all queries sent by harvester, bots and known evil user-agents
- Block all queries sent by Meta search engines like Google, MSN, Amazon, etc.
If activated, the corresponding search queries will be prevented.

The first option is controlled by the file
…/include/common/black_uas.txt
holding lists of user agents known to be evil.

Meta search engines are identified by their IP. The IPs could be entered as single IP, as well as IP ranges into the file
…/include/common/black_ips.txt

Prevented queries are answered with the text 'No results found'.

Instead, if the option 'Enable Debug mode for User interface' is activated in Admin backend, the IP (which caused the query) is also presented as result. If an evil user agent sent the query, the client user agent string is presented.

22.3 *Basic input validation against vulnerability attacks*

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.

As the protections against XSS attacks, Shell execution, Tag inclusions, as well as the suppression of JavaScript executions do avoid some words in the search query (exec, system, union, etc.) a special Admin setting is used to activate this protection. The setting is to be found in section "Search Settings" and is called:
Block all queries, which could cause an XSS attack, Shell execution, Tag inclusion,
or a JavaScript execution

23. Bound database

Entering a word into the search form will force Sphider-plus to scan the database for all links offering a result for this query. Already integrated had been a limit to present only x results. Never the less to find these x results, the complete database had to been browsed. Starting with version 2.5, an additional clean option is offered as part of the Admin backend. Main advantage of this option is a significant reduction of the search time for any query, because the content of the db could be limited to offer only x results. This option will be most useful especially for huge databases, holding the content of many links.
The setting in section 'Search Settings' called:
>   Define max. amount of results presented in result listing
will define the limitation for the database volume. In order to activate this limitation the 'Clean' menu presents the option:
>   Bound database
When activated, all keyword / link relationships - stored during index procedure – will be reduced to the above defined amount. All overhanging relationships will be deleted from the database. Consequently all further search enquires will be responded much faster, because only the relevant amount of results are available.

It is up to the admin to define how many results are relevant for the application Sphider-plus is integrated into. The 'Top keywords' table, as part of the 'Statistics' menu, could be helpful to define the limit. Once the database is bounded, also this table will only show the 'bounded' available results.

The 'Bound database' option should not be invoked, if the chronological order of result listing is defined to 'By hit counts in full text' or to 'By index date', because limitation of the database is performed by weighting.

Some patient is required for bounding the database. Once activated in 'Clean' menu, the following steps need to carried out:
>   - Get all keywords from database.
>   - Get all results from db for each keyword.
>   - Bound the results of each keyword to the defined limitation.
>   - Delete all possible results, exceeding the limitation (for each keyword).

24. Suggest framework

Sphider-plus offers an auto-complete function as part of the search form. Starting with version 3, the suggest framework was switched over from 'Prototype' to the JavaScript library 'jQuery'.

Suggestions are presented for single word queries, as well as for phrases, and also for media search. The suggest framework is configurable in Admin backend. As part of the 'Settings' menu in section 'Suggest Options', the following items are definable:
>   - Minimum count of query letters in order to get a suggestion.
>   - Search for suggestions in query log.
>   - Search for suggestions in keywords.
>   - Build suggestions also for 'Phrase' search.
>   - Obey the list of words not to be suggested.
>   - For 'Media' search get suggestions also from EXIF info and ID3 tags.
>   - Limit count of suggestions.

The option '**Obey the list of words not to be suggested**' is based on a list of words placed in the file
>   …/include/common/suggest_not.txt
All words placed in this file will be used to suppress this suggestion. All other suggestions will be presented. Even if the stop word in the list is only part of a suggestion, the suggestion will be also suppressed.
Example: If the list contains the word 'kinder', also 'kindergarten' will not be presented as a suggestion.

With respect to the different templates of Sphider-plus, also the design of the suggest field is adapted to the regarding template. Responsible for this adoption is the style sheet file

/templates/your_template/jquery-ui-1.10.2.custom.css

These style sheet files are based on the themes of jQuery. They are edited individually for the Sphider-plus templates 'Pure', 'Slade' and 'Sphider-plus'. Each of these style sheet files contains a link (see row 4) to the jQuery 'ThemeRoller'. Thus individual adoptions to personalized Sphider-plus templates are enabled by just editing the style sheet …/templates/your_template/jquery-ui-1.10.2.custom.css by means of the jQuery 'ThemeRoller'.

25. Integration of Sphider-plus into existing sites

There are 2 different ways of integrating Sphider-plus into existing sites:
1. Use layout and templates of Sphider-plus.
2. Embed the search engine into existing HTML code.

25.1 Integration into existing sites by use of Sphider-plus templates

This mode is simply invoked by calling the script 'search.php' in the root folder of the Sphider-plus installation. Assuming that the search engine is placed in a sub folder called 'sphider-plus', the according call would be something like:

http://www.abc.de/sphider-plus/search.php

Once called, the search engine will build up a complete HTML page with
- Headline
- Search form
- Result listing
- Footer

The design of this page is defined by one of the 3 templates delivered together with Sphider-plus. They are named:
- Pure   (close to Google design)
- Slade   (dark shadow design)
- Sphider-plus   (default, as on the project page)
and are selectable by the Admin backend of Sphider-plus.

Usually no one of these 3 templates will fulfil the requirements of an existing site design. Consequently the (activated) style sheet
…/templates/Pure/userstyle.css
needs to be individualized.

In order to create a new template it might be useful to copy one of the Sphider-plus template folders completely with all files, rename the folder with a new name and afterwards edit the personal style sheet
…/templates/your_template_name/userstyle.css

The new template will also be presented in the Admin settings, as one of the available selections.

25.2 Embed the search engine into existing HTML code

This mode extends the capabilities as described in the above chapter. There is an Admin setting called:
- Embed 'Search form' and 'Result listing' into an existing HTML page

If this checkbox is activated, the search engine will not create a complete HTML page, but needs to be embedded into an existing HTML code.

As the Sphider-plus scripts are based on charset UTF-8, in any case, the scripts for search form, result listing, etc. could only be embedded into UTF-8 coded HTML pages. No problem to index other coded content, but embedding is possible only on base of charset UTF-8.

As described in above chapter, the script 'search.php' is used to embed the search engine into the existing page. In general the script 'search.php' only consists of some definitions (prepared in …/search_ini.php) and the required 'include' directives.

The 'search.php' script contains complete documentation (comments) how and where to use the different include directives of this script, so that it needs not to be repeated here. In any case the

        include "$include_dir/search_10.php";

should be placed into the HTML header, before an already existing stylesheet.css file is called. This approach will ensure that the existing css will automatically overwrite the Sphider-plus style sheet. Consequently only the search engine specific settings need to be modified in the Sphider-plus userstyle.css.

On the same subject, another Admin setting might be also helpful:
        Name of search script

By means of this option, an individual script could be defined and used to control the search engine, by containing the required 'includes'. Sphider plus will automatically reference on this script, when searching and presenting the results.

The Sphider-plus functions
        - Intrusion Detection System
        - User may suggest URLs form
        - Admin backend
are always using the template as defined in the Sphider-plus Admin backend and (at present) are working non embedded.

25.3 The different style sheet files

Sphider-plus is delivered with two style sheet files:
        - adminstyle.css
        - userstyle.css

Both files are part of each template design (Pure, Slade and Sphider-plus) In order to adapt one of the three templates to an existing site design, only the userstyle.css file needs to be individualized. So the Admin backend remains stable and executable, even during development of the final design for the user interface.

With respect to the different templates of Sphider-plus, also the design of the suggest field is adapted to the regarding template. Responsible for this adoption is the style sheet file
        /templates/your_template/jquery-ui-1.10.2.custom.css

These style sheet files are based on the themes of jQuery. They are edited individually for the Sphider-plus templates 'Pure', 'Slade' and 'Sphider-plus'. Each of these style sheet files contains a link (see row 4) to the jQuery 'ThemeRoller'. Thus individual adoptions to personalized Sphider-plus templates are enabled by just

editing the style sheet …/templates/your_template/jquery-ui-1.10.2.custom.css by means of the jQuery 'ThemeRoller'.

26. JSON and XML result output

Usually the result listing is presented as HTML output for the client that has sent the query to the Sphider-plus scripts. Additional output is available as JSON and XML files. If requested by the search_ini.php script, the results will be presented as separated files in sub folder …/xml/

With respect to the query type, JSON and XML files will contain text, media, or link results. The according output files are called:

text_result.txt	=>	JSON output file
text_result.xml	=>	XML output file
media_result.txt		
media_result.xml		
link_result.txt		
link_result.xml		
multiple_link_result.txt		
multiple_link_result.xml		

As media results could be images, audio streams, as well as videos, the media type is marked by a type tag in the output files.

The additional JSON and XML output files are activated by the variable $out . In order to create the additional output files, the variable needs to be set to 'xml'. To be found in the script …/search_ini.php

The same script also offers the variable $xml_name (above set to 'result'), which may be used to define individual names for the output files. Individual for each search form. The names are always completed by one of the prefixes

    text, media, link, multiple_link

so that the complete name will become

    text_your-choice.xml
    media_your-choice.txt
    etc.

If a new query is sent to Sphider-plus, first of all the old JSON and XML result files are deleted from the sub folder .../xml. This is performed before searching the database for possible results. When new results are found in db, the search-script will store the results in new JSON and XML file, and also present the results as HTML output to the client's browser.

Additionally all JSON and XML output files are saved in sub folder

    .../xml/stored/

The file names in this folder additionally contain date and time of storage like

    _2014.03.02_04-47-10_PM_text_results.xml

The files in sub folder . . ./xml/stored/ will not be deleted automatically, but remain available until manually deleted by the admin backend. To be performed in menu 'Clear' by means of the item:

    Delete all files in 'XML' sub folder

An XML media result file may look like:

```xml
<?xml version="1.0" encoding="utf-8" ?>
-<media_results>
        <query>warp</query>
        <ip>127.0.0.1</ip>
        <host_name>www.007guard.com</host_name>
        <query_time>2014-03-02 05:25:04 PM</query_time>
        <consumed>0.447</consumed>
        <total_results>2</total_results>
        -<media_result>
        <num>1</num>
        <type>image</type>
        <url>http://www.abc.de/index.php</url>
        <link>http://www.abc.de/images/warp.gif</link>
        <title>warp.gif</title>
        <x_size>635</x_size>
        <y_size>98</y_size>
</media_result>
        -<media_result>
        <num>2</num>
        <type>audio</type>
        <url>http://www.abc.de/index.php</url>
        <link>http://www.abc.de/my_music/warp.m4a</link>
        <title>Warp.m4a</title>
</media_result>
</media_results>
```

Instead a JSON text file may contain the following content:

{"query":"colour","ip":"::1","host_name":"guard_007-hoster","query_time":"2016-03-18 10:56:23 AM","consumed":0.016,"total_results":2,"num_of_results":2,"from":1,"to":2,"text_results": [{"num":1,"weight":"100.0","url":"http:\/\/www.english.le-piaggie.info\/Info_eng.pdf","title":" Info_eng","fulltxt":" \r\n . . . Preamplifier and driver for connecting cable. HDTV receiver for digital TV and radio. DVB-S\/S2 and DVB-T, digital colour TV set with 24\" LED back light illuminated display. 7 Exterior Views 8 9 Interior Views Former, the cow and pig stable were placed on the ground floor, today's big kitchen with the heating fireplace for open and close mode of . . .\r\n ","page_size":"5,661.6kb","domain_name":"www.english.le-piaggie.info"},{"num":2,"weight":"50.0","url":"http:\/\/www.english.le-piaggie.info\/html\/description.html","title":" Description","fulltxt":" \r\n . . . cable. Additional active driven DVB-T antenna. HDTV receiver for digital TV and radio. DVB-S\/S2 and DVB-T digital colour TV set. First floor Upper floor Plot: Size: about 6,2 ha Hilly side, partly wooded View over the hills of the Chianti D.O.C.G. as far as to the Pratomagno mountains 160 olive-trees Detached house, about 800 m to reach the . . .\r\n ","page_size":"26.4 kb","domain_name":"www.english.le-piaggie.info"}]}

27. FAQs

27.1 Shouldn't the spider follow 301 http redirects?

Yes, Sphider-plus follows 301 redirects. But it might be necessary to enable
'Spider can leave domain'
in Sites /Options / Basic Indexing Options

27.2 Why do I get the message 'The search string was not found as part of the text'?

Only a warning message. Will be presented in result listing, if the found keywords are not part of the full text, but were found only in URL or meta tags of the indexed page.
You may disable this warning message in Admin / Settings/ Search Settings / by deselecting the checkbox:
Show warning message if query was not found in full text;
but only in 'Title' of page, 'Keywords' 'Meta tags' or 'URL'

27.3 Unable to log in as Admin. Always re-directed to the log in form. Why?

Verify that you really use the access authorization as defined in the script:

.../admin/settings/authentication.php

As the values are stored hashed, you will not be able to read them. In case you've forgotten them:
There is a backup file of 'authentication.php' in sub folder .../admin/backup/ containing name and password set to 'admin'.
Copy and paste this script over the existing .../admin/settings/authentication.php

if still the empty log in form is presented after entering 'Name' and 'Password', there might be a problem with session control. This option must be enabled for PHP scripts on the server. In case you are running Suhosin on the server, attention, as it encrypts sessions differently. Adding the following to the php.ini in the Sphider-plus root directory , will let you get access to the Admin backend
suhosin.session.encrypt = Off;

27.4 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 significant reduces the time required for index and re-index.

27.5 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 check-box 'Spider can leave domain' in Admin/Sites/Options/Edit/ . Otherwise Sphider will not follow the redirect directive of your .htaccess. file.

27.6 How to integrate the Sphider search form 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_t" 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 adoption. For example if you add
```
<input type="hidden" name="mark" value="markyellow" />
```
the found keywords will be marked yellow.

A complete search form is to be found in script
        …/templates/html/020_search-form.html

For more details about embedded operation of Sphider-plus, please notice the chapter
        Integration of Sphider-plus into an existing site
of this documentation.

27.7 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

27.8 *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
.../include/commonfuncs.php
and delete the row
$res = $db_con->query("FLUSH TABLE $row[0]");
Also the rows, used for debug output, should be deleted if (!$res) { . . . }

Additionally open the file
.../admin/spiderfuncs.php
and delete the row
$db_con->query("FLUSH QUERY CACHE");
Also the following rows, used for debug output, should be deleted.

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.

27.9 *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.

27.10 *Error message: "Access-Denied: You need the SUPER privilege for this operation."*

Another server limitation. This time facing a restriction concerning the MySQL server. In order to solve it, uncheck the setting:
"Enable 32 MByte MySQL query cache"

27.11 *Error message: "MySQL failure: INSERT command denied to user . . . "*

 Another server limitation. This time facing a restriction concerning the MySQL server.
On 'Shared Hosting' server, usually the size of databases is limited (e.g. 100 MB). For huge amount of indexed pages, this limitation will cause the error message. Even for search queries, because the statistics algorithm of Sphider-plus tries to save the current query.
For example and with respect to the amount of keyword/link relationships, a database containing 115 sites + 5.388 page links + 109.224 keywords, might occupy about 260 MB. There is no chance to overwrite the according provider settings.

27.12 Error message: "MySQL failure: Specified key was too long, max. key length is 767 bytes"

This is an issue with certain versions of MySQL (5.6.x) and utf8. In case you suffer from this MySQL bug, the installation of database tables will fail. Consequently the script …/admin//install_tables.php will throw this error message. Use 'latin1' instead of 'utf8' charset. The 767 byte limit still exists but MySQL will silently truncate the index for you as indicated in the MySQL documentation.
For more details, see here:
http://docs.oracle.com/cd/E17952_01/refman-5.6-en/innodb-restrictions.html

27.13 Error message: "MySQL failure: You have an error in your SQL syntax; . . . ."

The complete error message
> MySQL failure: You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '-upsep LIKE 'search1_%'' at line . . ."

It is known that some server do not accept a hyphen as part of the database name. Consequently names like
> db-sphider-plus
> mydb-123456

should not be used. As a precaution, the same restrictions should be attended for user names and passwords.
Even if provider like 'Hosteurope', who specify such db names, such names will fail for some mysqli commands.

27.14 Error message: "MySQL failure: MySQL server has gone away"

Might be a problem of too much data transferred to the MySQL server in one piece. Because the complete full text of each page been indexed, is stored in database (table: links, column: fulltxt).
In order to fix this issue, try the following:
> On your server in folder
> …/mysql/bin/
> find the script my.ini
> Here you may define the size of max. allowed packet.
> Usually set to 1M, you may alter the value to 20M like
> max_allowed_packet = 20M
> Afterwards you need to restart your server.

Not necessary to increase the value above 20M, because column 'fulltxt' in db is created as type 'mediumtext'. This type is limited to max. data of 16 MiB. This value is not the max. size of page content, which could be indexed, but the max. size for the extracted full text. Without images, HTML tags, JavaScript etc. 16 MiB of pure text, extracted by the Sphider-plus index procedure.

27.15 Error message: "Logging option is set, but cannot create folder for logging files."

Usually the admin backend of Sphider-plus creates all folders and log files with full write permission. But depending on your server and its restrictions for PHP scripts (like Sphider-plus), it might become necessary to chmod some sub-folders and files inside these sub-folders (chmod 777) to get full write permission.
In Sphider-plus forum it was discussed here
http://www.sphider-plus.eu/forum_529/read.php?5,2104,2106#msg-2106
and also here
http://www.sphider-plus.eu/forum_529/read.php?5,2104,2104#msg-2104
Please read my last post of this thread to understand, why I did not implement the suggestion from 'dsom73'. Sphider-plus has to survive on most different servers all over the world.
The log files during index procedure are stored in sub-folder
.../admin/log/

27.16 Error message: "Attention: Sphider-plus recognized a server error (clocks asynchronous)."

Usually the clocks for PHP and MySQL are running synchronously and without any offset. But on some servers, there is a difference between them. Differences of up to 5 hours were already noticed.
One way to fix this problem is to modify the 'date/timezone' in PHP environment. To be done in php.ini script as part of the PHP server environment by defining

    date.timezone = 'Europe/Berlin'

Of course 'Europe/Berlin' needs to be adjusted to the local requirements. A list of PHP supported time zones is to be found here http://www.php.net/manual/en/timezones.php

In case that php.ini is not accessible on the server, like on a 'Shared Hosting' server, the php.ini scripts in Sphider-plus root installation folder, as well as in sub folder …/admin/ could be used to define an individual timezone for Sphider-plus. Always both php.ini scripts need to be adapted.

27.17 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 initialization 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.

27.18 PDF documents are not indexed

Option 1:
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.

In any case the C++ standard library (libstdc++) needs to be enabled on the server. This library provides several generic containers, functions to utilize and manipulate these containers, function objects, generic strings and streams, so that the PDF converter, which is a binary, could be executed on the server.

Option 2: Only document titles are indexed
Are you sure it is a PDF document containing text? Or might it be an image, converted to PDF, which could not be indexed by Sphider-plus?
In order to check it out: Open the document in your PDF reader and try to copy and paste a single word from text content. In case that immediately the complete text is marked, and not only one word, the document is an image, which was just converted into PDF.

27.19 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 //

27.20 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.

Additionally an 'Intrusion Detection System' could be enabled as part of the Admin settings. If activated, all attempts to hack Sphider-plus are logged, a warning message is presented and further Internet traffic is blocked for the IP causing the attack. The IDS will additionally protect against:
- Cross-site request forgery
- Denial of service
- Information disclosure
- Local file inclusion
- Remote file execution
- Lightweight directory access

27.21 How to protect Database management against Admin access?

As per default, the sub-menu '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

27.22 Messages like: "Results from database1" are displayed on top of the result listing.

If in Admin settings the 'Debug' mode is enabled, several warnings and messages are displayed.
To suppress these messages, the check-box 'Enable Debug mode' in Admin settings needs to be unchecked.
Please keep in mind that there are separated setting available for 'Admin' and 'Search User'.

27.23   Unable to search for several words like clock, file and system. Why?

In order to prevent vulnerabilities like XSS attacks, SQL-injection etc, Sphider-plus is checking all user input as well as all client data sent to the server. Input containing 'bad' words is rejected
All input has to pass the function cleaninput($input) in the script .../include/commonfuns.php
By means of several preg_match(...) functions the bad words are detected and filtered. In order to avoid conflicts with common user queries, the corresponding filter words could be deleted. Always together with the following OR selector (for example **clock|**).

27.24   Indexing stopped after 20 links, but my site contains more than 650 pages.

Indexing with a search engine like Sphider-plus may become problematic on a 'Shared Hosting' server. Indexing huge amount of links might be interrupted, because the granted time slice can be finished before index procedure is finished. Sphider-plus tries 3 times to reconnect to the database. But if the script was canceled, it will become necessary to manually invoke again the index procedure to continue. Sphider-plus will remember the last indexed link and continue the suspended process.

27.25   Don't see the links, keywords and thumbnails on my screen during indexing, why?

There are some Admin settings that need to be attended:
    - Enable Debug mode   => must be activated
    - Suppress browser output of logging data during index/re-index   => must not be activated
If 'Multithreaded indexing' is activated, Sphider-plus takes control over these options. Because. in order to speed up the index procedure, all not required options will be switched off. But if you return to single thread indexing, Sphider-plus does not remember the old settings.
So it is up to the admin to reactivate the options with respect to his personal preferences

27.26 How to fasten the index procedure?

Deactivate all options in 'Settings' menu of the Admin backend, you don't really need. Each active option will take additional time to be performed.
Also you should create a sitemap.xml file of your site, so that the indexer does not need to search for all links in full text of all pages, but may use the much faster option:
    - If available follow sitemap.xml
Additionally, if you are sure about your index procedure, you may deactivate the 2 options:
    - Enable Debug mode for Admin backend.
    - Log spidering results into a Log-file.
Of course you will miss all details about index procedure, but you will save time.
Some other possibilities to reduce index time could be defining the following options to the really required minimum:
    - Define indexing depth.
    - Bound the length of full text indexed at each page.

27.27 In the search results I'm seeing the full text information repeated. Why?

There is an Admin settings (in section Search Settings) :
> "Define maximum count of result hits per page, displayed in search results
> (if multiple occurrence is available on a page)"

If you enter any value > 1 into this field, Sphider-plus may present several text extracts of one page. Because, if the keyword was found for example. 2 times in full text of a page, the result listing will present some text 'around' the found keyword position two times.

27.28 Receiving 'server error 500' on a fresh installed Sphider-plus (option 1)

Using the Apache suEXEC module, which allows users to run CGI and SSI applications as a different user, may cause this error message. Using the suEXEC may result in a conflict with the chmod 777 performed by the Sphider-plus Admin backend, which tries to get full write access to several sub-folders of the Sphider-plus installation. It may become necessary to disable all chmod 777 commands in …/admin/admin.php

27.29 Receiving 'server error 500' on a fresh installed Sphider-plus (option 2)

Sphider-plus is delivered with several .htaccess files in some folders. These scripts contain directives, which try to overwrite the server settings. Unfortunately some server do not accept these .htaccess directives. If receiving the above error message and the server settings do not allow 'overwriting', all .htaccess files of the Sphider-plus distribution need to be deleted.
This issue was reported e.g. for servers like Mageia 1, Mandriva 2010.2 and CentOS 6.0

27.30 Receiving 'server error 500' on a fresh installed Sphider-plus (option 3)

If the Sphider-plus scripts are installed on a server hosted by e.g. 'Hosteurope', it was reported to be a server conflict for the script …/admin/geoip.php
It might become necessary to disable this script and the GEOIP functions in Sphider-plus.

27.31 In the addurl form, is there a way to remove "none" as a category option?

Open the script …/addurl.php and delete the row:
> print "<option ".$selected." value=\"0\"> none</option>\n";

27.32 For the addurl form, how to make the Captcha text input not case sensitive?

Open the script …/addurl.php and find the row:
> if ($_SESSION['CAPTCHAString'] != $_POST['captchastring']){

Delete that row and replace it with
> if (strtolower($_SESSION['CAPTCHAString']) != strtolower($_POST['captchastring'])){

27.33 *Unable to rename the default search script. I am always redirected to search.php*

If .../search.php is no longer the default script, you will have to modify the .htaccess file in the root folder of your Sphider-plus installation for your personal requirements.

In .htaccess you will find:
>   #   2. Redirect client enquiries to search.php
>   RewriteEngine on
>   RewriteRule ^search\.html$ ./search.php
>   ...
>   ...
>   # 4. Always start with this file
>   DirectoryIndex search.php

27.34 *Parse error: syntax error, unexpected ';' in ..\settings\db1\conf_search1_.php on line 33*

This error message is presented, if someone manually edited the configuration file. It is not foreseen to edit any configuration file. All modifications need to be done in the Admin backend in menu 'Settings'.

The above error message is a total knock-out for Sphider-plus. Delete the corresponding configuration file in the sub folder as defined in your error message (e.g. .../sphider/settings/db1/conf_search1_.php).
Additionally restore the script
>   …/admin/configset.php

with the original script as of your Sphider-plus download.
Afterwards open the Admin backend and find the default settings replaced by Sphider-plus into your configuration file. Now modify the standard settings with all your individual settings in the 'Settings' menu. At the end of all, press any of the 'Save' buttons. If you stored a valid configuration backup file before starting your manual manipulation that causes the above error message, you may also restore this backup.

27.35 *Only the first part of a page gets indexed. The rest of the text got lost. Why?*

Might be a problem of incorrect defined HTML tags. In case that a tag is not closed correctly, indexing for that page will be ended with the incorrect tag. Words inside of tags are not part of the full text. But only the text of a page should be indexed. The indexer is using the PHP function strip_tags() to delete the tags from the page content.
Cit from the PHP manual:
>   "Because strip_tags() does not actually validate the HTML, partial or broken tags can result in the
>   removal of more text/data than expected."

In order to validate the HTML code, the following link might be helpful:
>   http://validator.w3.org/

This problem is solved since Sphider-plus version 2.7, because the PHP function strip_tags() is no longer used. A new function was created, now accepting also unclosed and invalid HTML and PHP tags.

27.36   Indexing from command line shows "Fatal error: Call to undefined function getHttpVars()"

The indexation script is placed in sub folder …/admin/sphider.php If the index procedure is invoked from command line, the error message will appear, because sphider.php was called from a folder different than …/admin/
Solved since version 3.2015c

27.37   Cron: Logging option is set, but cannot create folder for logging files.

When running from CRON you'll get "include()" errors because some files aren't being included.
That is because they're being relatively included with "../".
The full explanation for this issue, together with a solution, is described in the Sphider-plus forum 'Tips & Tricks & Mods'.

27.38   Admin backend returns blank page after login (Option1)

Getting a blank page after trying to login to the Sphider-plus Admin backend is sometimes reported on 'Shared Hosting' server. As first attempt, you should delete the .htaccess file in folder …/admin/

27.39   Admin backend returns blank page after login (Option2)

Some server do not accept the 'Intrusion Detection System' (IDS), which is part of the Sphider-plus scripts. In order to disable the IDS it is necessary to manually edit the scripts
        …/admin/settings/backup/Sphider-plus_default-configuration.php
        …/settings/db1/conf_search1_.php
In both files the variable $use_ids needs to be edited to
        $use_ids = 0;

27.40 Admin backend returns blank page after login (Option3)

It might be a problem of session control. If the provider agrees, you should change the PHP session save path. Additional info and a detailed description is to be found at
http://www.biggz.com/info/webmaster/php-login-error.html

27.41   Why does v.3 of Sphider-plus use the MySQLi interface?

The **MySQLi Extension** (MySQL Improved) is a relational database driver used in the PHP programming language to provide an interface with MySQL databases. MySQLi is an improved version of the older PHP MySQL driver, offering various benefits.
        - An object-oriented interface
        - Support for prepared statements
        - Support for multiple statements
        - Support for transactions
        - Enhanced debugging support
        - Embedded server support
Not all of these features are already implemented in Sphider-plus v.3.2013a. But all important features will be

added to future releases.

Released together with PHP 5, the MySQLi extension aimed to replace the old, reliable MySQL extension, which has been available in PHP since the mid-90s. After almost a decade of development activities, some questionable features had crept into ext/mysql. The extension became difficult to maintain and the feature set of ext/mysql started to differ from that of the underlying MySQL client library.

The "i" in the name of the extension stands for: improved, interface, incomplete or whatever you want to call it. The new extension ext/mysqli supports all new features of the MySQL Server Version 4.1 and higher, for example Prepared Statements and support for Character Sets. Prepared statements are a great step ahead, especially nowadays when everybody is concerned about security.

The MySQLi extension has a procedural interface which is very similar to the ext/mysql interface to make porting as simple as possible and it offers an additional object oriented interface for those, who prefer the object oriented programming style.

The new features are the reason why all PHP scripts should be switched over to the ext/mysqli. Looking forward at PHP 5.5, the old MySQL interface will become deprecated. Thus a lot of PHP scripts will be out of time, creating a myriad of 'deprecated function' messages.

27.42   *Why to use multiple databases and table sets?*

First of all to reduce the amount of data stored in db. By separating the overall content of the site to be indexed into different categories, you will be enabled to store only the part of the content belonging to one category into one database. Usually the total content of a site is already separated in some folders, which could be named like
- about_us
- chapter
- research
- store
- etc.

Sphider-plus offers to use a 'URL Must include' option during index procedure. Thus each category (folder name) per database will be indexed separately.

Most important advantages:
- Reduced response time for query input, because of smaller db.
- Reduced indexing time.
- More comfortable for the admin of the search engine,
  as you may index only the content subset (category),
   which was modified since last index procedure.
- More comfortable for the user of the search vehicle, because while indexing only one category,
   all other content remains accessible for the search algorithm.

By the way, Sphider-plus also offers the option to use an unlimited amount of table sets in each database. Thus you may place all the above discussed categories into one database, by defining one table set for each cat.

28. Change log

28.1 Version 1.0 – 1.9

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 sub-menu '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 sub-menu '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'

72

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 2016-03-22, 03/22/2016 or 03/22/2016

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 check-box 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 sub-menu for bulk indexing of all the new sites
    that were added since last index/re-index.

Erase & Re-index.
    Additional item in Admin Sites sub-menu 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 sub-menu: 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 check-box, 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 Sphider-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 separates 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 sub-menus:
      - Admin / Clean / Clean Spider log
      - Admin / Statistics /Spider logs

Search in categories: Four bugs fixed.
    The following items are modified for proper function of Sphider'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).

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

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 dialogue to user language.
        Improved version by ^demon

Bug fixed in language depending user dialog.

75

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

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' dialogue
        - '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

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

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 sub-menu *'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

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

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 sub-menu '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.

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 lower-case 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.

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 check-box. 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 time-stamp.
      - 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 Sphider-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

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.

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/GeoIp.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.

Version 1.9

Release date: not published; only internal developing version.

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

Version 2.1

Release date: September 03, 2009
Build up with Sphider: v.1.3.4
Sphider-plus vs. original Sphider: 154 items worked out

New item in Admin settings:
> Perform a segmentation of Chinese and Korean text during index / re-index procedure.
>> Will divide phrases like 帽子和服装 into the base words 帽子 and 和 and 服装 ,
>> so that all will become searchable.
>> Valid for Chinese sites with charset: GB2312, GBK and GB18030
>> Valid for Korean sites with charset: EUC-KR and ISO10646-1933

New item in Admin setting:
> Index password protected sites.
> If enabled, Sphider-plus will index also .htaccess protected sites (basic authorization).
> Up to 3 different zones could be registered in Admin settings and will be indexed.

New options in Admin settings:
> - Index framesets
> - Index iframes
>> If enabled, both options will index html and image frames.
>> Not available for dynamically reloaded frames (e.g. by JavaScript).

New item in Admin setting:
> Enable to decode BBCode during index / re-index into standard HTML
>> If selected, code like
>>> **[url=http://abc.de/][b]abc.de[/b][/url]**
>> will be converted to
>>> **abc.de**

New item in Admin settings:
> Enable to decode entity coded sites into standard HTML characters.
>> If selected, entity coded text like **&#268apek** and **Döhl**
>> will be converted to **Čapek** and **Döhl**,

New options in Admin settings:
> - Use whitelist in order to enable index / re-index only those pages
>   that include **any** the words in whitelist
> - Use whitelist in order to enable index / re-index only those pages
>   that include **all** the words in whitelist

Improved 'Follow sitemap.xml' procedure:
> If <sitemapindex . . > is detected in a sitemap.xml file, and if multiple Sitemap files are available,
> Sphider-plus will process the secondary Sitemaps and extract all links for index / re-index.
> Also gzip-compressed files (Index Sitemap files as well as the Sitemap files) will be processed.

Improved index / re-index procedure:
> If charset of a site to be indexed is undetectable, because it is not HTML standard conform
> or missing HTML tag, the index procedure will no longer been interrupted.
> Preferred charset as defined in Admin settings will be used for the involved link.

Improved index / re-index procedure:
> If Sphider-plus is relocated by http 301 or 302, links found at the relocated site
> will also be followed.

For new sites, as per default the spider-depth is now set to 'full'.

Improved UTF-8 support:
        Conversion into UTF-8 charset now is obligatory.

Improved index and re-index procedures for Cyrillic and Greek languages to support upper and lower case characters.

Bug fixed that prevented to continue suspended index procedures.

'Continue suspended index procedure' enabled now also for 'Re-index' and 'Erase & Re-index'.

Improved search functions for search with wildcards and for strict search.

Improved category search:
        - Selected category name is highlighted in headline of result listing.
        - If activated in Admin setting, categories which would also deliver results
          are presented individual for each result link in the result listing.
        - If search in category is performed, sub-categories which would also deliver results
          are presented individual for each result link in the result listing.

If media search is enabled in Admin settings, text search with wildcards will also present media results.

Improved search utility:
        Queries with and without hyphen will deliver the same results,
        so that queries like 'make-up' and 'make up' do have equal rights.
        The same behaviour is performed for queries containing dots, commas and question marks.

Maximum length for site and link URL's to be indexed is now increased to 1024 characters.

Maximum length for link 'title' increased to 255 characters.

Code rewritten to cooperate with PHP 5.3.x

Error corrected de-language file. Thanks to Carl D. Erling

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

        Nearly all, because of PHP 5.3 compatibility.

In order to enable the two new items:
        - For new sites, as per default, the spider-depth is now set to 'full'.
        - URLs will be accepted for a length of up to 1024 characters.
this release requires the installation of new table sets for each database.

Version 2.2

Release date: December 22, 2009
Build up with Sphider: v.1.3.5
Sphider-plus vs. original Sphider: 173 items worked out

Improved multiple database support:
        Results may now be collected from more than one database.
        1 - 5 databases could be configured to fetch results for the common result listing.
        Valid for text and media search, all search modes, taking into account category selection.

Improved RSS and Atom feed index procedure. Including now also a validation for the well-formed XML.

Index support added for RDF feeds.
        For a complete list of indexed items, please notice the documentation chapter:
                RDF, RSD, RSS and Atom feeds

Additional item in Admin settings:
        Follow CDATA directives for feed content.

Additional item in Admin settings:
        Index 'Dublin Core' and other individually marked tags in RDF feeds.

Additional item in Admin settings:
        Follow the 'preferred (true/false)' directive in RSD feeds.

Detection of encoding (charset) added for XML and XHTML files.

New item in Admin settings:
        During index procedure, convert all kind of single quotes like ` ´ ' ' into standard quotes '

New item in Admin settings:
        Reduce queries which contain quotes to the basic word?
                This will deliver the same results for queries like:
                d'information = **information** or **dei'largi** = **largi**
        Results will be highlighted for the base word. Exclusive noun, pronoun, etc.
        Works for all kinds of single quotes.

New Admin setting:
        For queries containing numbers, search with wildcards.
                Useful to search for complex article numbers,
                if the user only knows a part of the complete item description

New Admin setting:
        Index ZIP compressed files and archives.
                Supports (X)HTML, XML and also compressed PDFs and other document files,
                as well as all kind of feeds, frames and iframes. Links found in the compressed
                files will be followed.

New option to sort the result listing:
        Sort by last indexed (date and time). To be defined in Admin settings.

New option to limit result listing:
        Define max. amount of results presented in result listing.
        To be defined in Admin settings, the count of results will be limited for text and media results.

New item in Admin settings:
  Use list of div ids to ignore the corresponding div content during index/re-index
  A common list of div id values is used to ignore parts of a page.
    Content between <div id='this_value'> and </div> will be ignored,
    however links in it are followed. Multiple and nested divs will be attended.
    Values in common list may end with a wildcard, so that 'menu*' will work for
      menu1, menu2, menu_left, etc.
    Usable also for external pages, if it is impossible to add the <!--sphider_noindex--> tags.
    Details in chapter Ignoring parts of a page by <div id='abc'>

New option
  Common *'URL Must include'* and *'URL must Not include'* rules, which are valid for all new sites, may
  be placed now in 2 files. The contents will be transferred to the corresponding option fields when
  calling 'Add site' in Admin menu. Individually de-selectable by checkbox.
  Details in chapter Must include / must not include string list

Log output suppressed, if the indexer is only redirected from
  http://www.abc.de   to   http://www.abc.de/index.html

Improved response for 'canonical' links. Back references to the calling page are ignored now.

New option for iframe indexing in Admin settings:
  Instead of calling page, remember the link to iframe directly.

New Admin setting:
  If found on different pages, index also duplicate media content.
    If activated, all images, audio and video stream will be presented in result listing.
    Otherwise only the first occurrence (page/link) will be presented.

Index procedure improved for dynamical created links.

New option:
  Suppress zero results in 'Most popular searches' as presented at the button of result listing.
  To be selected in Admin settings

Self test whether all sub-folders of ../admin/ are writeable. Otherwise a chmod 777 is performed. Malfunction
will cause warning messages.

Self test for up to date table structure of MySQL database.

Self test of PDF converter for correct addressing of the cconverter and correct conversion of a test-file.
Failures and malfunction will cause warning messages.

Updated PDF converter for non-Latin text like Arabic, Cyrillic, Chinese, Greece and Hebrew documents.
  With special thanks for the assistance of Daniel Richard, cnmss.fr

New algorithm to create the CAPTCHA in 'Add URL' form.

Function renamed from replace() to replace_string() in …/commonfuncs.php

Bug fixed that prevented highlighting of keywords in result listing, if full text was shorter than 250 characters
(as to be defined in Admin settings: 'Maximum length of page summary displayed in search results').

Bug fixed that prevented highlighting of keywords in result listing for Strict search (!query), if keyword was
found in position 0 of full text.

Bug fixed that caused direct jump to iframe instead of linking to the calling page, when activating the link in
result listing.

Bug fixed that prevented to display long URLs (> 70 characters) in Admin sites view.

Updated Dutch language file. Thanks to Danny von Berg.

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/auth_db.php
     /admin/configset.php
     /admin/db_config.php
     /admin/index_media.php
     /admin/install_tables.php
     /admin/messages.php
     /admin/spider.php
     /admin/spiderfuncs.php
     /converter/dummy.pdf
     /converter/feed_parser.php
     /converter/pdftotext.exe
     /converter/pdftotext.script
     /converter/xpdfrc
     /include/click_counter.php
     /include/commonfuncs.php
     /include/make_captcha.php
     /include/media_counter.php
     /include/searchfuncs.php
     /include/search_links.php
     /include/search_media.php
     /include/common/must_include.txt
     /include/common/must_not_include.txt
     /include/common/not_div.txt
     /include/images/no_fonts.jpg
     /languages/ all files
     /templates/all folders/hdline.jpg
     /templates/al foldfers/thisstyle.css

     /converter/rss2html.php + rss.html + rss_parser.php => no longer required

Attention: This version requires an updated set of tables in the MySQL database. In order to create the new tables, please run the 'Install all tables' for all databases in 'Database Management / Configure' menu.

Version 2.3

Release date:	April 23, 2010
Build up with Sphider:	v.1.3.5
Sphider-plus vs. original Sphider:	192 items worked out

In front of Sphider-plus version 2.2 the following items have been added / modified:

In order to ease customer's integration of Sphider-plus into existing sites, HTML templates are prepared for
- Search form
- Text results
- Media results
- Most popular queries
- etc.

New feature:
Split words into their basic parts, separated at each hyphen, dot or comma inside the words.
For example 'sphider-plus.eu' will be divided into the 3 keywords: sphider plus eu
As also the original word is stored as keyword, all 4 words become searchable.
Alternatively the separation only at hyphens is selectable in Admin settings.

New feature:
Allow indexing of other hosts with same domain name for links
found during indexing. Also ignore TLD, SLD and www.
For details see chapter Allow other hosts in same domain

New feature:
Allow indexing of other hosts with same domain name but only
if the found links are redirected. Also ignore TLD, SLD and www.
For details see chapter Allow other hosts in same domain

New feature:
Index sites and follow links containing none 'Basic Latin' and none ASCII characters
as part of their URL.

2 new features of sorting the result listing:
- Results of a promoted / featured domain will be displayed on top of the search result listing.
  As part of the Admin settings, a domain name or part of the name could be entered.
  All search results belonging to this domain will be placed on top of result listing.
- Pages containing a catchword will be displayed on top of the search result listing.
  As part of the Admin settings, the catchword could be entered.
For details see chapter Chronological order for result listing

New feature:
Index the "Description" Meta tag in HTML header.
To be activated in Admin settings.

New feature:
Index of media files enabled for those servers that do not offer all PHP functions for remote files.
Bypassed PHP functions are: fopen()   file_get_contents()   md5_file()

3 new features for command line operation:
- Erase & Re-index all sites (-eall)
- Index all new URLs in database which had not yet been indexed (-new)
- Re-index all meanwhile erased sites (-erased)

New feature:

In order to index XLS files, a converter for Exel files was developed. Implemented as PHP script, the converter needs no adoption to the Operating System.

New Admin setting:

Index RAR compressed files and archives.

Supports (X)HTML, XML and also compressed PDFs and other document files, as well as all kind of feeds, frames and iframes. Links found in the compressed files will be followed.

15 language specific stemming algorithms implemented. Individually selectable for:

Bulgarian, Chinese, Czech, Dutch, English, Finnish, French, German, Greek, Hungarian, Italian, Portuguese, Russian, Spanish and Swedish.
For details see chapter Word stemming

New Admin setting:

Activate/disable: Create 'sitemap.xml' file of each indexed site.

New Site option in Admin menu:

Erase/clean site-specific data from MySQL database and thumbnails folder for a selected site.

New Admin setting:

Re-index all meanwhile erased sites.

New Admin setting:

Show complete list during import and export of URLs, or hide output.

24 language specific common files holding a list of words to be ignored during index (stop words).
Added or updated for:

Arabic, Bengali, Bulgarian, Catalan, Czech, Danish, Dutch, English, Farsi, Finnish, French, Greek, German, Hindi, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish and Turkish.
In order to speed up index procedure, they are to be activated individually in Admin settings.

New feature:

Self test for all required PHP libraries and extensions. If Debug mode is enabled, the corresponding warning messages will be presented on top of the Settings menu.

Improved database 'Activate / Disable' menu:

If multiple sets of tables are available, because they have been created for a database before, you will be able to activate any of these table sets by selecting the corresponding prefix.

New Admin setting:

Define directory for templates (relative to root directory of Sphider-plus)

Search and open media files enabled now for media links with up to 1024 characters.

Input settings for database configuration menus are now enabled for values up to 255 characters.

'Clean resources' improved for index procedure.

In case of failure, only warning messages will be created and indexing will not be aborted.

The feature 'Clean resources' is added now also for search procedure.

Common activation in Admin settings for search and index procedure.

If debug mode is enabled, during index procedure, the new keywords are presented in alphabetic order now.

Follow 'robots.txt' directive enabled also for localhost applications.

Bug fixed that causes the result listing to be presented only in lower case characters. Now presented like the original title and full text of the indexed pages.

Some more small bugs eliminated.

Updated Admin dialog. Thanks to Ian Bucklar.

Additional language file for Hebrew. Thanks to Noam Bercovitz.

Updated Russian languge file. Thanks to Uttkirbek Abdullaev.

Updated Romanian language file. Thanks to Lionel Geo Mischie.

Involved files that have been modified / added for this release:
.htaccess
search.php
…/admin/admin.php
…/admin/configset.php
…/admin/db_activate.php
…/admin/db_config.php
…/admin/index_media.php
…/admin/install_tables.php
…/admin/messages.php
…/admin/real-log.php
…/admin/spider.php
…/admin/spiderfuncs.php
…/admin/url_backup.php
…/admin/getid3/module.audio-video.asf.php
…/converter/xls_reader.php
…/converter/xls2csv.exe not required any longer
…/include/commonfuncs.php
…/include/searchfuncs.php
…/include/search_media.php
…/include/common/all common_xyz.txt files
…/include/stemming/all files
…/language/he-language.php
…/languages/ro-language.php
…/languages/ru-language.php
…/settings/conf.php
…/templates/all folder/thisstyle.css
…/templates/html/all files

Version 2.4

Release date: July 03, 2010
Build up with Sphider: v.1.3.5
Sphider-plus vs. original Sphider: 204 items worked out

In front of Sphider-plus version 2.3 the following items have been added / modified

New feature:
        In order to reduce the time for indexing, multi-threaded indexing was implemented.
        As part of the Admin settings, 1-10 threads are to be activated.
        Available also for command line operation without limitation of the thread counts.
        For details see chapter [Multithreaded indexing](#)

New feature:
        Segmentation of Japanese phrases. To be activated in Admin settings.
        Segmenting 5.724 kanji (new, old and half width), hiragana,
        katakana and jinmeiyo Japanese character writing systems.
        Available for Japanese sites with charset: Shift_JIS, EUC-JP and UTF-8

New feature:
        Index CVS files. To be activated in Admin settings.
        Implemented as PHP script, the converter needs no adoption to the Operating System.

New feature:
        In order to index 'OpenDocument files, the corresponding converter were integrated.
        Implemented as PHP scripts, the converter need no adoption to the Operating System.
        To be activated separately for spreadsheet files (.ods) and text files (.odt) in Admin settings.

New command line options:
        - Erase content of database <-erase>
        - Set 'Last indexed' date and time to 0000 <-preall>

Improved support for Japanese coded sites (charset: SHIFT_JIS, EUC-JP and UTF-8).

Template design 'Pure' reduced to 'like Google'.

Improved search algorithm: significantly reduced search time.

Improved support for Greek language:
    1. Transliterate queries with Latin characters into their Greek equivalents.
       Will for example transform query input **alla** to find **ἀλλὰ** and **baptismatos** to find **βαπτίσματος**.
    2. Accept Greek queries containing vowels without accents.
       Query input of letter **α** will be valid also for
       ἀ, ἁ, ᾶ, ᾳ, ἄ, ἅ, ᾆ, ᾇ, ὰ, ά, ά, ᾀ, ᾁ, ᾂ, ᾃ, ᾄ, ᾅ, ᾲ, ᾳ, ᾰ, ᾱ, ὰ, ᾳ, ᾴ, ᾶ and ᾷ
       The same behavior for all other Greek vowels, as well as for the upper case vowels.

New options in 'Add site' and 'Edit site' menus:
        - Enter URL of individual Sitemap
         If Sitemap is not in root folder, the URL of the individual Sitemap could be entered.

New option to manipulate the result listing:
        For result sorting **'By URL names'** the number of results shown per domain is selectable.
        Offers result presentation similar to 'Like Google', but additionally offers a selectable count of links.

Search option 'More results from this domain' not only enabled for result sorting 'Like Google', but also for

'By URL names'.

Bug fixed that prevented correct interpretation of http 301 redirects.

Bug fixed that causes invalid results for multiple word queries, which contain numbers, like 'price 25 euro'.

Bug fixed that prevented highlighting of keywords, if found in position 0 of title or full text.

Bug fixed that prevented suppressing of 'Show result scores' in Admin settings.

Bug fixed that prevented to follow the option 'temporary ignore no-index'.

Additional language file for Japanese. Thanks to Sano Tomonori.

Updated Arabic language file. Thanks to Naif Alanazi.

Updated Italian language file. Thanks to Giorgio Nanni.

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

    …/search.php
    …/admin/ all files
    .../converter/ods_reader.php
    …/converter/odt_reader.php
    …/converter/dictionaries/jp_shiftJIS.dic
    …/converter/OpenDocumentSheet/ all files
    …/include/commonfuncs.php
    …/include/searchfuncs.php
    …/include/search_media.php
    …/include/suggest.php
    …/languages/ all files
    …/templates/html/all files
    …/templates/Pure/thisstyle.css
    …/templates/Slade/thisstyle.css
    …/templates/Sphider-plus/thisstyle.css

Attention: This version requires an updated set of database tables. In order to rescue the current sites in admin interface, backup the existing URL's by means of the admin menu 'Import / export URL list'. Also store the existing configuration file .../settings/conf.php outside of the Sphider-plus installation. Then copy all the new scripts of version 2.4 over the existing installation. Replace the new file .../settings/conf.php with the old file. In order to create the new set of tables, run the 'Install all tables' for all databases in 'Database Management / Configure' menu. Finally restore the URL list and re-index all.

Version 2.5

Release date: November 30, 2010
Build up with Sphider: v.1.3.5
Sphider-plus vs. original Sphider: 215 items worked out

New feature:
      Bound database.
      This option will delete all keyword relationships, exceeding a definable amount of query results.
      Beside the result cache, this option will significantly speed up the search procedure for huge
      databases.
      For details see chapter Bound database

New feature:
      In order to get indexed, user suggested sites optionally need a meta tag in header.
      Defined by the Sphider-plus admin during approval, the tag values could be used to verify the
      ownership of the suggested URLs, offer commercial dependencies, or perform a membership
      verification.
      For details see chapter User suggested sites

New feature:
      Intrusion Detection System (IDS) included to protect Sphider-plus against hacking attempts.
      It includes extensive regex rules to tags like
          XSS, SQLI, RCE, LFI, DT, CSRF, LDAP Injections, and DoS.
      Admin selectable, the IDS will block further user input, create a log-file,
      present a warning message, or even block any traffic of IP's known to be evil.
      For details see chapter Intrusion Detection System (IDS)

New feature
      Index only links and their link text.
      If activated in Admin settings, full text and media content will not be indexed,
      but only the link text (titles) of all links. Will also work for image links and their 'title' and 'alt' tags:
      title="this text", alternatively alt="this text". Result listing presents the (active) links with respect to the
      page at which they were found. If searching for a link text, the different search modes are available.

New feature in Admin settings:
      Add new domains found during index procedure to 'Approve Sites' table.
      To be activated in section 'General Settings', this option is available for those sites,
      having activated 'Spider can leave domain' in their options.

New feature in Admin sites menu:
      Index all the suspended.
      Will continue the index procedure for all the sites that are marked as 'Unfinished'.

New feature:
      Index media content with respect to frame/iframe position.
      To be activated in Admin settings, this option allows indexing media links, which are addressed
      as links relative to the frame/iframe position (folder).

Improved URL import/export function:
      Now all options of each site will be stored in backup file and re-imported.

New Admin setting:
>Clean query log during index / re-index and also for all erase functions.
>>Will reset all 'Search' statistics in Admin backend, as well as the
>>'Most popular search' tag-cloud / table at the bottom of result listing

New feature in Admin backend:
>When opening the Admin interface, a warning message will be presented about new
>suggested sites, waiting for approval. Working independent from the currently activated databases,
>so that suggestions of any databases will create an alert.

Dynamically created description tag in result page header. Build up with the titles of the most important results, presented on the different result pages.

Some small bugs eliminated.

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

>…/addurl.php
>…/search.php
>…/admin/admin.php
>…/admin/admin_footer.php
>…/admin/auth.php
>.../admin/configset.php
>…/admin/GeoIP.dat
>…/admin/index_media.php
>…/admin/install_tables.php
>…/admin/spider.php
>…/admin/spiderfuncs.php
>.../admin/url_backup.php
>.../include/commonfuncs.php
>…/include/ids_handler.php
>…/include/search_links.php
>.../include/searchfuncs.php
>…/include/search_media.php
>…/include/suggest.php
>…/include/swfobject.js
>…/include/tagcloud.swf
>…/include/IDS/all files and sub folder
>…/languages/all files
>.../templates/html/010_html_header.html
>.../templates/html/020_html_search-form.html
>.../templates/html/021_html_search-form.html
>.../templates/html/022_html_search-form.html
>…/templates/html/050_result-header.html
>…/templates/html/060_text-results.html
>…/templates/html/070_more-results.html
>…/templates/html/080_most_pop.html
>.../templates/html/081_3D_tag_cloud.html
>…/templates/html/100_all-media result-header.html
>…/templates/Pure/all files
>…/templates/Slade/thisstyle.css
>…/templates/Sphider-plus/thisstyle.css

Attention: This version requires an updated set of database tables. In order to create the new set of tables, run the 'Install all tables' for all databases in 'Database Management / Configure' menu.

Version 2.6

Release date of version 2.6:   2011-03-08
Build up with Sphider: v.1.3.5
Sphider-plus vs. original Sphider: 233 items worked out

New feature:
       Result output is available now also as an XML file. If requested in search.php script,
       the results will be presented as XML file in sub folder …/xml/
       For details see chapter <u>XML result output</u>

New feature:
       Index only parts of a page by <div id='abc'>
       If enabled in Admin settings, the values as defined in the list-file …/include/common/divs_use.txt
       will be used to index only the content between <div id='abc'> and </div> .
       This is the contrary function to
               Ignoring parts of a page by <div id='abc'>
       which is controlled by the list file …/include/common/divs_not.txt
       For details see chapter <u>Indexing only parts of a page by <div id='abc'></u>

New feature:
       Individual (Admin) settings for each database and each set of tables.
       Automatically activated by selecting any of the 5 databases and any set of tables in the dbs.

New feature in Admin backend:
       Search functions are available now in order to query for:
               - sites
               - links
               - keywords
               - categories

New Admin setting:
       Define number of sites shown per page in Admin backend (pagination 10, 20, 30, 50, 100).
       Used for:
               - Sites view
               - Approve URLs
               - Banned domains
               - Statistic results

Improved Admin settings:
       The table in Admin backend 'Sites' view could be sorted:
               - by index-date, latest on top
               - by index-date, oldest on top
               - by title as personally defined when adding the site
               - in alphabetic order (URL)

New feature:
       Additional option to Re-index only the sites that are currently shown in 'Sites' view.
       By selecting (pagination) 10, 20, 30, 50 or 100 sites per page, it is possible
       to re-index only the URLs presented on page 1, and later on those URLs of page 2 etc.

New Admin setting:
       Obligatory use the preferred charset as defined in 'General Settings' for indexing.
       The corresponding option is to be found in sites 'Edit' option, so that individual sites
       could be influenced. If activated, the header information like

<meta http-equiv='Content-Type' content='text/html; charset=windows-1256 />
of the site to be indexed, will be overwritten by the preferred charset.

New Admin setting:
Separated activation of debug mode for Admin backend and User interface.

New Admin setting:
Do not index the full text. If activated, only the page 'Title', the 'Keywords' Meta tag,
as well as the 'Description' Meta tag will be indexed.
Never the less, links found in full text will be followed.

New feature:
Queries containing ' && ' will overwrite the advanced search settings to AND.
Queries containing ' || ' will overwrite the advanced search settings to OR.

Complete redesign of all search files for easier integration of Sphider-plus scripts into an existing HTML site.
With special thanks for the suggestions, ideas and the participation of Carl Erling
http://www.tba-berlin.de

New Admin setting:
Define whether the 'Search form' and the 'Result listing' of Sphider-plus is embedded into an existing
page HTML layout and design, or whether it is used as an independent page.
For details see chapter Integration of Sphider-plus into existing sites

New Admin setting:
Define the name of the search script in root folder of Sphider-plus (default: search.php).

Separated style sheet files are now included for Admin backend and for the User interface. This enables to
individualize the User style sheet without destroying the Admin design.
For details see chapter Integration of Sphider-plus into existing sites

Improved 'Did you mean' algorithm. Now searching for a wider range of potential keywords.

Break character (­) inside of words will now be ignored, so that the complete word becomes indexed
and searchable.

Output of Intrusion Detection System now is presented with respect to the currently activated template
design.

Improved backup for 'Settings'. The name of the backup file will now consist of:
- Date and time
- Number of database
- Name of table prefix
Consequently, all details for the allocation of the backup files are available now.

Automatically add "http://" for new sites in Admin backend.

Bug fixed, which prevented limiting of search results. Occurred, if multiple databases were selected to deliver
search results.

Bug fixed that created multiple wildcards, if searching for numbers.

Bug fixed that suppressed the HTML header in link search.

Bug fixed, which has overwritten the Admin setting
"Show x results per page in result listing"
caused in …/include/searchfuncs.php by the row
if ($all_wild && $greek != '1') $max_hits = '99';

Bug fixed to prevent a blank display on first opening the Admin backend
(if mb_string functions are not available).

Bug fixed, which causes invalid URL parsing for relative links with ../../ indication.

Bug fixed that prevented domain search for localhost applications

Bug fixed to prevent invalid character size for 'Like Google' result listing

Bug fixed in database 'Backup & Restore' function.

Some additional small bugs removed.

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

      …/addurl.php
      …/search.php
      …/search_ini.php
      …/admin/admin.php
      …/admin/admin_header.php
      …/admin/admin_footer.php
      …/admin/auth.php
      …/admin/configset.php
      …/admin/db_main.php
      .../admin/GeoIP.dat
      …/admin/install_tables.php
      …/admin/messages.php
      …/admin/real_get.php
      …/admin/real_log.php
      …/admin/settings/backup/all files
      …/admin/setting_backup.php
      …/admin/spider.php
      …/admin/spiderfuncs.php
      …/admin/url_backup.php
      …/include/commonfuncs.php
      …/include/media_counter.php
      …/include/search_10.php
      …/include/search_20.php
      …/include/search_30.php
      …/include/search_40.php
      …/include/search_50.php
      …/include/search_links.php
      …/include/search_media.php
      …/include/searchfuncs.php
      …/include/show_id3.php
      …/include/suggest.php
      …/include/common/audio.txt
      …/include/common/divs.txt
      …/include/IDS/Config/Config.ini.php
      …/settings/all files and folders
      …/templates/html/all files
      …/templates/Pure/adminstyle.css
      …/templates/Pure/userstyle.css
      …/templates/Slade/adminstyle.css
      …/templates/Slade/userstyle.css
      …/templates/Sphider-plus/adminstyle.css
      …/templates/Sphider-plus/userstyle.css

Attention: This version requires an updated set of database tables. In order to create the new set of tables, run the 'Install all tables' for all databases in 'Database Management / Configure' menu.

Version 2.6a

Release date: March 14, 2011

In front of version 2.6 the following modifications had been added:

> - Bug fixed that prevented renaming of the default search script.
> - Bug fixed in multi-threaded indexing.
> - Bug fixed to prevent creation of duplicate sub-folders in …/admin/
> - Media search enabled for multiple database support.
> - User debug mode enabled for link search.
> - Indexing of https:// sites enabled.

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

> …/search.php
> …/admin/admin.php
> …/admin/spider.php
> …/admin/url_backup.php
> :::/admin/settings/backup/Sphider-plus_default-configuration.php
> …/include/categoryfuncs.php
> …/include/search_links.php
> …/include/search_media.php
> …/templates/html/ all files

Version 2.6b

Release date: March 25, 2011

In front of version 2.6a the following modifications had been added:

New Admin setting:
> 'Protect the .../admin/ folder by means of a .htaccess file'
> If activated, and if the .htaccess file is not yet available, the script will automatically detect the IP of the admin and create your .htaccess file in the ../admin/ folder.
> If the setting is deactivated (check-box), the .htaccess file will be deleted by the script, so that afterwards the admin folder is freed again for IP independent access.

New feature:
> Result listings 'By URL names' and 'Like Google' are sorted in alphabetic order now.

New feature:
> The words specified in common list (to be ignored during index procedure) are no longer interpreted case sensitive. Consequently words like 'Sphider' and 'sphider' need not to be included both.

Improved calculation of keyword weighting. Now working independent from lower case and/or upper case written text.

> - Bug fixed for applications not using the advanced search options (</form> and </div> missing)
> - Bug fixed for embedded application.
> - Bug fixed for result sorting (By URL names).
> - Bug fixed in 'More results from URL'.
> - Bug fixed for 'Use commonlist for words to be ignored during index / re-index'.
> - Bug fixed in 'Use blacklist to prevent index of pages'.

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

> …/search_ini.php
> …/admin/configset.php
> …/admin/spider.php
> …/admin/spiderfuncs.php
> …admin/thumbs/.htaccess
> …/include/commonfuncs.php
> …/include/search_10.php
> …/include/search_40.php
> …/include/searchfuncs.php
> …/include/common/common_de.txt
> …/templates/html/020_search-form.html
> …/templates/html/060_text_results.html
> …/templates/html/090_footer.html

Version 2.7

Version   2.7
Release date: October 18, 2011
Sphider-plus vs. original Sphider: 251 items worked out

New indexing feature:
 Re-indexing could be performed periodically. Once started, this mode will automatically re-index
 all sites periodically. The time interval is Admin selectable for
 3 hours, 12 hours, 1 day, 1 week or 1 month.
 Also the count of periodically performed re-indexing procedures is Admin selectable.
 For details see chapter Periodical Re-indexing

New feature for media search:
 Find media results not only by media 'tile', but also by EXIF and ID3 info
 To be activated in Admin backend.

New option in Admin settings:
 "Use string list of 'URL Must Not include' also to prevent erasing of involved URLs"
 If activated, also erasing of the involved sites and pages (links) will be prevented.
 In order to erase all sites and all pages completely, it might become necessary to uncheck this option

New option in Admin settings:
 "Limit the amount of media results presented together with text results"
 Defined as maximum count of media results per page. The image results are counted separately
 from audio + video streams.

Improved search form. Now offering separated search buttons for 'text' and 'media' queries, as well as a button for combined search.

Improved search procedure for combined search of text and media, in order to speed up the search procedure.

Improved delete function in Admin backend:
 If a site is deleted from the admin backend, now also all keyword relationships to that site
 are withdrawn from the database. Site-specific links, category relationships and other dependencies,
 like registrations in temporary and pending tables, had been already observed before.

Improved Admin search function:
 Searching for 'Sites', the result listing now will present also the 'Options' button to select
  Edit, Re-Index, Erase & Re-index, Erase, Delete, Pages, Browse and Statistics

Improved index procedure for media indexing:
 No longer accepting dead links. In order to become indexed, the media file must be present.

Improved index procedure to speed up indexing.

Improved index procedure to cooperate with those servers that do not accept basic authentication strings.

Improved index procedure:
 If the 'User Agent String' as defined in Sphider-plus Admin backend is not accepted by the site
 to be indexed, Sphider-plus will use a standard browser HTTP_USER_AGENT to connect to the site.

New algorithm to delete the content of HTML and PHP tags
 No longer using the PHP function strip_tags(); now also unclosed and invalid tags will be observed
 during index procedure. As result, also the text following an unclosed or invalid tag will become
 indexed. This part of the full text was cut off by the PHP function strip_tags().

Modified index procedure:
 The instructions 'RESET QUERY CACHE' and 'FLUSH TABLE' will only be used, if the following
 Admin setting is activated:
  'Clean resources during index / re-index and also for search function'

Improved 'Settings' interface in Admin backend. After pressing 'Save', now additionally presenting the eventually existing dependencies and the necessarily modified settings.

Improved 'Approve sites' menu. If categories are available, as per default the new sites are placed in category 'none'.

Improved search function:
 If in admin backend the option
 'Delete special characters like dots, commas, exclamation and question marks etc. as part of words'
 is activated, also the search query will be cleaned from secondary characters.
 Consequently queries like 'book: kellner' and 'kellner, rolf' will no longer fail.
 This modification will not be active for 'Phrase' search.

Improved search function for queries containing hyphens.

Improved HTML files. Now loading faster the search form.

Improved display output for main categories.

Improved 'addurl' form. Now accepting URL's without www.

Improved 'addurl' form. If categories are available, as per default the new suggested site will be placed in category 'none'.

Common word list added for Chinese language. With thanks to Jame Sian 孙 春淦

Updated framework for ID3 and EXIF extraction during media indexing.

Updated GeoIP database, used to provide the IP of the search user.

Updated IDS configuration file, default filter and converter.

Updated language files for Czech and Slovenian language. Thanks to Peter Krupa.

Updated suffix list, holding all the file suffixes, which will not to be indexed.

Bug fixed in Database backup script that prevented correct storage of index-date.

Bug fixed in suggest framework to enable suggestions for queries using main-categories.

Bugs fixed, which prevented disabling the IDS framework for 'Search User' and 'Suggest User'.

Bug fixed in option "Ignoring parts of a page by <div id='abc'>" for multiple nested divs.

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

     …/addurl.php
     …/search_ini.php
     …/admin/admin.php
     …/admin/admin_header.php
     …/admin/admin_search.php
     …/admin/auto_index.php
     …/admin/db_common.php
     …/admin/configset.php
     …/admin/index_media.php
     …/admin/messages.php
     …/admin/spider.php
     …/admin/spiderfuncs.php
     …/admin/url_backup.php
     …/admin/getid3/all files
     …/include/commonfuncs.php
     …/include/search_10.php
     …/include/search_40.php
     …/include/search_media.php
     …/include/searchfuncs.php
     …/include/suggest.php
     …/include/common/common_cn.txt
     …/include/common/suffix.txt
     …/languages/all files
     …/templates/010_html-header.html
     …/templates/011_html-header.html
     …/templates/html/020_search-form.html
     …/templates/html/040_category_tree.html
     …/templates/html/060_text_results.html
     …/templates/html/110_media-only header.html
     …/templates/html/200_no media found.html
     …/templates/html/sphider-plus.ico

Version 2.8

Version	2.8
Release date:	March 31, 2012
Sphider-plus vs. original Sphider:	264 items worked out

New feature:
> Same results for queries typed with pure vowels or with accents.
> Will deliver the same results for queries like: caf**e** and caf**é.**
> To be activated in Admin backend.

New feature for AND and OR search:
> If the length of the text extract in result listing is too short to highlight all search words, additional text extract are build up to highlight all search words of the total query.

New feature:
> Besides bulk Re-indexing of all sites, the periodical Re-indexer is now available also site specific.
> To be activated individual in "Options" menu of each site.

New feature:
> Bound the length of full text indexed at each page. Will limit the indexed keywords to be extracted only from the first part of the full text, if set to values like 500 or 1000.
> [0 = unlimited text becomes indexed].

New option to be set in Admin backend:
> Block all queries sent by Meta search engines like Google, MSN, Amazon, etc
> For details see chapter Prevent queries from Meta search engines and crawler known to be evil

New option to be set in Admin backend:
> Block all queries sent by crawler known to be evil.
> For details see chapter Prevent queries from Meta search engines and crawler known to be evil

New option to be set in Admin backend:
> Delete special characters inside of words. Underscores, hyphens and symbols like ' "
> etc. as part of words are deleted. So only the pure words will be indexed.

New feature:
> The indexer could be interrupted periodically after indexing a predefined count of pages (links).
> Configurable in Admin settings.

New option to be activated in Admin backend:
> Convert all kind of double quotes like **"** and **"** into standard quotes **"**

New option to be activated/disabled in Admin backend:
> Show time elapsed (to fetch the results) in result header.

New option to be activated/disabled in Admin backend:
> In result listing show the actual result number of each result.

New option to be activated/disabled in Admin backend:
> In result listing show the URL of each result in a separate row.

New option to be activated/disabled in Admin backend:
> Index the 'Title' Meta tag in HTML header

New option to be selected in Admin backend :
Define the default chronological order for media result listing
- By title (alphabetic)
- By image size
- By 'Last queried'
- By 'Most popular'
- By file suffix

New option to be activated in Admin backend :
Limit the amount of media results presented together with text results.
Defined as maximum count of media results per page. The image results are counted separately from audio + video streams.

New method of thumbnail storage:
The thumbnails are no longer stored in a sub folder of the Sphider-plus installation, but now are stored in database table "media" in field "thumbnail".

Improved media search:
AND, OR and TOLERANT modes are now selectable for media search, while the PHRASE mode will be interpreted as an AND search.

Improved media search:
Henceforward file name as well as the title will be queried to find media results.

New options to be defined in Admin backend:
The following basic indexing options are globally definable for all sites:
- Spidering depth: Full Index or folder depth definition
- Spider can leave domain
- Use preferred charset for indexing
Afterwards individual settings could be performed site specific in the advanced option of each site URL. The global settings will also be used for suggested sites (addurl form).

New option in Admin 'Clear' menu:
Clear all entries in Addurl' table.

New option in Admin 'Clear' menu:
Clear all entries in 'Banned' table.

Improved option:
Ignoring parts of a page defined by <div id='abc'>
now is working alternately also for <div class='abc'>
Besides the string list in divs_not.txt file, the file now alternatively may contain regexp patterns.

Improved option:
Indexing only parts of a page defined by <div id='abc'>
now is working alternately also for <div class='abc'>
Besides the string list in divs_use.txt file, the file now alternatively may contain regexp patterns.

Presenting of multiple hits in result listing enabled now also for strict search.

Language files added for Norwegian (nynorsk and bokmål). Thanks to Geir Kleiveland.

White- and blacklist, as well as the other lists in …/include/common/ folder
now are tolerating (ignoring) blank rows.

Improved index procedure, now also accepting links containing "blank" characters.

Improved "Erase & Re-index all" function. Now deleting also the "pending" and "temp" tables.

Support for Greek language totally rewritten. Now accepting Latin characters for old and new Greek transcription.

For details see chapter [Greek language support](#)

Improved parser for RSS v.2.0 feeds.

Bug fixed in index procedure, which prevented correct indexing of text placed behind multiple tabs.

Bug fixed in search function for searching in multiple databases.

Bug fixed in result listing when presenting multiple hits per page.

Some more small bugs killed.

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

> …/admin/admin.php
> …/admin/admin_header.php
> …/admin/auth.php
> …/admin/auto_index.php
> …/admin/configset.php
> .../admin/db_copy.php
> .../admin/db_main.php
> …/admin/index_media.php
> …/admin/install_tables.php
> …/admin/messages.php
> …/admin/real_log.php
> …/admin/spider.php
> …/admin/spiderfuncs.php
> …/converter/feed_parser.php
> …/include/click_counter.php
> …/include/commonfuncs.php
> …/include/make_captcha.php
> …/include/media_counter.php
> …/include/search_10.php
> …/include/search_40.php
> …/include/searchfuncs.php
> …/include/search_media.php
> …/include/show_id3.php
> …/include/suggest.php
> …/include/common/black_ips.txt
> …/include/common/black_uas.txt
> …/languages/nn-language.php
> …/languages/no-language.php
> …/templates/html/010_html_header.html
> …/templates/html/011_html_header.html
> …/templates/html/020_search-form.html
> …/templates/html/021_search-form.html
> …/templates/html/022_search-form.html
> …/templates/html/050_result-header.html
> …/templates/html/060_text-results.html
> …/templates/html/070_more-results.html
> …/templates/html/090_footer.html
> …/templates/html/120_media-only results.html
> …/templates/html/140_image-results.html

Attention: This version requires an updated set of database tables. It is strongly recommended to follow the instructions as described in chapter [Updating from 2.x to 2.y](#) for the actual version 2.8

Version 2.9

Version	2.9
Release date:	November 01, 2012
Sphider-plus vs. original Sphider:	284 items worked out

New feature:
> Support for non-ASCII URLs using 'Internationalized Domain Names' (IDN).
> It is a standard described in RFC 3490, RFC 3491 and RFC 3492.
> If activated, internationalized domain names like 'http://президент.рф/' and 'http://müller.de/'
> will be accepted as new sites in Admin backend, as well as in User's addurl form.

New feature:
> Support Punycode URLs like   http://xn--90aoqlh7c4a.xn--d1abbgf6aiiy.xn--p1ai/
> Converted into the readable form   http://события.президент.рф/
> To be activated in Admin settings.

New feature:
> Besides the usual HTML elements <element> , also delete from full text all those HTML elements,
> which are defined like < element >
> To be activated in Admin settings.

New feature:
> Index only parts of a page, defined by <element> . . . </element>
> This feature is foreseen to cooperate with the new HTML5 elements like
>   section, nav, aside, hgroup, article, header, footer, etc
> If enabled in Admin settings, the values as defined in the list-file
> …/include/common/elements_use.txt will be used to index only the page content between
> <element> . . . </element>
> For details see chapter Indexing only parts of a page defined by <element> . . . . </element>

New feature:
> Ignore parts of a page, defined by <element> . . . </element>
> This feature is foreseen to cooperate with the new HTML5 elements like
>   section, nav, aside, hgroup, article, header, footer, etc
> If enabled in Admin settings, the values as defined in the list-file
> …/include/common/elements_not.txt will be used to remove the content between
> <element> . . . </element> from the page content.
> This is the contrary function to 'Index only parts of a page, defined by <element> . . . </element>'
> For details see chapter Ignoring parts of a page defined by <element> . . . . <element>

New feature:
> Index only files and documents with defined suffix :
> If activated, all pages of the site will be searched for links,
> but only files with suffixes as defined in the docs list will be indexed.
> For details see chapter Index only files and documents with defined suffix

New feature:
> 1. Perform a WHOIS check for sites waiting for approval in Admin backend.
> 2. Perform a WHOIS check for suggested URLs direct in the addurl form,
>   so that invalid URLs will automatically be rejected.
> For both tests a basic list of WHOIS servers for the generic top level domains
> and some important country codes (supporting 30 suffixes),
> or an extended list (supporting 155 suffixes) are selectable.

New option to be activated in Admin backend:
> Crawler can leave domain during index procedure, but only for canonical links.
> Only the canonical link will be indexed, but links found there will be ignored.

New feature:

      Obey the 'refresh' meta tags as part of HTML headers.

      Now following the redirection and delayed indexing.

New option:

      Support UTF-16 coded sites. Will convert UTF-16 coded sites into UTF-8.

      To be activated in Admin settings

New option:

      For index procedure always use the standard Firefox HTTP_USER_AGENT string

      and ignore the individual defined Sphider-plus string. To be activated in Admin backend.

New feature

      Follow redirections, which are invoked by JavaScript, when sent as HTTP content.

      Will obey directives like:

                              <SCRIPT language="javascript">window.location="mp.php?mcv=59";</SCRIPT>

New feature:

      Follow URL redirections caused by HTTP 301, 302, 303 and 307 status codes.

New feature:

      Separated PDF converter supplied for 32 and 64 bit Operating Systems.

      For details, please notice chapter [PDF converter for Linux/UNIX systems](#)

New feature:

      Follow links placed in JavaScript files. Will detect and follow links like

            document.write(' **All news 2012** ');

      Also the complete content of

            document.write(**this text in all rows'**);

      will be indexed and stored as keywords in db.

New feature:

      Now indexing also sites, which do send a obligatory request for a cookie, to be set by the crawler.

New feature:

      In order to reduce transmission time, the crawler now requests gzip-formatted data transfer

      from the remote server for the URL to be indexed.

New option:

      In order to convert the text into UTF-8, use the charset definition as supplied via HTTP by the client server.

      If this option is not activated in Admin Settings, the charset will be extracted from the header of the files to be indexed. If not found, like in PDF documents, the preferred charset will be used.

New option:

      Delete duplicate parts of the URL path found in the indexed page URL and the new links.

      Unfortunately some CMS seem to be unable to build up a correct path for relative links.

      If activated in Admin backend, these duplicate parts of the path will be deleted from the link URL.

      Should be activated only, if sites are indexed created by dedicated CMS.

New feature:

      Show summary of actually active User database at the bottom of result listing.

      To be activated in Admin backend, the count of sites, categories, page links and keywords are displayed.

New feature:

      Automatically deleting invalid URLs from Admin 'Sites' view.

Improved 'Add site' function in Admin backend.

Now treating URLs with and without 'www' as equal, and excluding them as duplicate sites.
Improved image indexing procedure
Now also indexing phpBB images, linked by php command files.

New option
Suppress the file suffix from image file names for indexing.

Improved media indexing procedure
In case of missing title tag, now the alt tag is used to define the name of the media.
In case that also the alt tag is missing, the file name will be used as keyword.

Improved "banned domain" management
Now holding name and suffix of the banned domains, and no longer the URLs.

Improved index procedure
Now ignoring links that try to link to the calling URI (self back linking).

Improved link detection for relative links, which are to be found in full text.

Improved input protection against SQL injections.

Improved Admin statistics
Now providing also the IP, country code and country name for
- Search log
- Most popular searches
- Most popular page links
- Most popular media links

Updated GeoIP database, used to provide the IP, CC and country name for the Admin statistics.
Now also supporting Ipv6 URLs.

Support on Windows systems temporary removed for ppt files, as the converter causes failures on large PowerPoint documents.

Bug fixed, which prevented category selection without activating the "Advanced search form" option.

Bug fixed that caused invalid URL encoding in result listing.

Bug fixed causing the error output "Unknown column 'naame' in field list" during media indexing.

Bug fixed that caused MySQL warning messages during index procedure at some older MySQL versions, if the URL to be indexed contained blank characters.

Bug fixed, which caused invalid URL creation for relative links containing a file name and/or query.

Bug fixed in option 'Crawler can leave domain'.

Bug fixed in option 'Use list of div ids to ignore the div content during index/re-index'.

Bug fixed in option 'Enable to decode entity coded sites into standard HTML characters'.

Bug fixed in 'addurl' form, which prevented input of words containing accents in 'title' and 'description' fields.

Some additional small bugs killed.

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

…/addurl.php
…/admin/admin.php
…/admin/admin_header.php
…/admin/admin_search.php
…/admin/auth.php
…/admin/auth_bypass.php
…/admin/auth_db.php
…/admin/configset.php
…/admin/db_activate.php
…/admin/db_config.php
…/admin/db_main.php
…/admin/geoip.php
…/admin/GeoIPv6.dat
…/admin/http.php
…/admin/index_media.php
…/admin/install_tables.php
…/admin/messages.php
…/admin/spider.php
…/admin/spiderfuncs.php
…/admin/url_backup.php
…/converter/feed_parser.php
.../converter/pdftotext32.script
.../converter/pdftotext64.script
…/include/click_counter.php
…/include/commonfuncs.php
…/include/domain_whois.php
…/include/idna_converter.php
…/include/media_counter.php
…/include/search_10.php
…/include/search_40.php
…/include/search_50.php
…/include/search_media.php
…/include/searchfuncs.php
…/include/suggest.php
…/include/common/docs.txt
…/languages/ all files
…/templates/html/020_search-form.html
…/templates/html/090_footer.html
…/templates/html/091_footer.html

Attention: This version requires an updated set of database tables. It is strongly recommended to follow the instructions as described in chapter <u>Updating from 2.x to 2.y</u> for version 2.9

Version 3.2013a

Version: 3.2013a
Release date: July 31, 2013
Sphider-plus vs. original Sphider: 314 items worked out

Honour:
    This version of Sphider-plus is dedicated to Anton Cygankov. For all the month, he followed the
    development with an ongoing testing. Verifying especially the index procedure with the amount of
    593(!) URLs and reviewing all the bugs, I intermediately implemented. Thank you very much for this
    big effort. I wouldn't be able to get up to the current status of development, without your enormous
    support.
    It took several months to come up with the v.3 release, but together with the SQLi support,
    it seems to be a reasonable step into the future of this search engine.

New feature:
    Index DOCX files. To be activated in Admin settings.
    Implemented as PHP script, the converter needs no adoption to the Operating System.

New feature:
    Index XLSX files. To be activated in Admin settings.
    Implemented as PHP script, the converter needs no adoption to the Operating System.

New feature:
    Index only preferred sites. Level depended re-index of only those URLs, containing the according
    level. For details please notice the chapter Preferred indexing

New option:
    Admin's 'Sites' table sorted by index priority.

New feature:
    Create a thumbnail of all Internet URLs during index procedure.
    Will be presented as part of the text result listing for each link.
    To be activated in Admin backend.
    For details please notice the chapter Create thumbnails during index procedure

New feature:
    If the blacklist is met too often, automatically abort the indexation of the regarding site.
    Defined to a count of 20.

New option:
    Check correct converting of content into UTF-8
    Will detect invalid charset definitions in Meta tags of HTML header,
    or invalid charset definition supplied via HTTP by the client server.
    If an invalid charset is detected, the index procedure will be aborted for the regarding link.

New feature:
    The addurl form now will only store domain name and TLD. Something like 'sphider-plus.eu'
    Thus, www. and any sub folder of the suggested URL will be ignored.

New feature:
    Ignore the content of style="display:none" in div elements. Something like:
    <div style="display:none">ignore_this_content</div>

New feature:

In order to enable immediate query input, auto focus is set to the search form.

New suggest framework.
The auto-complete feature of Sphider-plus is now based on the JavaScript library jQuery
For details please notice the chapter Suggest framework

New feature:
Separate search fields for text and media queries. Consequently also separate suggestions
will be offered. To be activated in Admin 'Settings'.

New feature:
Restrict the search results by means of up to 5 categories simultaneously.
Import and export of URLs with multiple category definitions assigned to each site.
For details please notice Parallel structure of search in categories.

New feature:
Now indexing also site URLs containing the https scheme.

Improved index procedure:
- Now treating link URLs with and without 'www' as equal, and excluding them as duplicate pages.
- Linking in it selves caused by HTTP 301/302/307 redirections are intercepted.
  Thus infinite indexation is prevented.
- Multiple attempts to redirect in it selves will force Sphider-plus to abort the index procedure
  for the involved site.

New option in Admin 'Settings' menu:
Define count of redirections followed for each link (1-9) while indexing.

New options in Admin 'Settings' menu:
Follow URL redirections, which are invoked by JavaScript like
<sript . . . 'window.location.replace . . . . '
<script . . . var cURL = ...'
<script . . . window.location = ... AND " + location.host + "
ând several other script directives

New options in Admin 'Settings' menu:
Follow URL redirections, which are invoked by body tags like
<BODY onLoad = "parent.location = 'home.asp'">
'HTTP-EQUIV= . . refresh . . content= . . .'
and several other tags

New option in Admin 'Settings' menu:
Obey refresh delay directives, placed in meta tags like
<meta http-equiv=**"refresh" content="180**;url=http://www.moodys.com.ar">

New option in Admin 'Settings' menu:
Do not index comment parts <!-- this text --> and scripts outside the HTML tags

New option in Admin 'Settings' menu:
If not already exist, add a final slash to the path for all detected links.
If a file name exists as part of the path, this option will be bypassed.
Also, if the http request for the main URL is only excepted without slash,
this option will not be obeyed.

New option in Admin 'Settings' menu:
Convert all link URLs to lower case characters.

New option in Admin 'Settings' menu:
Convert all link URLs found during indexation into UTF-8
Will convert URLs like:
/3v/catalog/%C1%E0%E2%E0%F0%E8%FF+%E8%E7%F0%E0%E7%E5%F6/

into:
/3v/catalog/Бавария+изразец/

Improved link detection:
- Invalid URLs containing duplicate slashes in its path will be ignored.
- The following links are followed now:
<script>window.document.location ="/this.path";</script>
<script>window.document.location.href="/this.path";</script>
<script>window.location.replace("/this.path")</script>
<script>"https|http this URL"</script>
<body onload "/this.path">
and several other.

New option in Admin backend 'Clean' menu:
Truncate all tables in database.

Improved 'NOHOST' detection during index procedure:
Now trying 5 times to get in contact with the server.
Each attempt is performed by 2 different HTTP requests.

Improved 'Add site' function in Admin backend.
Now treating URLs with the scheme 'http' and 'https' as equal, and excluding them as duplicate sites.

Support added for Windows-31J (CP932) charset as extension of Shift JIS.
(CP932 contains standard 7-bit ASCII codes, and Japanese characters are indicated by the high bit set to 1)

UTF-8 support implemented for media titles, file names and ID-3 tags.

SQLi connector implemented between PHP and a MySQL database. Performed by OOP.

Bug fixed in option: Do not index the full text.

Bug fixed for URLs containing CP1252 coded paths.

Bug fixed in detection of www/non www links. Now preventing double indexing.

Bug fixed in 'Strip session ids'

Bug fixed in Korean word segmentation

Some small bugs killed.

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

As the SQLi connector was implemented between PHP and a MySQL database,
nearly all scripts are renewed.
It is strongly recommended to perform a fresh installation for this version.

Version 3.2013b

Version: 3.2013b
Release date: 2013-09-19
Sphider-plus vs. original Sphider: 316 items worked out

New option in Admin 'Settings' menu:
    Use words in white- and blacklist as trimmed values.
    If activated, for example the word ' cailis' in blacklist will match with words like 'specialist'. Instead,
    if **not** activated, the leading blank of ' cailis' will prevent to match the word 'specialist' in full text.

Improved index procedure for URLs, containing quotes.

Improved black list comparison.

Improved xlsx converter.

Bug fixed in option 'Use list of div ids to ignore the complete div content during index/re-index'.

Bug fixed in 'Most Popular search' table at the bottom of result listing.

Bug fixed while indexing <!--sphider_noindex-->

Bug fixed in result listing, while sorted 'Like Google'.

Debug info removed from index procedure.

MySQLi error messages now are correctly assigned to the according SQL queries.

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

    …/admin/admin.php
    …/admin/auto_index.php
    …/admin/configset.php
    …/admin/spider.php
    …/admin/spiderfuncs.php
    …/converter/xlsx_reader.php
    …/include/commonfuncs.php
    .../include/search_10.php
    .../include/search_40.php
    …/include/searchfuncs.php
    …/templates/html/060_text-results.html
    …/templates/html/080_most_pop.html

Version 3.2014a

Version: 3.2014a
Release date: 2014.03.03
Sphider-plus vs. original Sphider: 321 items worked out

New feature:
      Equalize text results for query terms with and without ligatures.
      Worked out for Latin ligatures in Unicode (Latin-derived alphabets) and also ligatures
      used only in phonetic transcription, but not taking into consideration medieval ligatures.
      Will present the same results for:
            cœur and **coeur**
            dʒ = dezh = ʤ
      To be activated in Admin settings.

New feature:
      Delete multiple used letters from page content.
      Will delete decoratively used duplicates, like *****, -----, =====, etc.
      To be activated in Admin settings.

New feature:
      Query strictly for search results. Available for AND, OR, and PHRASE search modus.
      Searching for **data** it will not present results for **data**base or C**DATA**.
      To be activated in Admin settings.

New feature:
      Index the meta contents in .xlsx and .docx files. If available in file, it will index:
            Title, Subject, Keywords, Creator, Description,
            Last Modified By, Revision, Date Created, and Date Modified.
      To be activated in Admin settings.

New method implemented, when manually aborting the index procedure.
      Now obligatory a repair, optimize and table flush will be performed.
      Worked out also for multi threaded indexing.

Thumbnails, created of the indexed URLs, will now be taken from top of the page; no longer centered.

UTF-8 charset ensured for data transfer from/to MySQL database.

Updated IDNA converter.

Updated English language file.

Several new session ids implemented to be stripped during index procedure.

Improved XML result output:
      Now caching the XML results and additionally offering
      - Date and time of query.
      - User IP.
      - Name of remote host.
      Also all XML output files are now stored for permanent usage in sub folder …/xml/stored/
      To be deleted in Admin backend in 'Clear' menu.

Improved support for multiple table sets in one database.

If the option 'Do not index the full text' is activated in Settings menu, the 'Required number of words in a page in order to get indexed' is obligatory set to zero

Bug fixed in option:
Show 'Description' Meta Tag (if exists) on results page

Bug fixed in PHP error messages. Strict warnings will no longer be presented.

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

.../addurl.php
.../admin/admin.php
.../admin/admin_header.php
.../admin/admin_search.php
.../admin/auto_index.php
.../admin/db_activate.php
.../admin/db_common.php
.../admin/ad_copy.php
.../admin/configset.php
.../admin/db_main.php
.../admin/install_tables.php
.../admin/messages.php
.../admin/real_get.php
.../admin/real_log.php
.../admin/settings_backup.php
.../admin/spider.php
.../admin/spiderfuncs.php
.../include/click_counter.php
.../include/commonfuncs.php
.../include/idna_converter.php
.../include/media_counter.php
.../include/search_10.php
.../include/search_links.php
.../include/search_media.php
.../include/searchfuncs.php
.../include/show_id3.php
.../include/suggest.php
.../include/xml.php
.../languages/en-language.php

Version 3.2014b

Version: 3.2014b
Release date: 2014.06.03
Sphider-plus vs. original Sphider: 323 items worked out

New option:
        Obey the list of words not to be suggested.
        To be activated in Admin settings.
        For details please notice the chapter Suggest framework

New option:
        Suppress IDS warning messages in Admin backend.
        If activated in Admin settings, all warning messages delivered by the 'Intrusion Detection System'
        are blocked in dialogue of Admin backend.

Improved search algorithm for queries containing numbers.
        - If the search term **12.34** does not deliver any result, **12,34** will be used alternately.
         The same vice versa.
        - If the search term **test123** does not deliver any result, **test 123** will be used alternately
        ..The same vice versa

Improved detection of title tag in HTML head. If the <title> tag is not available, now also parsing <meta name
'title' = 'this title' />

Warning message added, if the cURL library as part of the PHP environment is not installed.

Improved suggest framework. Now delivering suggestions also for integer numbers.

Improved image search. Now searching in title of image, as well as in file name. Also thumbnails are created
for .png images, which are marked to be .jpg

Improved word stemming implementation. Now allowing to search for original and stemmed words.

Improved .doc converter. Now supporting 32 bit and 64 bit Windows OS.

Improved .ppt converter. Now supporting 32 bit and 64 bit Windows OS.

Improved Admin interface. If a site is member of multiple categories, all of them are presented in 'Sites' view
for the according URL.

Length of 'Name of promoted domain' enlarged to 255 characters.

Length of 'Promoted catchword in text' enlarged to 255 characters.

Modified title extraction for PDF, DOC, RTF and XLS files. In result listing, no longer presenting the file suffix
as part of the title.

Bug fixed in category search for option 'More results from category xyz'.

Bug fixed in defining an external template folder, outside of the Sphider-plus installation folder..

Bug fixed in German word stemming for upper case vocals.

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

  …/admin/admin.php
  …/admin/auth.php
  …/admin/configset.php
  …/admin/index_media.php
  …/admin/messages.php
  …/admin/spiderfuncs.php
  …/converter/catdoc.exe
  …/converter/catppt.exe
  …/include/commonfuncs.php
  …/include/search_10.php
  …/include/search_40.php
  …/include/searchfuncs.php
  :::/include/search_media.php
  …/include/suggest.php
  …/include/stemming/de-stem.php
  …/templates/html/010_html_header.html

Version 3.2014c

Version: 3.2014c
Release date: 2014.09.28
Sphider-plus vs. original Sphider: 327 items worked out

New option:
    Index only media content and ignore any text.
    To be activated in Admin settings.

New option:
    Do not index content, which is placed as div class="s-hidden" like:
        <div id=" . . ." class="s-hidden"> **this content** </div>

New option:
    Treat localhost URLs RFC 1808 conform.
        If activated, http:/localhost/ will always be used as root directory.
        Otherwise URLs like http://localhost/public_html/my_site/
        will also be accepted as root directory.

New option:
    Admin and database access authorization are now available
    as part of the 'Settings' menu in Admin backend.
    'User name' and 'Password' for Admin access, and also for the database configuration,
    are no longer stored as readable values.
    Now they are stored hashed in a separate file.

Accelerated index procedure for media content like images.

New feature in index procedure:
    Indexation of media content enabled, even if no text content is available
    (page contains less than x words).

Improved index procedure:
        - Now extracting the 'option' values in < select > tags.
        - Now splitting words also at multiple special characters.
        - Remove all tag content from full text.
        - Improved charset detection.

Improved option
    'Convert all kind of accents like á, ç, ê, ì, ü, into their basic vowels'
        Now additionally all special HTML character codes for Czech, Romanian, Slovak and Slovenian
        languages will be treated as accents and translated into their basic letters.
        Examples: Ę => E
                        ů=> u
                        ť => t
                        ž => z
                        etc.

Improved intrusion protection:
 - Prevent 'session fixation attacks'
 - Improved protection against SQL injection, even without activated IDS

Updated link and charset detection for HTML5 coded URLs.

Updated Danish language file. Thanks to 'incognito'.

Bug fixed in result listing for title presentation, containing %20 blanks.

Some small bugs fixed.

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

 …/admin/auth.php
 …/admin/auth_db.php
 …/admin/configset.php
 …/admin/db_config.php
 …/admin/index_media.php
 …/admin/messages.php
 …/admin/new_authent.php
 …/admin/spiderfuncs.php
 …/admin/settings/authentication.php
 …/include/commonfuncs.php
 …/include/search_40.php
 …/include/search_media.php
 …/include/searchfuncs.php
 …/languages/dk-language.php

Version 3.2015a

Version: 3.2015a
Release date: 2015.01.06
Sphider-plus vs. original Sphider: 332 items worked out

New feature:
      Responsive design for search form, result listing and addurl form.
      Automatically adapting to display size of computer, tablet, smartphone, etc.

New feature:
      Use list of **ul** tag classes to ignore the corresponding ul content during index/re-index.
      A common list of ul class values is used to ignore parts of a page.
          Content between <ul class='this_value'> and will be ignored,
          however links in it are followed. Multiple and nested ul tags will be attended.
          Values in common list may end with a wildcard, so that 'menu*' will work for
             menu1, menu2, menu_left, etc.
          Usable also for external pages, if it is impossible to add the <!--sphider_noindex--> tags.
          Details in chapter: Ignoring parts of a page defined by <ul class='abc'> . . . .

New feature:
      Use list of **pre** tag classes to ignore the corresponding pre content during index/re-index.
      Details in chapter: Ignoring parts of a page defined by <pre class='abc'> . . . . </pre>

New feature:
      User of the search engine may select, whether the advanced part of the search form
      should be presented.

New feature in admin backend:
      Verify the time required to create a web shot of any desired URL.

Improved protection against intrusion attempts.

Accelerated search algorithm, especially for presentation of multiple results per page.

Common footer file for addurl.php script (templates/html/092_footer.html).

Updated scripts for ID3 extraction. Now parsing ID3v1(v1.0 & v1.1) as well as ID3v2(v2.2 & v2.3 & v2.4).

Updated IDS scripts.

Updated bot and harvester list (black_ips.txt) to prevent unwanted queries.

Enhanced .htaccess file. For details see item 11 in …/.htaccess

Bug fixed in option 'Use list of div ids and classes to ignore the div content during index/re-index'.

Some small bugs fixed.

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

.../.htaccess
.../addurl.php
.../search.php
.../search_ini.php
.../admin/admin.php
.../admin/admin_header.php
.../admin/admin_search.php
.../admin/auth_db.php
.../admin/auto_index.php
.../admin/configset.php
.../admin/db_activate.php
.../admin/db_config.php
.../admin/db_copy.php
.../admin/db_main.php
.../admin/settings_backup.php
.../admin/spider.php
.../admin/spiderfuncs.php
.../admin/url_backup.php
.../admin/ID3/all_files
.../include/commonfuncs.php
.../include/commons.php
.../include/click_counter.php
.../include/ids_handler.php
.../include/search_10.php
.../include/search_40.php
.../include/search_media.php
.../include/show_id3.php
.../include/common/black_ips.txt
.../include/IDS/all scripts
.../languages/all scripts
.../templates/html/015_headline.html
.../templates/html/020_search-form.html
.../templates/html/025_search-form.html
.../templates/html/030_category-selection.html
.../templates/html/040_category-tree.html
.../templates/html/050_result-header.html
.../templates/html/060_text-results.html
.../templates/html/080_most_pop.html
.../templates/html/092_footer.html
.../templates/html/120_media-only results.html
.../templates/Pure/hdline.jpg
.../templates/Pure/userstyle.css
.../templates/Slade/hdline.jpg
.../templates/Slade/userstyle.css
.../templates/Sphider-plus/hdline.jpg
.../templates/Sphider-plus/userstyle.css

Version 3.2015b

Version: 3.2015b
Release date: 2015.03.08
Sphider-plus vs. original Sphider: 336 items worked out

New feature for index procedure:
      - Instead of the HTML tags 'title' and 'description', use admin edited title and description.
      Will be overwritten, if one of the following new features is selected:
      - Use admin edited title, if count of words in HTML title tag is less than: xxx
      - Use admin edited description, if count of words in HTML description tag is less than: xxx

New option in admin backend:
      Delete all sitemaps created during index procedures, which were performed in debug mode.
      To be found in sub menu 'Clean'.

User and admin interface now are HTML 5 conform. With special thanks to Ulf Dunkel for his code input.

Updated scripts for ID3 extraction. Now parsing ID3v1(v1.0 & v1.1) as well as ID3v2(v2.2 & v2.3 & v2.4).

Bug fixed in 'Multithreaded indexing' in conjunction with sitemap.xml files.

Bug fixed in option
      Allow other hosts with same domain name for all links
      found during indexing. Also ignore TLD, SLD and www

Bug fixed in 'Ignoring parts of a page defined by <div id='abc'> or <div class='abc'>'
in conjunction with nested div tags.

Bug fixed in 'Activate/disable database' menu for multiple databases containing the same table prefix.

Bug fixed in 'Import / Export URL list' for multiple categories per site.

Some small bugs fixed.

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

 …/addurl.php
 …/root.php
 …/search.php
 …/search_ini.php
 …/admin/admin.php
 …/admin/configset.php
 …/admin/db_main.php
 …/admin/http.php
 …/admin/index_media.php
 …/admin/install_tables.php
 …/admin/messages.php
 …/admin/setting_backup.php
 …/admin/spider.php
 …/admin/spiderfuncs.php
 …/admin/url_backup.php
 …/admin/getid3/all scripts
 …/converter/docx/AutoLoader.inc.php
 …/converter/docx/CreateDocx.inc.php
 …/converter/docx/Docx2Text.inc.php
 …/include/categoryfuncs.php
 …/include/commonfuncs.php
 …/include/search_10.php
 …/include/search_media.php
 …/include/show_id3.php
 …/templates/html/all files

Version 3.2015c

Version: 3.2015c
Release date: 2015-xx-xx
Sphider-plus vs. original Sphider: 342 items worked out

New option to define the chronological order of text result listing:
      Single result per page (ignoring arguments in URL)
      Will present only one result for URLs like:
      . . . /pizza-restaurants.html
      . . . /pizza-restaurants.html?page=1
      . . . /pizza-restaurants.html?page=2

New method of script addressing:
      No longer relative addressing, but the scripts now contain server based addressing.
      Consequently cron jobs may call the index procedure from anywhere,
      wherever the Sphider-plus scripts live on the server.

New feature:
      Admin backend protected against XSRF attacks (Cross-Site-Forgery-Request).
      Independent from IDS.

New feature:
      Admin backend protected against SFA attacks (Session-Fixation-Attacks).
      Independent from IDS.

New feature:
      Limit the duration of a session (time-out) in admin backend.
      To be defined in 'Settings' menu as seconds of inactivity.

New feature:
      Prevent search form from being flooded by too many queries per unit of time.
      To be activated in admin settings.

Improved suggest framework:
      Now starting the search function immediately after selecting any keyword.
      Without clicking the 'search' button.

Improved user statistics in admin backend:
      Now delivering all available Geo info, including Google map for IP location.

Improved detection of Operating System.

Updated data file for GeoIP.

Updated black list for Meta search engines.

Updated jQuery framework.

Bug fixed in function 'Index media'.

Bug fixed in 'Multithreaded indexing'.

Some more small bugs fixed.

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

Nearly all. Please use all scripts as per download for your installation.
Except all files in

…/templates/Pure/
…/templates/Slade/
…/templates/Sphider-plus/

These files remained unchanged since last version of Sphider-plus.

Version 3.2015d

Version: 3.2015d
Release date: 2015-07-05
Sphider-plus vs. original Sphider: 343 items worked out

New feature for command line operation:
       Enabled to index with respect to preference level.
       To be invoked by:
           -preferred <level>

Improved admin backend:
       Verification of PHP and MySQL server clocks to run synchronously and without offset.

Modified 'Pure' template.

Bug fixed in sub-menu of 'Server Info'.

Bug fixed in warning messages as part of index procedure.

Bug fixed in sub-folder creation for first log-in.

Some more small bugs fixed.

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

      …/addurl.php
      …/search.php
      …/admin/admin.php
      …/admin/admin_header.php
      …/admin/admin_search.php
      …/admin/auth.php
      …/admin/auth_db.php
      …/admin/auto_index.php
      …/admin/configset.php
      …/admin/db_copy.php
      …/admin/db_main.php
      …/admin/geo_show.php
      …/admin/install_tables.php
      …/admin/messages.php
      …/adminspider.php
      …/admin/spiderfuncs.php
      …/admin/url_backup.php
      …/templates/html/020_search-form.html
      …/templates/Pure/adminstyle.css
      …/templates/Pure/userstyle.css

Version 3.2015e

Version: 3.2015e
Release date: 2015-09-24
Sphider-plus vs. original Sphider: 349 items worked out

New feature:
        Block all queries for e-mail accounts like 'my-name@gmail.com'
        To be activated in admin backend.

New feature in admin backend:
        Create a default configuration file as backup of current settings and options.

New feature:
        Sphider-plus may now be installed on server using HTTP as well as HTTPS protocols.

Improved index procedure:
        Now automatically limiting the amount of data (full text) stored in database
        to the max. size of database table field.

Ready to run in PHP7 environment.
        Proven with PHP 7.0.0 RC3

Updated bad bot list. Now containing 1040 bot names.

Updated list of Meta search engines.

Admin backend updated for Opera and Chrome browser.

Bug fixed for 'Show EXIF info' in admin backend statistics 'Indexed Images'.

Bug fixed while simulating a web shot in admin backend.

Bug fixed for presenting multiple result pages in media search.

Some small bugs fixed.

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

.../admin/admin.php
.../admin/admin_header.php
.../admin/auth.php
.../admin/auto_index.php
.../admin/configset.php
.../admin/confirm.js
.../admin/geoip.php
.../admin/messages.php
.../admin/setting_backup.php
.../admin/spiderfuncs.php
.../include/commonfuncs.php
.../include/commons.php
.../include/search_10.php
.../include/common/black_ips.txt
.../include/commom/black_uas.txt
.../templates/Pure/adminstyle.css
.../templates/Slade/adminstyle.css
.../templates/Sphider-plus/adminstyle.css

Version 3.2016a

Version: 3.2016a
Release date: 2016-02-16
Sphider-plus vs. original Sphider: 360 items worked out

New feature:
     For promoted result presentation, now multiple keywords (catchwords)
     and also multiple domain names could be defined.

New feature:
     If available in result listing, show a complete phrases as text extract
     around the found keyword (search term).
     To be activated in admin backend.

New feature:
     Index only feeds and ignore all other page content like text and media.
     Never the less all links will be followed.
     To be activated in admin backend.

New feature:
     Index Youtube hosted videos.
     To be activated in admin backend.

New feature:
     If the option 'Search strictly for search results' is not activated in admin settings,
     domains could be searched with and also without the www prefix.

New feature:
     Ignore the content inside of meta tags like <meta THIS CONTENT />, which are placed
     in body part of the HTML. Never the less all links will be followed.
     To be activated in admin backend.

New feature:
     Ignore the content of noscript tags like <noscript> THIS CONTENT </noscript>,
     which might be placed in body part of the HTML.
     To be activated in admin backend.

New feature:
     Ignore the content of cookies, which might be added to the page content.
     To be activated in admin backend.

New feature:
     Do not store links and their attributes as keywords.
     To be activated in admin backend.

New feature:
     Database support for full UNICODE, including astral symbols.
     Requires MySQL server version 5.5.3+

New feature:
     Compressed transfer on the Internet enabled for page content and PHP scripts.
     Depending on server environment this feature may not work on all servers.

Improved MySQL database support:
      - Now creating tables in compressed format.
      - Protection to prevent error # 1071 (Specified key was too long, max key length is 767 bytes).

Improved index procedure:
      If 'Spider can leave domain during index procedure' is not activated in admin settings,
      the external links are not stored as keywords.

Improved UP and DOWN buttons in admin 'Settings' menu, and also in result listing.

Wrapper added to bypass the PHP bug (error known since PHP v.5.3)
gzopen() => gzopen64() and all other gz functions.

Bug fixed to store the admin and dispatcher e-mail account in admin backend.

Bug fixed in <!--sphider_noindex--> directive.

Bug fixed for search terms with a length < 3 characters.

Some more small bugs fixed.

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

…/.htaccess
…/addurl.php
…/php.ini
…/admin/admin.php
…/admin/admin_header.php
…/admin/admin_search.php
…/admin/configset.php
…/admin/db_activate.php
…/admin/db_common.php
…/admin/db_config.php
…/admin/ad_copy.php
…/admin/db_main.db
…/admin/index_media.php
…/admin/install_tables.php
…/admin/messages.php
…/admin/php.ini
…/admin/real_get.php
…/admin/spider.php
…/admin/spiderfuncs.php
…/admin/url_backup.php
…/include/categoryfuncs.php
…/include/click_counter.php
…/include/commonfuncs.php
…/include/media_counter.php
…/include/search_10.php
…/include/search_40.php
…/include/search_links.php
…/include/search_media.php
…/include/searchfuncs.php
…/include/show_id3.php
…/include/suggest.php
…/templates/html/010_html_header.html
…/templates/html/011_html_header.html
…/templates/html/060_text-results.html
…/templates/html/070_more-results.html
…/templates/html/120_media-only results.html
…/templates/html/160_stream-results header.html
…/templates/html/170_stream-results.html
…/templates/Pure/adminstyle.css
…/templates/Pure/userstyle.css
…/templates/Slade/adminstyle.css
…/templates/Slade/userstyle.css
…/templates/Sphider-plus/adminstyle.css
…/templates/Sphider-plus/userstyle.css

Version 3.2016b

Version: 3.2016b
Release date: 2016-03-22
Sphider-plus vs. original Sphider: 364 items worked out

New feature:
        Besides XML result output file, now also a JSON file is created.

New feature:
        Protect the .../admin/ folder against external access by foreign clients (Ips).
        To be activated in admin backend.

New feature:
        'Protect the .../admin/ folder by means of a .htaccess file' now is activated by default.

New feature:
        Suppress usage of user's right mouse click (context menu) in result listing.
        To be activated in admin backend.

Improved media search. Now obeying the minimum word length as defined in admin backend.

Improved suggest framework. Now offering independent suggestions for text and media queries.

Bug fixed, which prevented caching of queries that contain numbers.

Bug fixed for media search, if multiple databases are involved.

Bug fixed for media search, now offering media suggestions.

Some small bugs fixed.

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

…/.htaccess
…/search_ini.php

…/admin/admin.php
…/admin/admin_header.php
…/admin/configset.php
…/admin/messages.php
…/admin/settings_backup.php
…/admin/spider.php
…/admin/spiderfuncs.php

…/include/commonfuncs.php
…/include/search_10.php
…/include/search_40.php
…/include/search_links.php
…/include/search_media.php
…/include/searchfuncs.php
…/include/suggest.php
…/include/xml.php

…/settings/.htaccess

…/templates/html/010_html_header.html
…/templates/html/011_html_header.html
…/templates/html/020_search-form.html
…/templates/html/025_search-form.html
…/templates/html/070_more-results.html
…/templates/html/200_no media-found.html