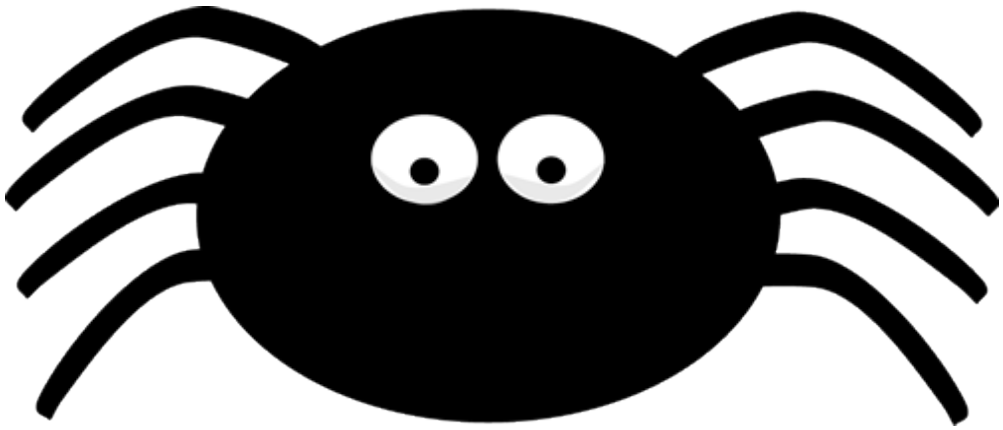
Sphider

User's Guide

Versions 5.5
and Lite 2.6



Contents

Introduction

Sphider is a lightweight web spider and search engine written in PHP, using MySQL as its back end database. It is a great tool for adding search functionality to your web site or building your custom search engine. Sphider is small, easy to set up and modify, and is used in thousands of websites across the world

.

Sphider not only supports all standard search options, but also includes a plethora of advanced features such as word auto-completion, spelling suggestions etc. The sophisticated administration interface makes administering the system easy. The full list of Sphider features can be seen on the About Sphider page.

The current official version is 1.3.6 and was released 6 April 2013, and it only was a security update to address a critical issue. The last release with any functional changes was 1.3.5, and that dates to 2009. Version 1.3.6 may be obtained from the Sphider PHP search engine site. The official version is a) no longer supported,[1] b) built upon earlier versions of PHP which contain much deprecated code, c) is highly vulnerable to SQL injection attacks as well as other forms of remote code execution, d) uses a suggest system which has grown increasingly unstable and unreliable as browsers change, and e) has several uncorrected bugs.

This version, 5.x.x or Lite 2.x.x, has been updated to use prepared statements and works with the latest current PHP (8.1 at this writing) and MySQL 5.6 or greater. MariaDB may be used in lieu of MySQL.

All queries, which in the official version use the now deprecated MySQL extension, have been updated since 1.5.1 to use the MySQLi /MySQLnd extension and prepared statements, virtually eliminating SQL injection attacks. The unstable and insecure SuggestFramework has been replaced by jQuery, making spelling suggestions dependable once again. All HTML is now HTML5 compliant. Configuration settings are now contained in the database, eliminating the horrendous danger presented when an entire page was completely rewritten using unfiltered $_GET data every time the configuration settings changed.

Windows operating systems, which was only partially supported in the official versions are now fully supported. All this represents only SOME of the improvements made in 1.5.1 and later.

1. The official Sphider site also has a forum, which supposedly provides support, although much of the "advice" is aimed at directing individuals to a paid Sphider-plus version , rather than giving genuine help and discussion for the free version.

About Sphider

Sphider is a popular open-source web spider and search engine. It includes an automated crawler, which can follow links found on a site, and an indexer which builds an index of all the search terms found in the pages. It also catalogs images occurring on each page (link) scanned, as well as the ability to store links found in a RSS feed. It is written in PHP and uses MySQL as its back end database (requires version 5.5 or above for both). For the standard 5.x.x and Lite 2.x.x versions, both MySQLi and MySQLnd are required.

Features

Spidering and indexing

- Performs full text indexing.
- Can index both static and dynamic pages.
- Finds links in *href*, *frame*, *area* and *meta* tags, and can also follow links given in javascript as strings via *window.location* and *window.open*.
- Respects robots.txt protocol, and nofollow and noindex tags.
- Follows server side redirections.
- Allows spidering to be limited by depth (ie maximum number of clicks from the starting page), by (sub)domain or by directory.
- Allows spidering only the urls matching (or not matching) certain keywords or regular expressions.
- Supports indexing of pdf, doc, xls and ppt files (using external binaries for file conversion).
- Allows resuming paused spidering.
- Possibility to exclude common words from being indexed.
- Indexes images occurring either directly or by reference to each link spidered.
- Indexes RSS feed links.

Searching

- **Default search**
  - Supports AND, OR and Phrase searches.
  - Supports excluding words (by putting a '-' in front of a word, any page including that word
  - will be omitted from the results).
  - Supports wildcard (*) searches.
  - Option to add and group sites into categories.
  - Possible to limit searches to a given category and its subcategories.
  - Possible to search all or a single specified domain.
  - "Did you mean" search suggestion on mistyped queries.
  - Context-sensitive auto-completion on search terms (a la Google Suggest).
  - Word stemming for English (searching for "run" finds "runnings", "runs", etc.).
  - Optional word stemming for eleven other languages, such as French, German, Italian, or Spanish.
- **RSS search**
  - Support AND and OR searches.
  - Supports wildcard (*) searches.
  - Can search all publication dates, a specific date, or a date range.
  - Can retrieve all feed items by leaving the query blank.
  - Possible to search all feed sources or a specific one.
- **Image search**
  - Can search by the occurrence of a word in the image name, in the image URL, or in the image 'alt' tag.
  - Can retrieve all images by leaving the query blank.
  - Supports wildcard (*) searches.
  - Possible to search all indexed sites or a specified site.

Administering

- Includes a sophisticated web based administration interface.
- Supports indexing via a web interface as well as from command line.
- Easy to set up cron jobs (or in Windows Task Manager).
- Comprehensive site and search statistics.
- Simple template system - easy to integrate into a site.

Installation

New installation

1. Unpack the files, and copy them to the server, for example to /home/youruser/public_html/sphider. This will be the '[path_of_sphider]'.

2. In the server, create a database in MySQL to hold Sphider data.

a) at command prompt type (to log into MySQL):
`mysql -u` <your username> `-p`
Enter your password when prompted.

b) in MySQL, type:

```
CREATE DATABASE `sphider_db` CHARACTER SET utf8mb4 COLLATE
  utf8mb4_0900_ai_ci;
```

Of course you can use some other name for database instead of sphider_db.

c) Use `exit` to exit MySQL.

At this point, it would be advisable to create a another user and password for use in the next step. For more information on how to create a database and give/get the necessary permissions, check MySQL.com

Note that creating the database can also be done using phpMyAdmin ,if available.

3. In **settings** directory, edit **database.php** file and change $database, $mysql_user, $mysql_password and $mysql_host to correct values. If you don't know what $mysql_host should be, it should probably stay as it is - 'localhost'. There is also $mysql_table_prefix, defaulted to a null value. If you desire to change this, the names of the soon to be created tables will all begin with the value of $mysql_table_prefix. For example, if you set $mysql_table_prefix = "sph_", the table "keywords" will be created as "sph_keywords". The prefix is optional.

4. in **settings** directory, edit **my.cnf** file with the appropriate host, user, and password.

5. Open **install.php** script (**admin** directory) in your browser, which will create the tables necessary for Sphider to operate.

Alternatively, the tables can be created by hand using **tables.sql** script provided in the **sql** directory of the Sphider distribution. At the prompt, type:
mysql -u USERNAME -p sphider_db < [path_of_sphider]**/sql/tables.sql**
You will be prompted for you password.
** Realize that creating the tables in this manner will NOT recognize any prefix designated by $mysql_table_prefix in the **database.php** file.

6. In **admin** directory, edit **auth.php** to change the administrator user name and password (default values are 'admin' and 'admin').

7. It is highly recommended that the **admin** and **settings** directories be password protected. If at all possible, the **admin** directory should also be set to only allow SSL access. When logging into the **admin** directory using standard http access, your directory user name and password are not encrypted. With https access, these items are encrypted and the risk of unauthorized access to the **admin** directory is greatly reduced. The **common_template** and **include** directories should also be protected. Do NOT restrict **js_suggest** or **templates**!

8. On Linux machines, you should check to bew sure your web server has read/write/delete permissions for the **admin/backup**, **admin/log**, **admin/reports**, **admin/sitemaps**, and **admin/tmp** directories. There is also a **tmp** directory in Sphider home that needs web server permissions. (Not all of these directories exist in SphiderLite).

9. Open **admin/admin.php** in a browser and start using Sphider.

10. The first step to take after getting the admin screen should be to click on the "Database" tab to ensure that all 29 tables (26 tables in SphiderLite) have been successfully created.

Upgrading an existing installation

1. If you already have an earlier installation of Sphider, you should first make a backup of your existing database and store it in a safe place.

2. In the server, alter your database in MySQL (or use phpMyAdmin) to current standards.

a) at command prompt type (to log into MySQL):
mysql -u <your username> -p
Enter your password when prompted.

b) in MySQL, type:
ALTER DATABASE `sphider_db` CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci;

Use your current database name in place of sphider_db.

c) Use exit to exit MySQL.

3. Delete these current directories and their contents:
admin
include
common_template
js_suggest
languages
settings
sql
templates
upgrade (if it exists)
Then delete the current files: changelog, install.txt, search.php, and SphiderUserGuid.pdf.
4. Unpack the new files to your existing sphider directory which you have just cleaned out.

5. In **settings** directory, edit **database.php** file and change $database, $mysql_user, $mysql_password and $mysql_host to correct values. If you don't know what $mysql_host should be, it should probably stay as it is - 'localhost'. There is also $mysql_table_prefix, defaulted to a null value. If you desire to change this, the names of the soon to be created tables will all begin with the value of $mysql_table_prefix. For example, if you set $mysql_table_prefix = "sph_", the table "keywords" will be created as "sph_keywords". The prefix is optional. Edit the **my.cnf** file to match values in **database.php**.

6. Open **version_update.php** script (**admin** directory) in your browser, which will update the tables necessary for Sphider to operate. Your existing data should be preserved.

7. In **admin** directory, edit **auth.php** to change the administrator user name and password (default values are 'admin' and 'admin').

8. It is highly recommended that the **admin** directory be password protected. If at all possible, the **admin** directory should also be set to only allow SSL access. When logging into the **admin** directory using standard http access, your directory user name and password are not encrypted. With https access, these items are encrypted and the risk of unauthorized access to the **admin** directory is greatly reduced.The **common_template** and **include** directories should also be protected. Do NOT restrict **js_suggest** or **templates**!

9. Open **admin/admin.php** in a browser and start using your updated Sphider.

NOTE ABOUT UPGRADING - The **changelog** lists which files have changed. It may be tempting to ONLY replace the changed files and be done with it. While this may be fine on a base level, if you do so, PLEASE DO RUN the **version_update.php**. It will make needed changes to your database.

FINAL NOTE ABOUT INSTALLATION - When you have completed installing or upgrading Sphider, the install.php and update_rollup.php scripts should be deleted. You won't be needing them and there is no sense leaving them around for someone else to misuse.

Using the Admin Panel

Settings Tab

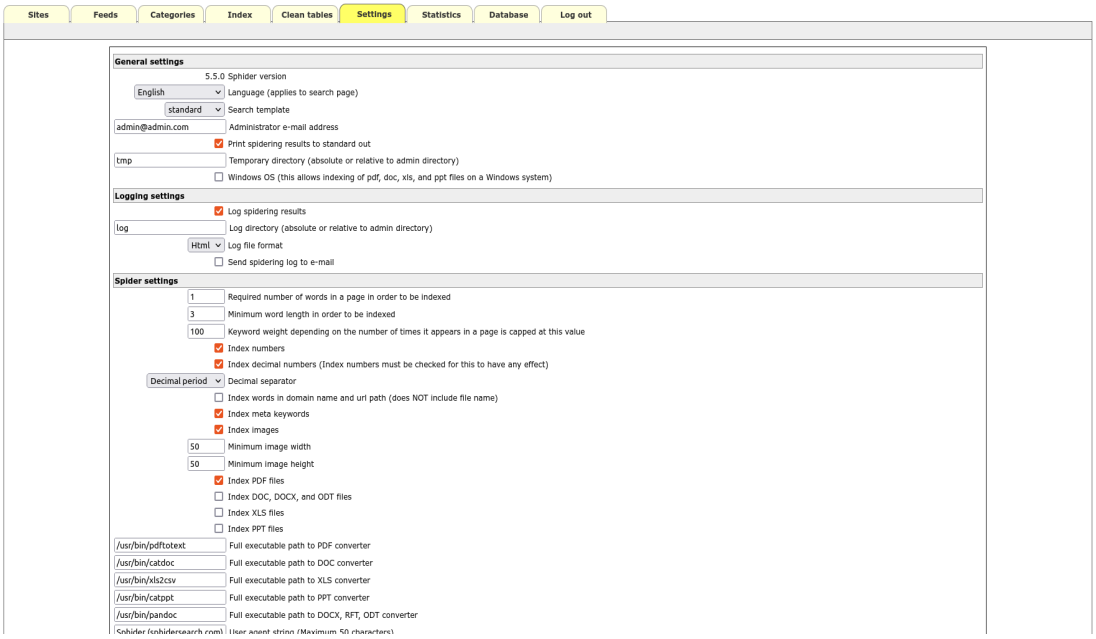


Figure 1: Settings tab

There are 68 user configurable settings (62 in SphiderLite) on this page.

GENERAL SETTINGS

• Language	A drop down list of available languages is provided. This is the language which will appear to the user on the search page.
• Search template	This drop down list shows available templates. Each template uses a CSS file to determine the look of the user search and search results pages.
• Administrator e-mail address	The e-mail address to which spidering log files may be sent.
• Print spidering logs to standard out	If this is checked, the spidering results will be displayed in the browser as spidering progresses.

• Temporary directory	This is the name and relative or absolute path to the temporary directory. This directory is used by Sphider during the parsing of url's during indexing. If a Windows path containing backslashes is used, the next setting, Windows OS, must be enabled. The path must exist.
• Windows OS	Check this box if Sphider is to be run in a Windows environment.

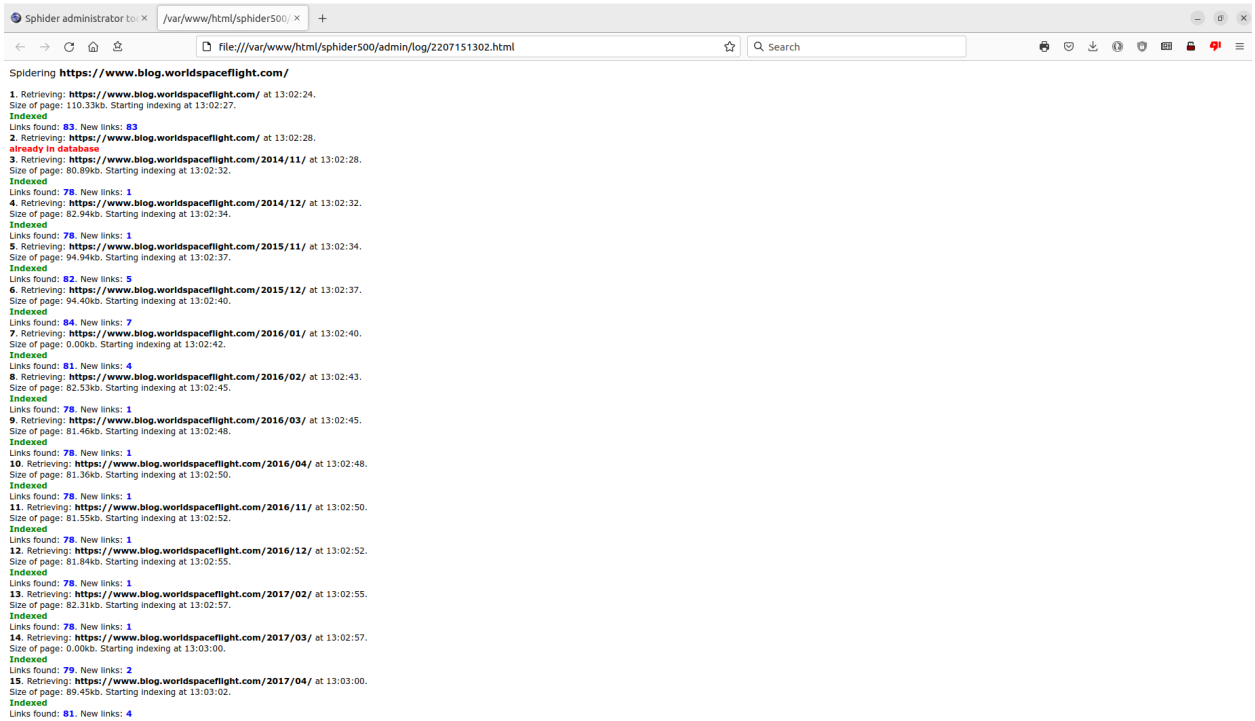


Figure 2: Example of a spidering log printed to standard output

LOGGING SETTINGS

• Log spidering results	If checked, a log file will be created for each occurrence of indexing or re-indexing.
• Log directory	This is the name and relative or absolute path to the log file directory. This directory is where spidering log files are stored. If a Windows path containing backslashes is used, the next setting, Windows OS must be enabled in General Settings.. The path must exist.
• Log file format	Log file may be in either HTML or text format.
• Send spidering log to e-mail	If checked, the spidering log will be e-mailed to the Administrator.

SPIDER SETTINGS

• Required number of words in a page to be indexed	This sets the minimum number of words which must appear on a page for it to be indexed.
• Minimum word length in order to be indexed	This sets the minimum length of a word before it can be indexed.
• Keyword weight depending on the number of times it appears in a page is capped at this value	A keywords weight is increased by the number of times it is used on a page. This caps the weight of a keyword.
• Index numbers	If checked, numbers will be indexed. (They are subject to minimum word length rules.)
• Index decimal numbers	If checked, decimal numbers will be indexed. (This setting will be ignored if the above 'Index numbers' is not also checked.)
• Decimal separator	Decimal period is default, but decimal comma may be chosen. Choice affects thousands separator.
• Index words in domain name and url path	If checked, words appearing in the domain name or path to a page will be indexed.
• Index meta keywords	If enabled, keywords appearing in meta tags are indexed.
• Index images	If checked, each page being indexed will be checked for images, and if found, the images will also be indexed. (Not in SphiderLite)
• Minimum image width	If Sphider can determine this size, this is the minimum width which will be accepted. (Not in SphiderLite)
• Minimum image height	If Sphider can determine this size, this is the minimum height which will be accepted. (Not in SphiderLite)

• Index PDF files	If checked, PDF files will be parsed and indexed.
• Index DOC files	If checked, DOC, DOCX, and ODT files will be parsed and indexed.
• Index XLS files	If checked, XLS files will be parsed and indexed.
• Index PPT files	If checked, PPT files will be parsed and indexed.
• Full executable path to PDF converter	This is the full path to the PDF converter. For a Windows OS, backslashes may be used. NOTE: The converter is not provided as a part of Sphider.
• Full executable path to catdoc converter	This is the full path to the catdoc converter. For a Windows OS, backslashes may be used. NOTE: The converter is not provided as a part of Sphider.
• Full executable path to XLS converter	This is the full path to the XLS converter. For a Windows OS, backslashes may be used. NOTE: The converter is not provided as a part of Sphider.
• Full executable path to PPT converter	This is the full path to the PPT converter. For a Windows OS, backslashes may be used. NOTE: The converter is not provided as a part of Sphider.
• Full executable path to Pandoc converter	This is the full path to the Pandoc converter. For a Windows OS, backslashes may be used. NOTE: The converter is not provided as a part of Sphider. Pandoc is needed to convert DOCX or ODT files.
• User agent string	This is the user agent string which will appear in the log files of the domain being spidered and indexed. It can be up to 50 characters in length.
• Minimal delay between page downloads	The minimum time, in seconds, between page downloads during spidering. Increasing this number will increase the amount of time required to spider a site, but may reduce the number of time-out errors.
• Pause	When checked, Sphider will pause for (1, 2, or 5) minutes after indexing (10, 20, 30, or 50) pages
• Use word stemming	If used, this should be enabled BEFORE indexing. It allows, for example, a search for the word "run" to also return "runs" or "running".
• Language to stem	Each language has its own algorithm
• Strip session ids	If enabled (recommended), session ids are removed from spidering results.

SEARCH SETTINGS

• Default results per page	This sets the number of results shown per page to 10, 20, or 50. (it can be overriddeen on the search screen.)
• Number of columns in category list	If categories are shown on the search page, this determines the number of columns to be used in their display.
• Bound number of search results	This limits the number of search results returned. When set to 0, the limit is removed.
• The length of the description string	This limits the length of the description string retrieved from the database. Visually, it will have no impact on the length of the description shown in search results unless the value is less than "Maximum length of page summary" (below). A 0 removes the limits.
• Number of links shown to "previous" and "next" pages	This limits the number of links shown for "Previous" and/or "Next" pages when the number of results returned exceeds the maximum number of results per page.
• Floor for query scores	Limits results to this minimum score. 0 means no limit.
• Show meta description in a results page	If enabled, the meta description will be used if available. If not available, the normal page extract will be shown in the result descriptions.
• Advanced search	Changes the default AND search to a AND/OR/Phrase search.
• Show result number	Toggles the result number on the results report
• Show index date	Displays the index date of the page reported
• Show url	Shows the url of the reported result
• Show query scores	This shows the query scores (chance of relevance) for each returned search result.
• Show stars	Shows scores using a 5 star system. Show query scores must also be enabled.
• Show categories	If enabled, categories will be displayed on the search form.
• Maximum length of page summary	This controls the length of the page summary for each search result.
• Enable spelling suggestions (Did you mean...)	If enabled, when a search returns empty but Sphider finds a similar word or phrase in the database, it will be suggested.
• Show the 2 most relevant links from each site	If enabled, only the 2 most relevant links in each domain are returned.

FORM SELECTION (Full version only)

• Display the classic search form	This will make the classic search form available. If no forms are selected, classic will auto-select.
• Display the RSS search form	This will make the RSS search form available.
• Display the image search form	This will make the image search form available.

SUGGEST

• Enable Sphider Suggest	This turns the suggestion feature on. If unchecked, none of the next five items are of any effect.
• Search for suggestions in query log	This enables suggestions from the query log. Only successful queries appear. Query log suggestions take priority over keyword or phrase suggestions.
• Search for suggestions in keywords	Enable suggestions from keywords. By default, suggestions are returned alphabetically.
• Use weighting when suggesting keywords	If suggestions for keywords is enabled, this alters their return from alphabetical to weighted. Keywords are weighted by frequency of occurrence.
• Search for suggestions in phrases	Enables suggestions from keyword phrases. This setting **overrides** any keyword settings, although phrase suggestions do not occur unless more than one word is entered.
• Limit number of suggestions	Controls the number of suggestions in the drop down from the query.

WEIGHTS

• Relative weight of a word in the title of a webpage	Assigns a relative weight to words appearing in a page title.
• Relative weight of a word in the domain name	Assigns a relative weight to words appearing in the domain name.
• Relative weight of a word in the path name	Assigns a relative weight to words appearing in a url path name.
• Relative weight of a word in meta_keywords	Assigns a relative weight to words appearing in meta tag keywords.

If any of the options in the Setting tab are altered, click the "Save Settings" button at the bottom of the page. The page will automatically refresh with the new settings.

Sites Tab

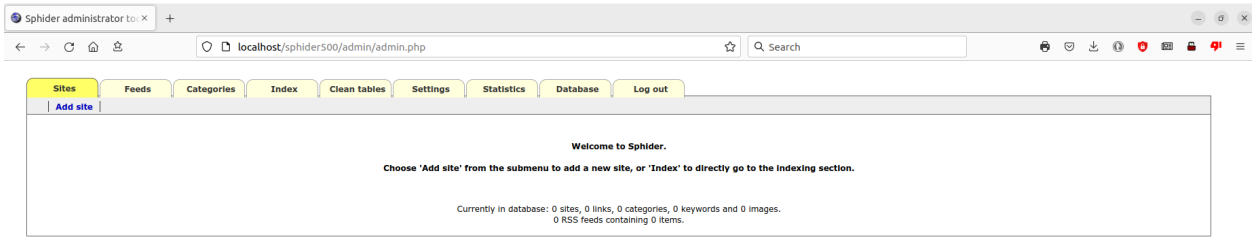


Figure 3: Initial appearance of the Sites screen

This tab shows information on all sites in the database. If this is a new installation, this tab appear as in Figure 3. When one or more sites have been added, you will see each site, one per line, showing Site name, URL, Indexing status, and a link to Options so you may edit the site. On the upper left of the Sites tab, you will initially have an additional link, Add site. Once one or more sites have been added to the database, a second link, Reindex all, will appear. See Figure 4.

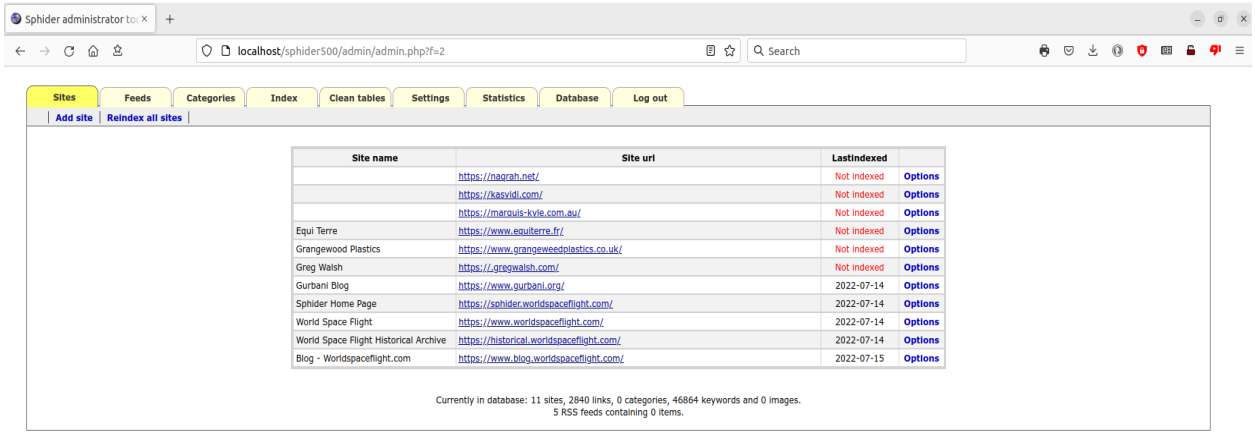


Figure 4: Sites tab after several sites have been added

Add site:

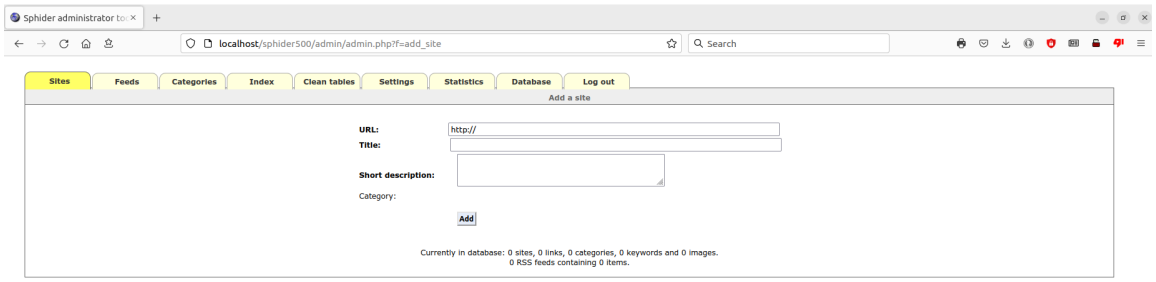


Figure 5: Add a site screen

From this screen, you can add sites to the database. For URL, enter the complete url of the site you want to add, for example, "http://www.bobbuilder.com/".

For the Title, enter the title of the site, for example, "Bob the Builder".

The Short description is a description of the site, for example, "Bob Smith, builder of fine custom homes in the Red River Valley".

If any categories exist, they will be displayed and you may choose which category or categories best fit this site.

Click **"Add"** to save the site. You will be taken to a new page showing the information you have entered about the site. Except for the "Site added" caption, this is the Options page accessed from the main Sites screen with each site listed. See Figure 6.

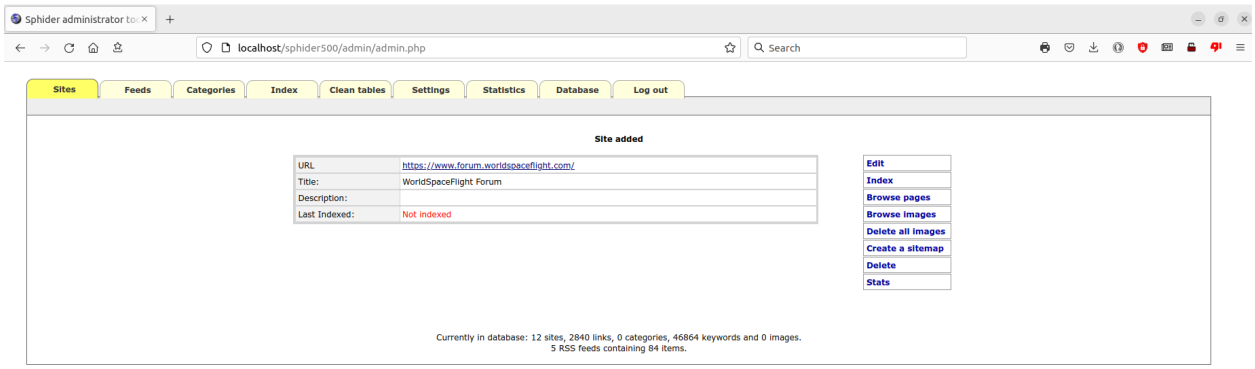


Figure 6: Site added screen showing options

On the right, there will be several options.

Edit takes you to the Edit site screen (Figure 7) which allows you to make changes to the site. You can change the title, description, or even change the selected categories. Most importantly, there are several other changes which may be made. Spidering options allows you to control how deep into a web site you wish to spider. The default, 2, means spider will search no more than two clicks away from the home page. Setting this option to Full removes any limitation.

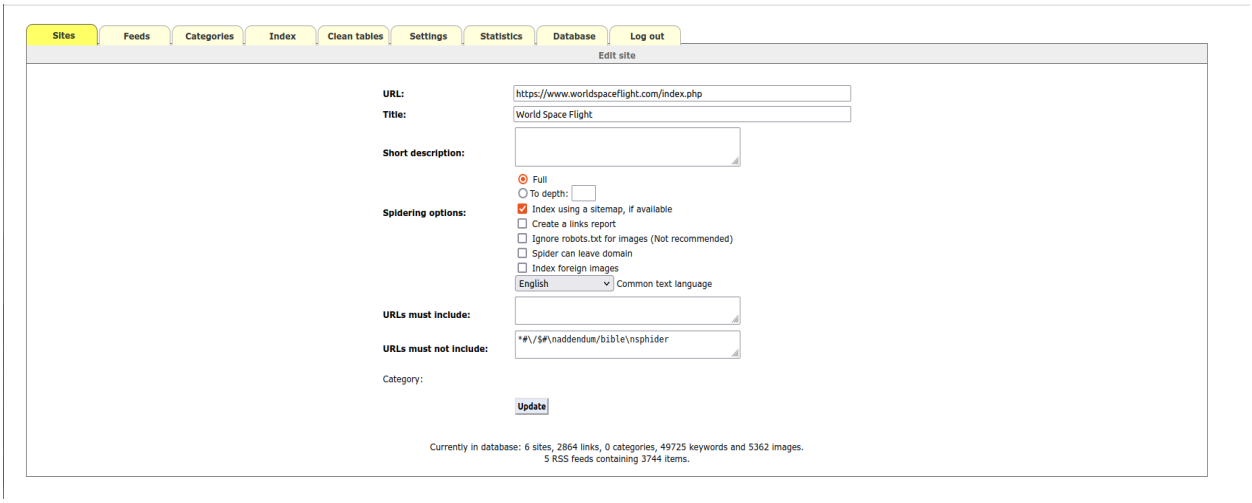


Figure 7: Edit site screen

Index using a sitemap, if available causes the site to be indexed by using the contents of the sites sitemap.xml (if it exists and is valid) instead of crawling and following links.

Ignore robots.txt for images causes to Spider to ignore the same rules as apply to indexing of links. Some sites may allow a page to be indexed, but ask that you keep hands off indexing images. This allows that to be overridden, but is not a recommended thing to do. Respect the site owners. (If you ARE the owner, then go ahead and index away!) (Not in SphiderLite.)

The Spider can leave domain means the search can include links to other sites.

Index foreign images allows you to index referenced images which are not native to the domain being indexed. (Not in SphiderLite.)

Common text language: This allows you to choose the common text language for the selected website. Common text words are excluded from indexing.

URL's must include is a list, one per line, of url's which must be included in the spidering. For example, you may want www.mysite.com/gotta-see-this to be indexed, so you would enter "/gotta-see-this" in the text box.

URL's must not include is a list, one per line, of url's which are not to be included in the spidering. If you have a set of pages in www.mysite.com/donot-search-here, you would enter "/donot-search-here" in the text box.

Both the must and must not lists may optionally use Perl style regular expressions in lieu of literal strings. Every string starting with a '*' in front is considered as a regular expression, so that '*/[a]+/' denotes a string with one or more a's in it. The delimiter used does not need to be a '/' (slash), but it is recommended that the character used not be one occurring in the regular expression.

When finished editing the site, be sure to click **"Update"** to save your changes. This will take you back to the main page on the Sites tab.

The Index (or Re-index) option takes you to a page where you may enter or change indexing options. This is initially a subset of the spider options given on the Edit page. Advanced options in the upper left will expand to show all indexing options. When you are ready, click **"Start indexing"**. Be patient. It may appear nothing is happening, but you may notice your browser indicating activity. If you enabled "Print spidering results to standard out" on the Settings tab, you will soon begin to see the spidering log appear. It will indicate when spidering is complete. If you did not enable "Print spidering results to standard out", just wait it out. Depending on the size of the site being crawled, it may be from a minute to an hour or more. When images are being indexed, this can add significantly to the time required.

Clear site allows all links and keywords associated with the site to be deleted. This essentially resets the site to a “Not indexed” status. (Images associated with the site are NOT deleted.) Clear site may be absent if the site hasn’t yet been indexed.

The Browse pages option lets you view a list of pages indexed on the site. If there is a long list, there is a filter which you can use to narrow the results. For example, putting "/contacts" in the filter and clicking the **"Filter"** button will restrict the pages listed to those containing "/contacts" in the url. You can change the number of urls listed per page. The default is 10. You also have the option to delete an indexed page from the database.

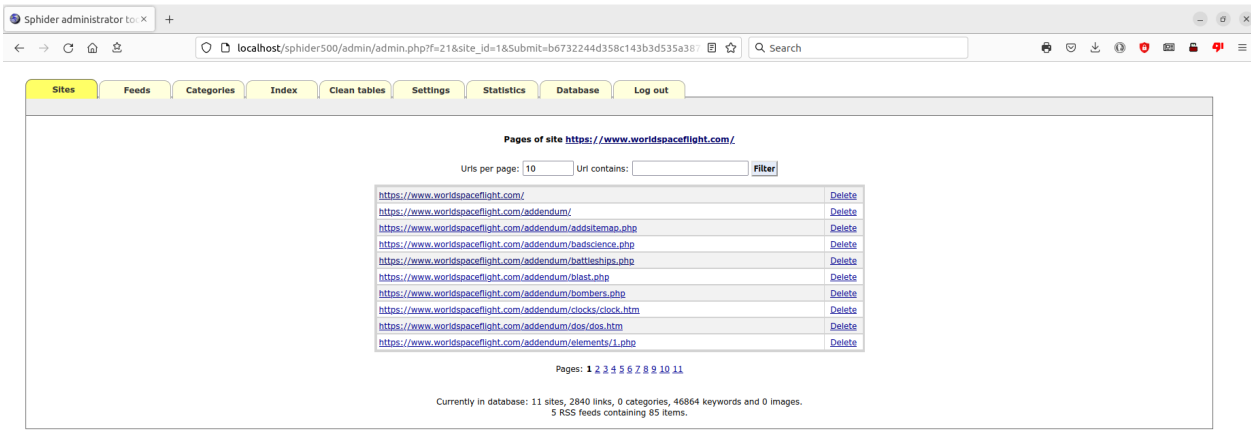


Figure 8: Browse pages

The Browse images option, like Browse pages, shows a list of the urls for images indexed for the site. The functionality is the same as Browse pages except it only applies to images. (Not in SphiderLite.)

Delete all images deletes all images associated with the site. When used with the Clear site option, ALL data associated with the site is deleted except the site settings. The site itself is not deleted. (Not in SphiderLite.)

The Delete option deletes the site and any indexed pages from the database.

The Stats option gives database information about the site indexing. It gives Last index date, number of Pages indexed, Total index size, Cached texts, Total number of keywords, and Site size.

Reindex all:

This link does exactly what it says. It re-indexes EVERY site in your database! In you have several sites in your database, this could take awhile! Don't click on Reindex all *just to see what happens!* You may be in for a rude awakening.

Feeds Tab

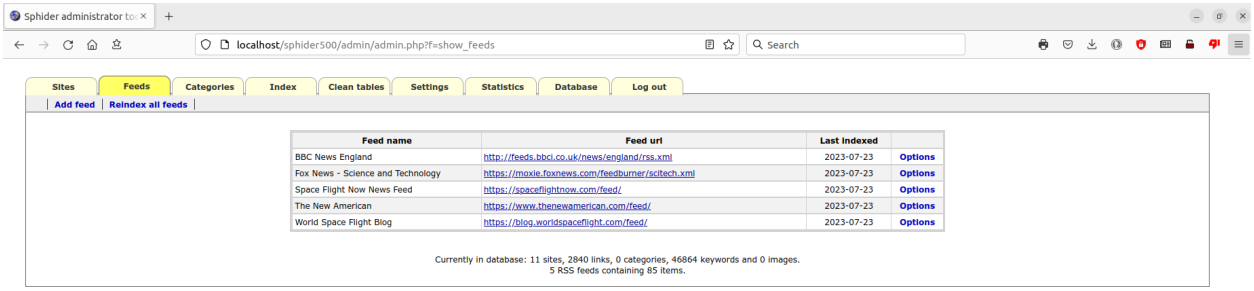


Figure 9: Show RSS Feeds

Just as with the main Sites tab, this page will initially show just a Welcome screen until you start adding RSS Feeds.

Feeds are added by clicking on the Add feed link in the upper left of the screen.

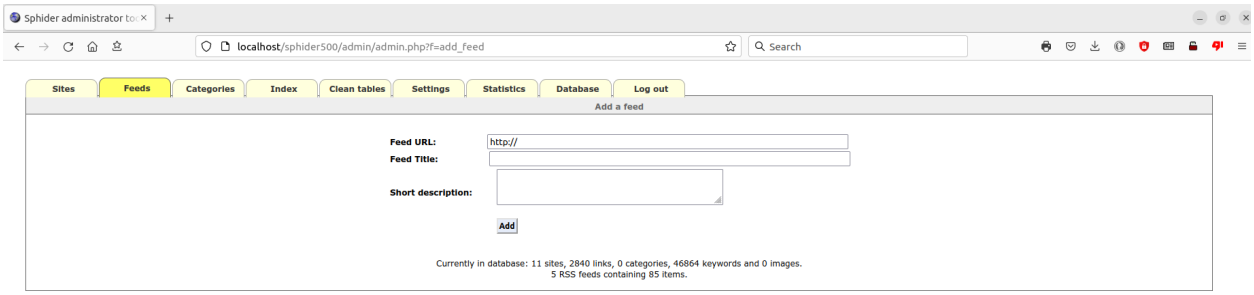


Figure 10: Add feed screen

Reindex all feeds is also an available option once there are feeds added. Unlike re-indexing all sites, re-indexing all feeds is not a time consuming task. Since feeds are volatile and change often, the individual items can change many times a day. It is recommended that all feeds be re-indexed regularly using a cron (or, in Windows, a scheduled task.) The Feeds tab does not occur in SphiderLite.

As an example for running a cron in a Linux environment which runs every 30 minutes, make a shell named “rssspider.sh” containing the following:

```
#!/bin/bash
cd varwww/html/sphider/admin
php rss_spider.php -all
```

Then create a cron job such as this:

```
MAILTO=""
*/30 * * * * homedan/Scripts/rssspider.sh
```

In Windows, Task Manager must be used. You can run a batch file on a daily basis, starting at 12:01 AM and repeating every 30 minutes. The batch file will look something like this named "rss_spider.bat":

```
cd "C:\Users\Dan\Documents\My Web Sites\sphider\admin"
php rss_spider.php -all
```

As an added tip, set this task to run as "SYSTEM" to prevent seeing a black command box flash open for a few seconds every half an hour!

Categories Tab

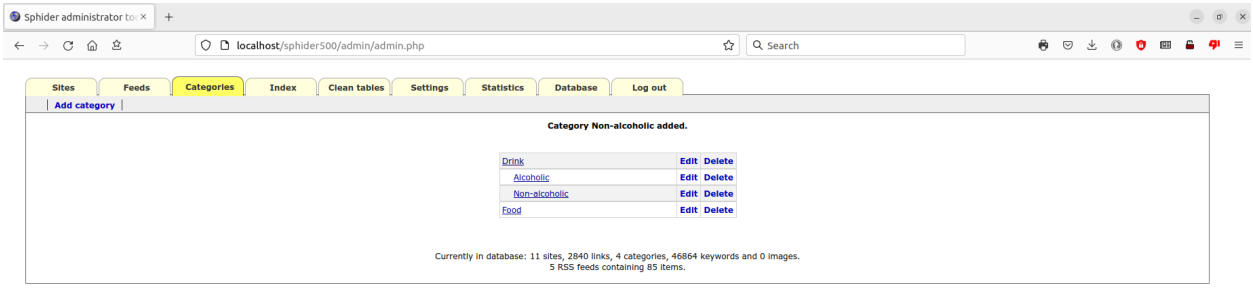


Figure 11: Categories tab

Categories provide a way of grouping web sites by category. Please do note, categories work at a site level, not a page level! You cannot assign some pages of a site to "Category One" and others to "Category Two".

This tab will initially be blank, except for the statistics at the bottom of the screen.

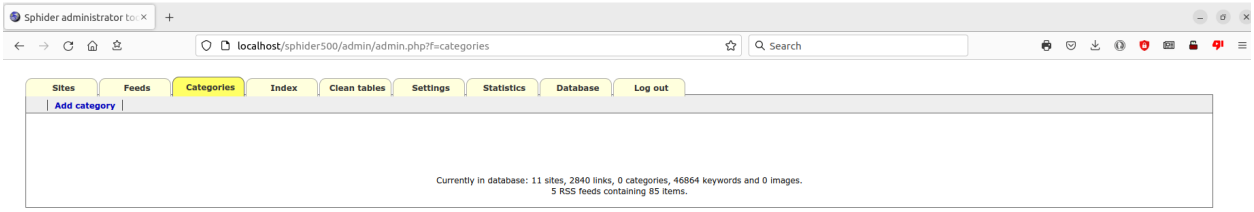


Figure 12: Initial blank category tab

Using the Add category link in the upper left corner of the page, enter the name of the category you wish to create, for example "Food". Click **"Add"**. The newly created category will appear. Repeat the process to add more categories. To add a sub-category, click the Add category link, then click on the category under which you wish to create the sub-category, then click **"Add"**.

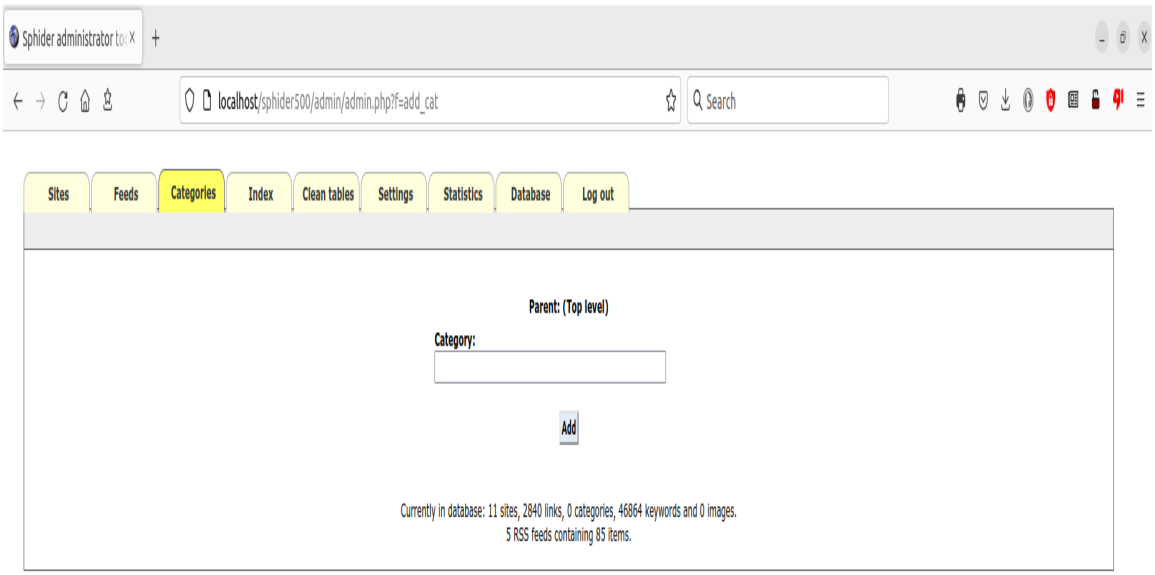


Figure 13: Add category screen

In the category list, Edit permits you to modify the category name. Delete removes the category from the list. Deleting a top level category automatically deletes all sub-categories under it.

Index Tab

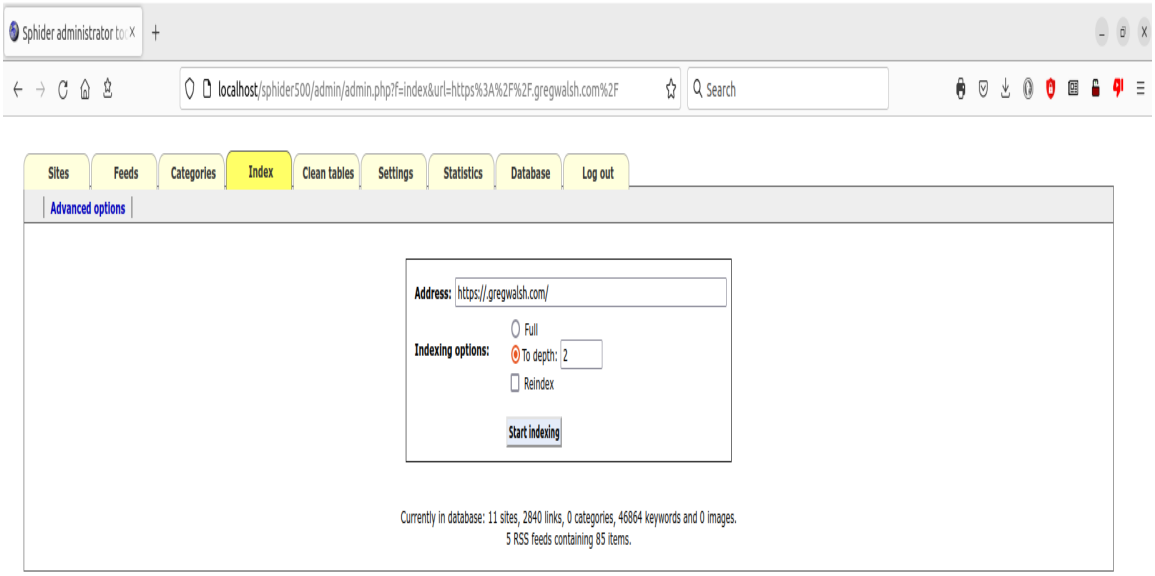


Figure 14: Advanced options hidden

On this tab, you may enter the url to any web site. Complete the indexing options as desired. Click **"Start indexing"** and the site will be indexed. If the site is not already in the database, it will be automatically added and will appear on the Sites Tab, although Site Name will be blank. Choosing Options to the right of the new site will allow you to change that.

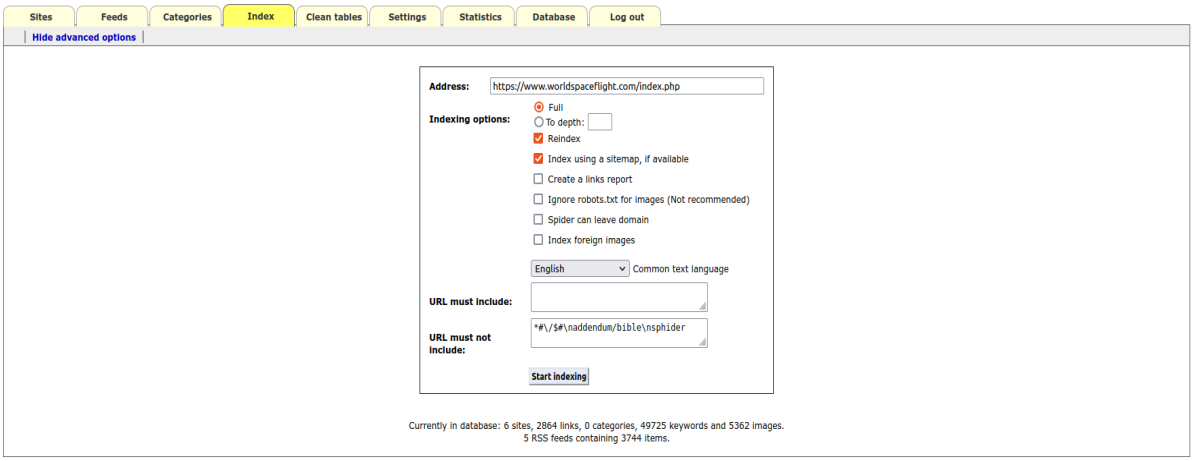


Figure 15: Showing advanced options

"Advanced Options" or **"Hide Advanced Options"** in the upper left toggles the screen between showing and hiding Index using a sitemap, Create a links report, Ignore robots.txt, Spider can leave domain, and Index foreign images, as well as the URL must include and URL

must not include boxes. Any url containing a string in the 'must not include' list is ignored. Any url that does not contain any string in the 'must include' list is likewise ignored.
‘

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

Clean Tables Tab

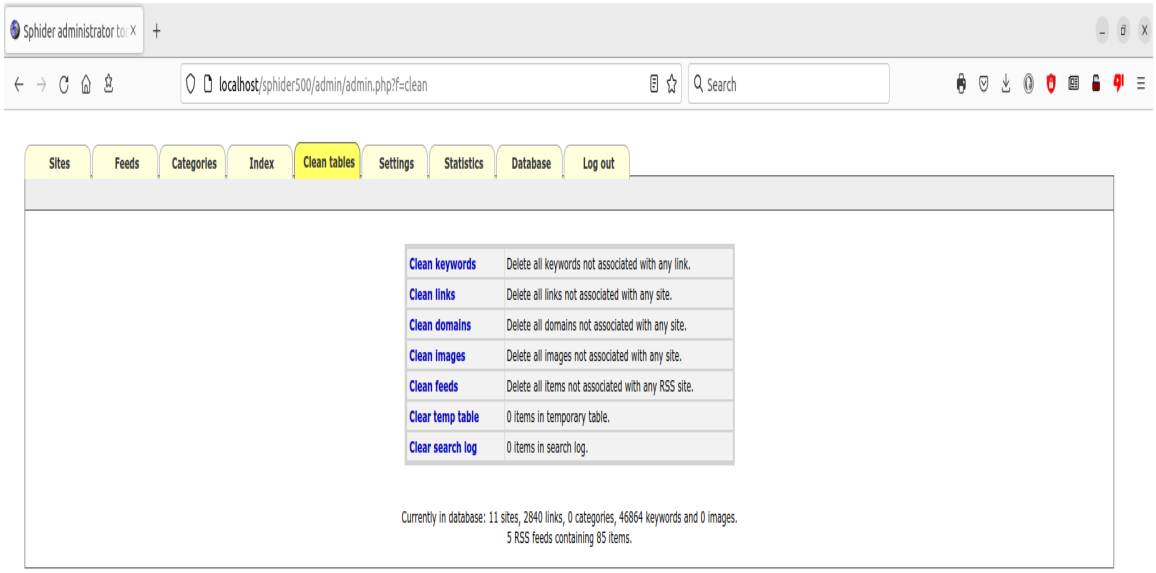


Figure 16: Clean tables tab

On this page, there are seven links (five in SphiderLite).
Clean keywords will remove any keywords not associated with any links in the database.
Clean links deletes any links not associated with any site in the database.
Clean domains deletes any domains not connected with any sites in the database.
Clean images deletes any images not associated with any sites in the database. (Mot in SpiderLite)
Clean feeds deletes any feed items not associated with any RSS feeds in the database. (Not in SphiderLite)
Clear temp tables cleans out the database temporary table, which is used by Sphider during indexing and re-indexing.
Clear search log deletes all entries in the search history.

Statistics Tab

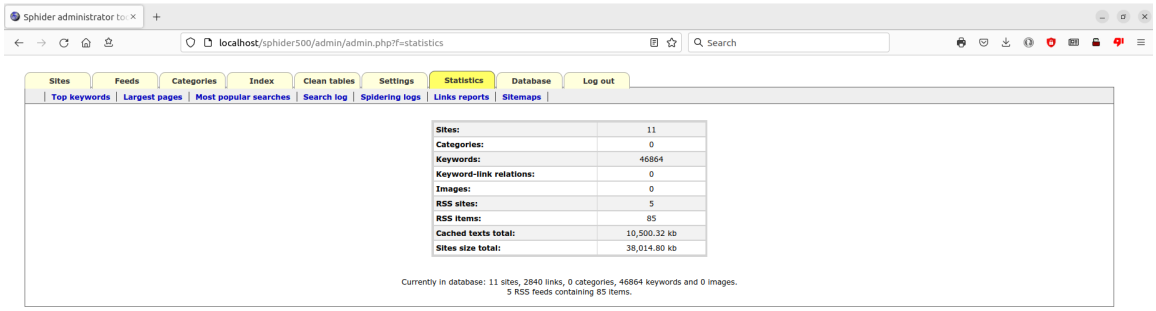


Figure 17: Statistics tab

The main screen on this tab provides overall data on the contents of the database.

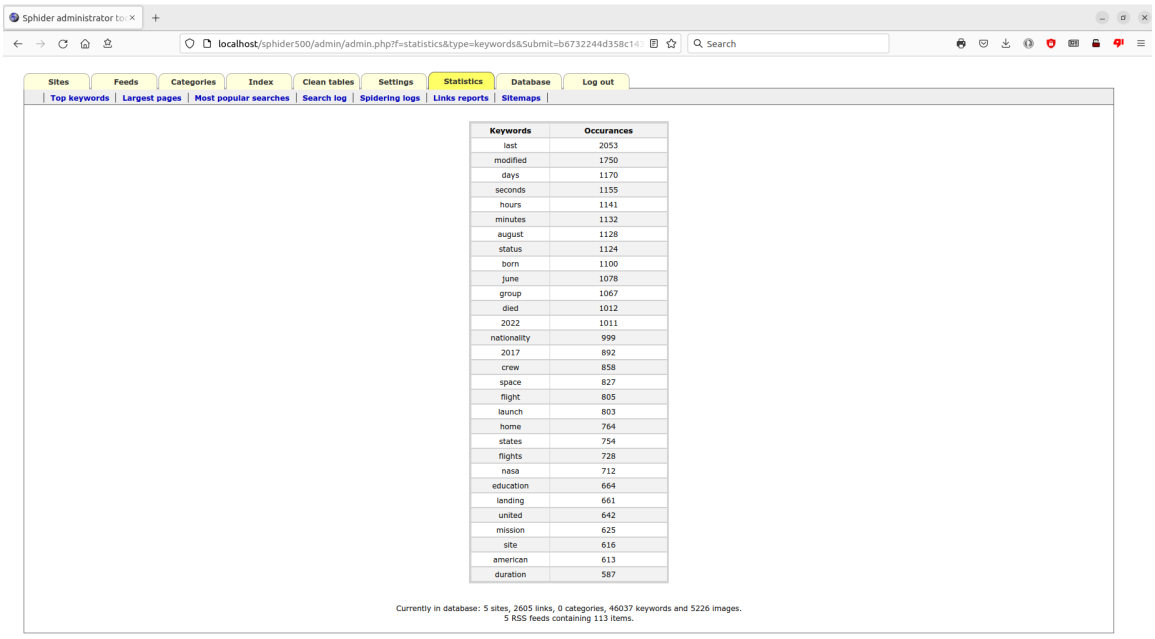


Figure 18: Top keywords

The Top keywords link lists the 30 most common keywords in the database and how many times each one occurs.

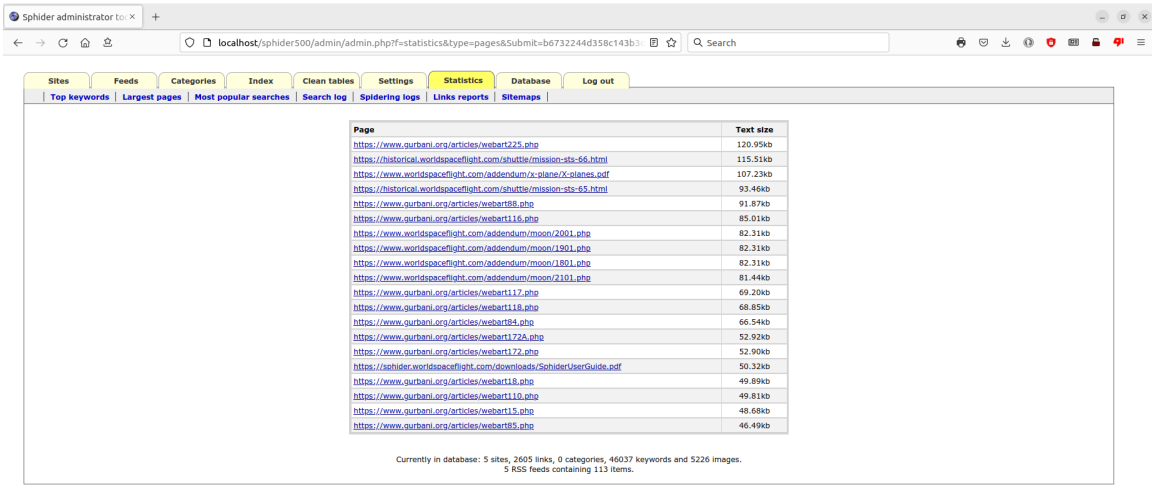


Figure 19: Largest pages

Largest pages lists the 20 largest pages in the database and their text size.

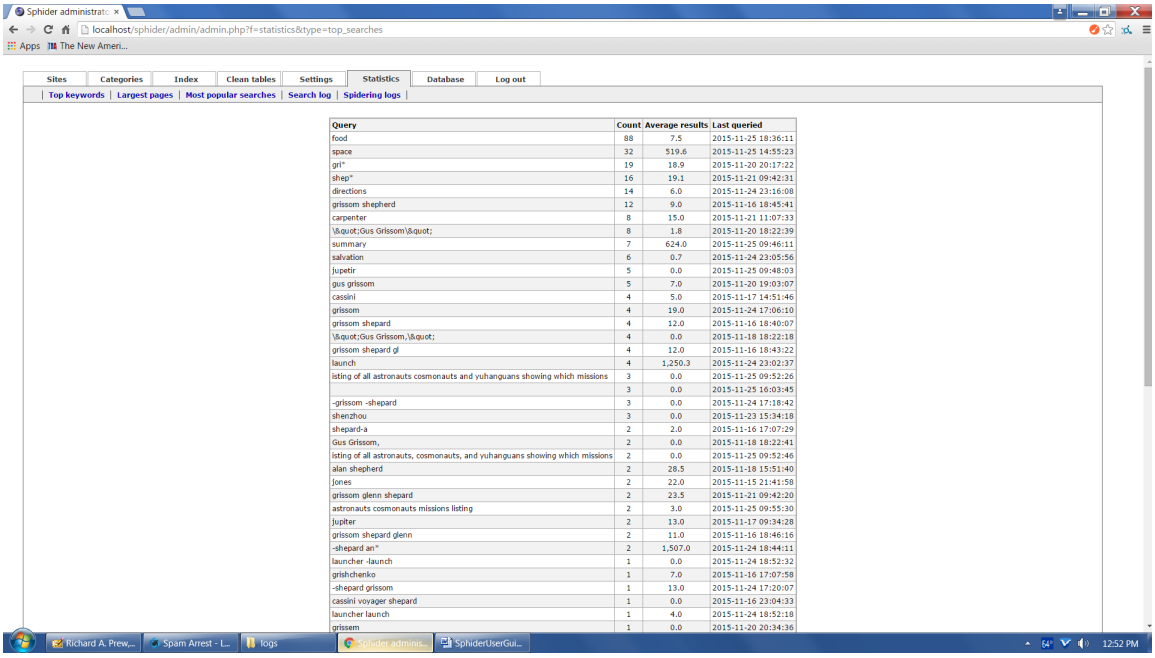


Figure 20: Most popular searches

The Most popular searches link lists the most popular queries, the number of times that query has been used, the average number of results returned, and date and time it was last used.

The Search log link is a dump of the database query_log table and contains the query, the number of results returned, the date and time the query took place, and how long the query took.

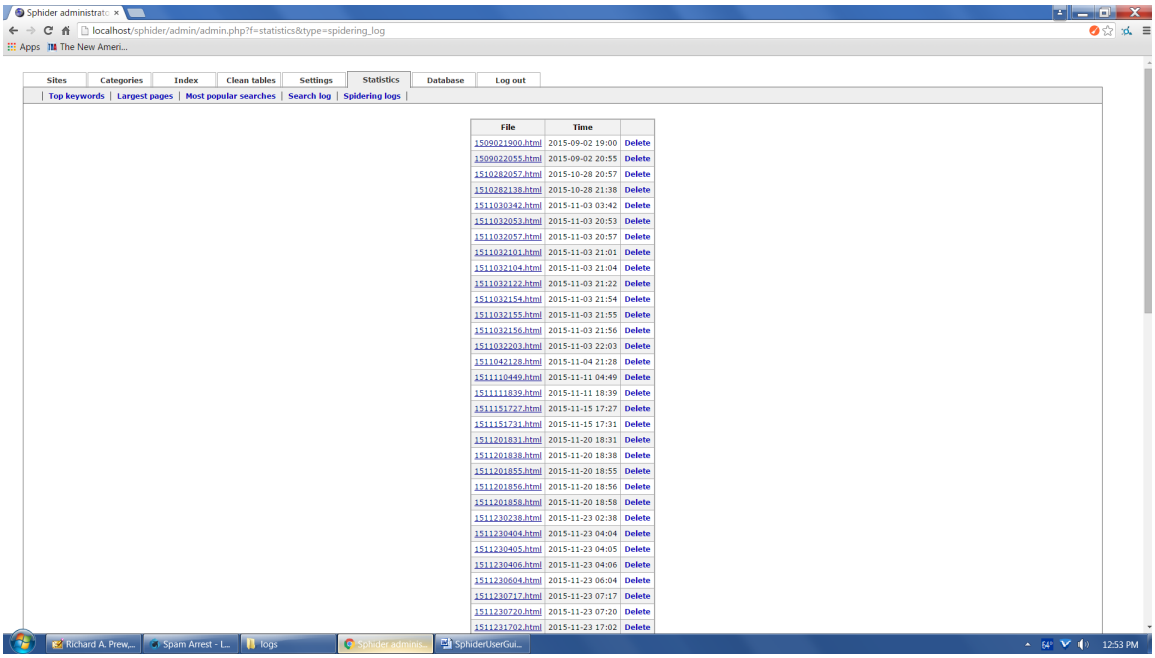


Figure 21: Search log

Spidering logs is a list, starting with the most recent, of all spidering log files in the log directory. It lists the file name and the date and time it was created. You can view a log by clicking on it. You may also **Delete** the file.

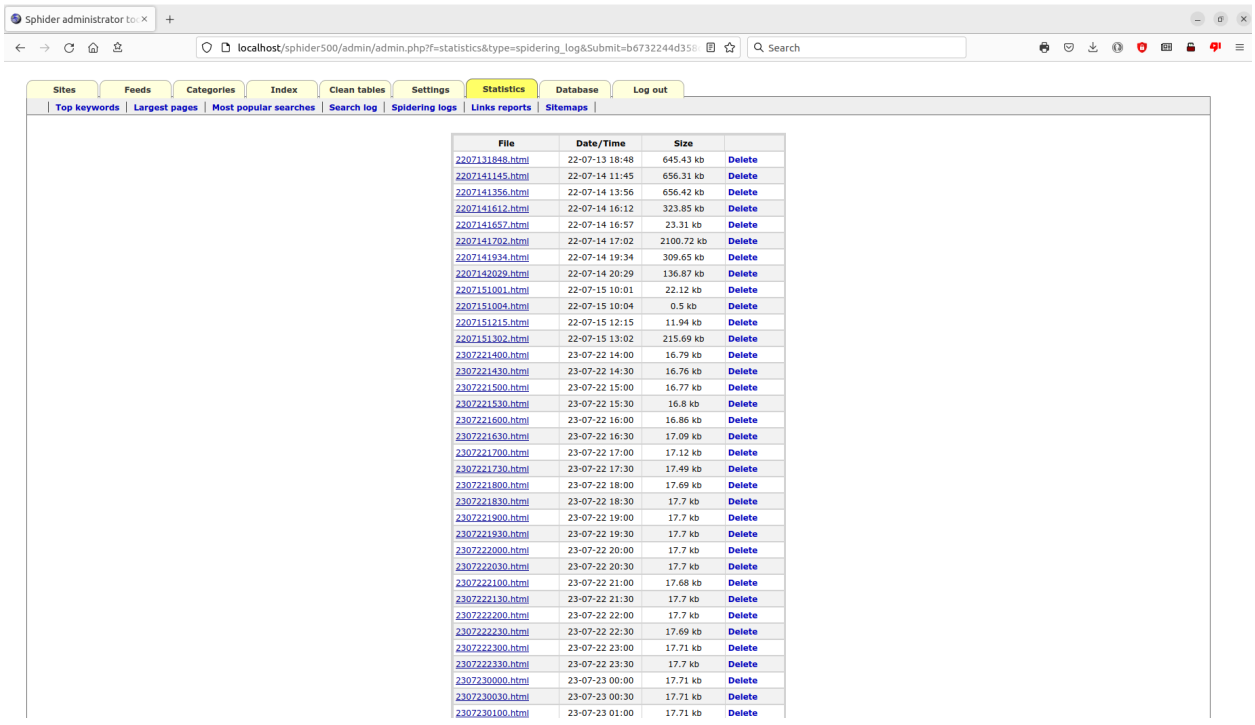


Figure 22: Spidering logs

Database Tab

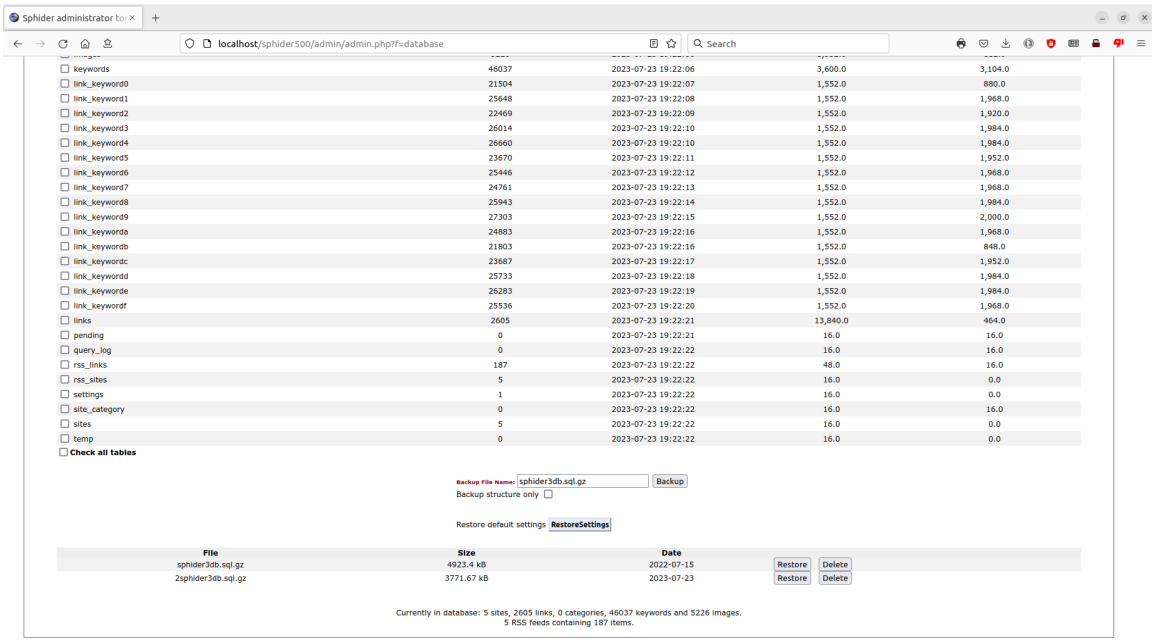


Figure 23: Bottom portion of the Database tab

This page lists all the tables in the database, the number of rows contained in each, date and time the table was created, the data size in Kb, and the index size in kB.

You may select tables individually, or click **Check all tables** to select all.

Selected tables may be **backed up**, or have only their **structure** backed up.

If you have done a structure-only restore, your setting table will be empty. Clicking the **Restore Settings** button will restore default configuration settings. You can also click **Restore Settings** if you simply want to go back to the default settings.

You may also change the default backup file name, although it is HIGHLY recommended you retain the .sql.gz at the end of the name. If a file with the same name already exists in the backup directory, it will be overwritten.

These backup files are stored in /admin/backup unless overridden on the **Settings** tab.

If there are existing backup files, they will be listed at the bottom of the page. You have the option to **Delete** or **Restore** any of these files. After any **Restore** has been run, you may need to refresh the page to see any changes.

Note that restoring a structure-only backup will delete ALL the data in the tables.

FINAL NOTE CONCERNING DATABASE BACKUP AND RESTORE: The backup and restore procedures have been completely rewritten in Sphider 4.0.0 (Lite 2.0.0) resulting in an

improvement in restore times. The original restore procedure restored the database a single row at a time. The new procedure uses mysqldump. What this means is that the number of individual sql statements to be processed is dramatically decreased, resulting in much faster times.

Our test database contains 15 sites, 10 categories, over 238,000 keywords, 57,000+ links (pages), almost 38,000kb of cached text, and has a cumulative size of over 247,000kb (gzip size ~18,000kb). This database was backed up in just under 20 seconds and was fully restored in approximately 30 seconds. This is down from over 7 hours restore time in the original version. The restore procedure was rewritten to accommodate the new rbackup method.

Log Out Tab

In case you haven't figured out what clicking on this tab might do, it logs you out of the Administration screen. This performs a secure log out and presents you with a generic page.

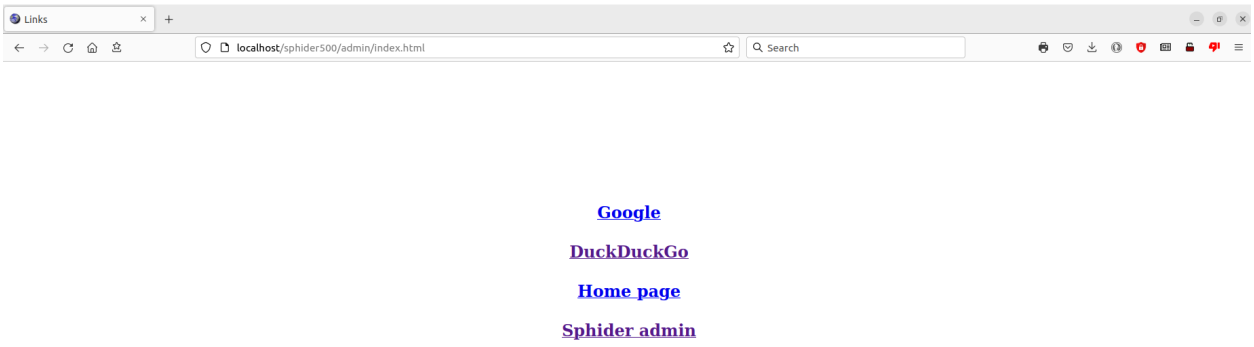


Figure 24: Generic page displayed after log out

Clicking on Sphider admin takes you back to the Sphider log in page. It ALSO starts a new session! While that isn’t enough to permit a malicious attack, it is a vulnerability that gives anyone with malicious intent a piece to the puzzle. This screen denies them that piece.

Using the Search Features

This is a screenshot of an example advanced search page. This example is a case with multiple domains and categories.

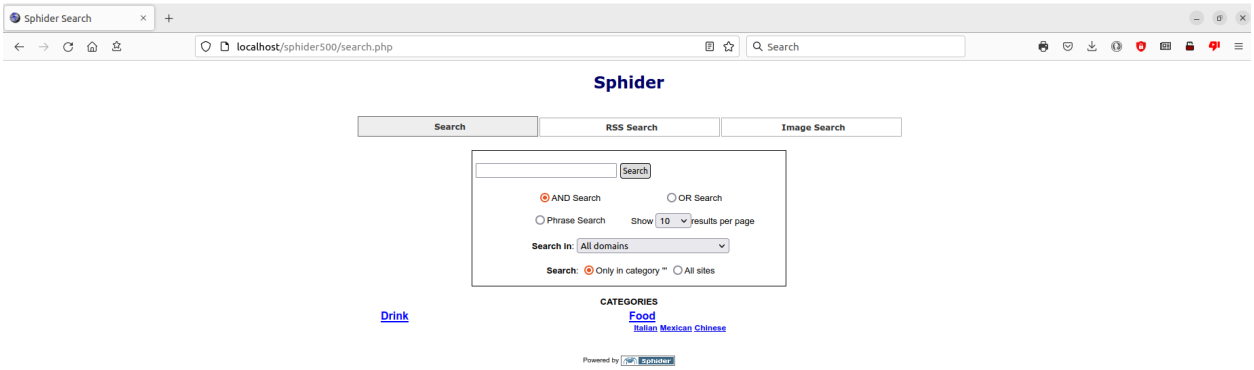


Figure 25: Default search screen with advanced options

It consists of a text box into which your query will be entered, options to choose the type of search to be performed (AND/OR/Phrase), and the option to search all sites (default) or to choose an individual site in which to search.

When search criteria are entered and set and the Search button clicked, one of several things may happen. If Spelling suggestions has been enabled in Setting and you fat fingered the search, for example you typed “spase”, no results will be returned but you will see the message “Did you mean: space”, at which point you can click on the suggest and redo the search with the other criteria remaining the same.

If nothing was found to match your search, you will see the message “No results found”. You can then click on the Reset for a new search button to try different criteria. Please remember, this is NOT Google! You are searching for specific words or phrases, and questions don’t work. For example,
searching with the phrase “What are the names of the seven dwarfs” as an AND search probably will get no results, and as an OR search will return every page in which ANY of the words appear!

The third scenario is that you get results.

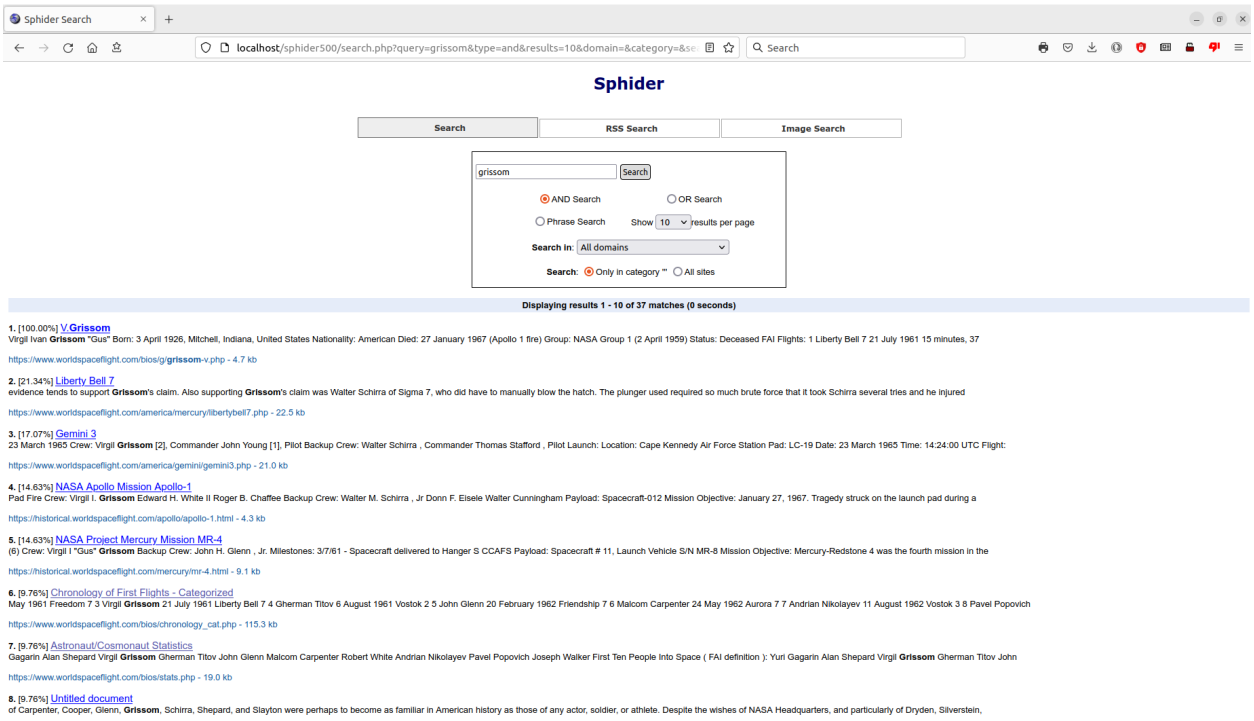


Figure 26: Results page

Alternatively, you may also click on a listed category. If you do so, you may then be present with the opportunity to choose a sub-category, it one exists.

Choosing a category search, your screen will look something Figure 27. Again, you will have the ability to select the type of search (AND/OR/Phrase) and whether to search only on the selected category, or to search all sites (default).

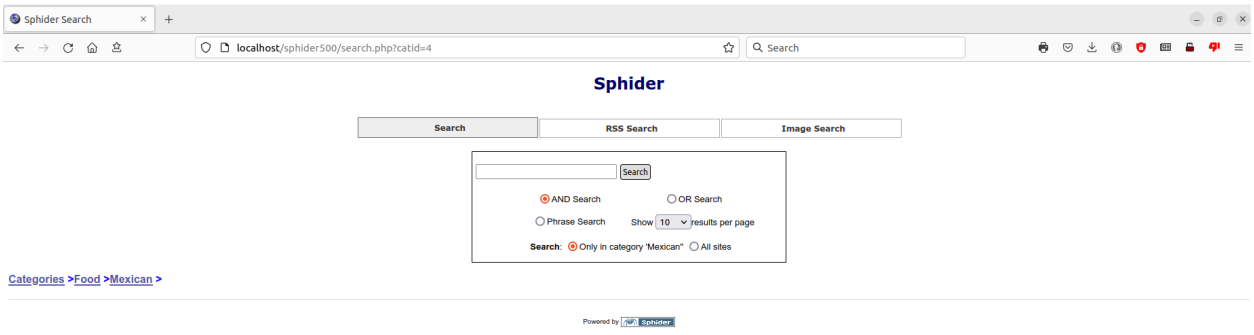


Figure 27: Search by category

If you do not have Advanced Search enabled in the configuration settings, the ability to choose the type of search will not be available and the search will default to type AND.

An AND search will require ALL words entered in to the query to appear in any results.

The OR query will return results for any page containing any of the search terms.

A Phrase search demands that not only all words must appear in the results, they must appear in the same order as in the query.

If Enable Sphider Suggest is enabled in the configuration settings, by the time you enter the third character into you search, you should see something like this:

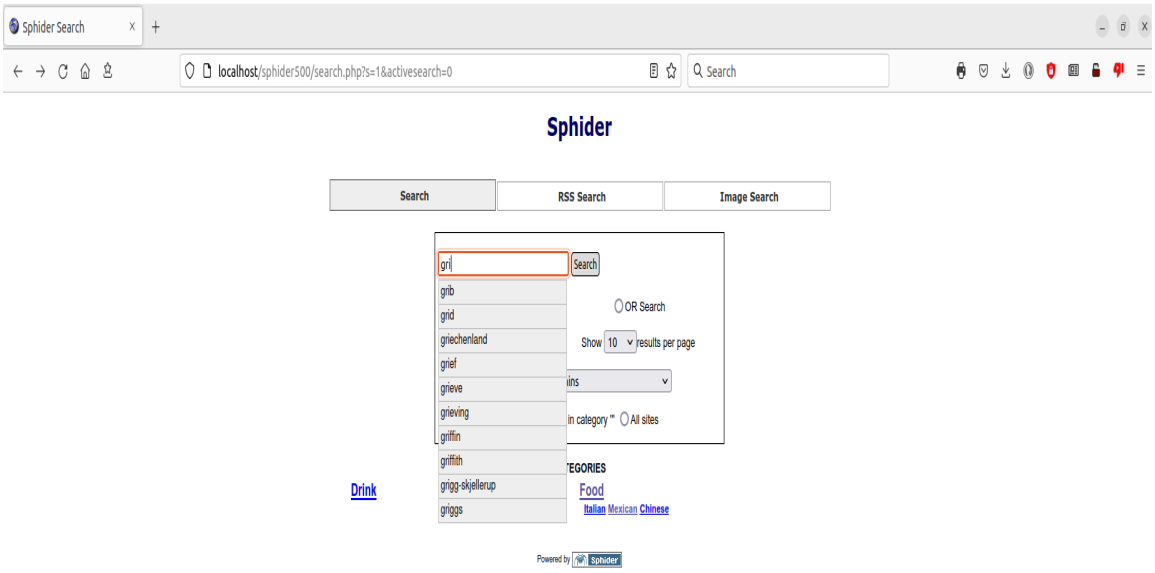


Figure 28: Sphider suggest enabled using keywords

What appears in the drop drown box below the query depends on your configuration settings. You can also set the maximum length of the list.

Queries may also contain a wildcard (*).

*ium will return words like medium, premium, and stadium (provided those words exist in your database.

Cho* will return the like of chop, choose, and chocolate.

St*p will return stop, step, or strip.

A "-" in front of a word will return pages which do NOT contain that word. The negate word cannot be used alone and must contain at least one other word you DO want to appear in the results. Example: "red -blue" will return results with pages which contain the word "red" but do NOT contain the word "blue". If the "-" is not preceded by white space, it will be part of the search term, such as in a hyphenated name or the word "x-ray".

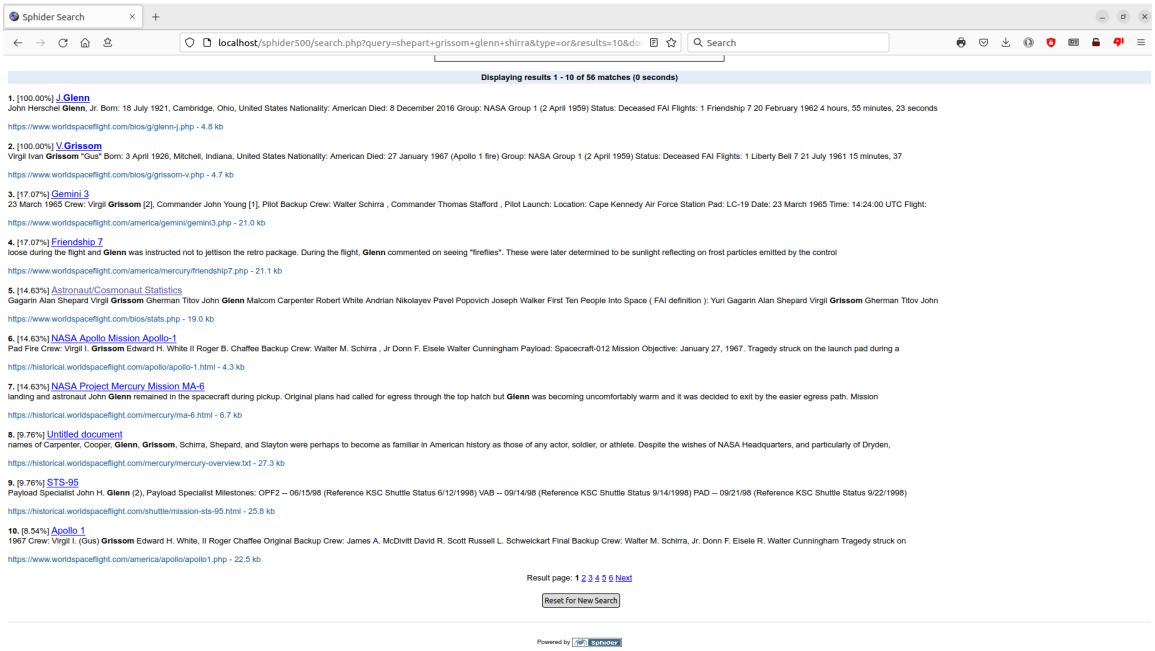


Figure 29: Results with multiple pages

When a search is successful, the results are displayed. You can control (from settings) whether to display 10, 20, or 50 results per page.

If more than results are returned than can be displayed on a single page, links to more pages will appear at the bottom in a Previous/Next format. From settings, you can control how many links can be provided.

If Advanced search is not enabled in settings, the search defaults to an AND type search.

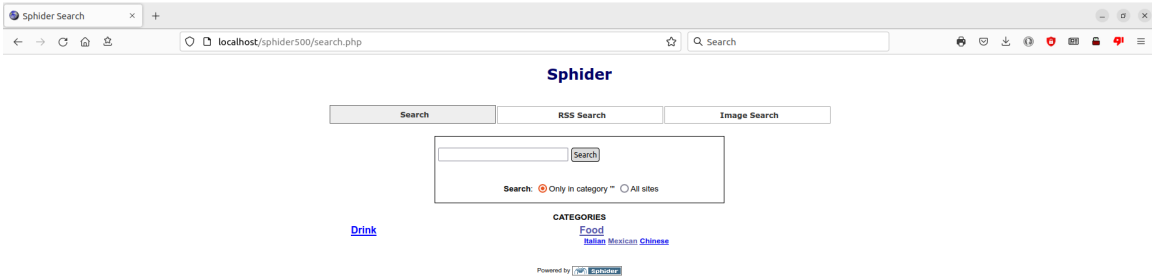


Figure 30: Default search with advanced search options turned off

When linking to your search page, even when Advanced search is not enabled, you may still display the advanced format by using "/search.php?adv=1" in your link.

The default search, with or without advanced search options, enable you to search the contents of pages of the sites you have actually spidered.

You may also do a search of all the RSS Feeds you indexed. (MB version only) An RSS search allows you to do either an AND or an OR search on feed titles. You can also enter '*' (wildcard) in the query box, in which case ALL items in the database are returned based upon other criteria you may have entered.

You may search All Dates, a specific date, or a date range. You may also specify to search All Feed sources, or a specific source.

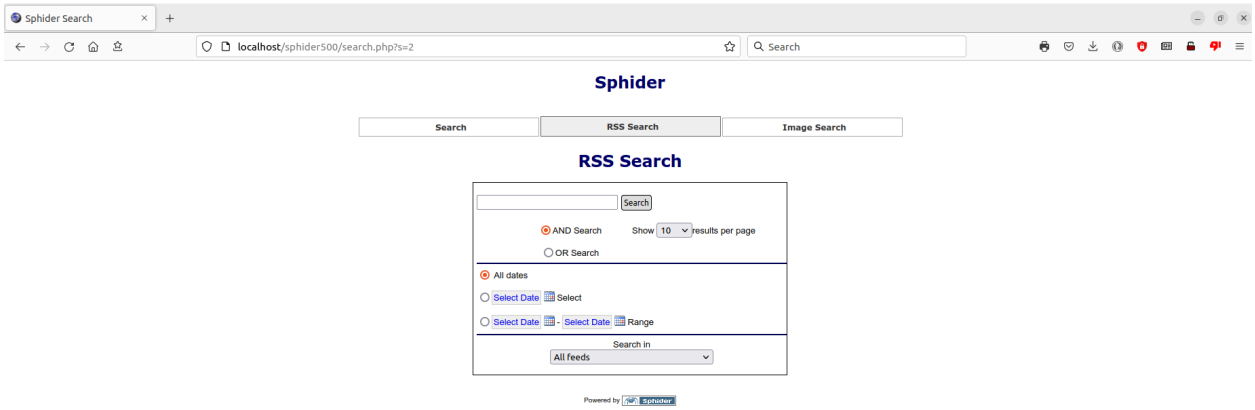


Figure 31: Initial RSS Search screen

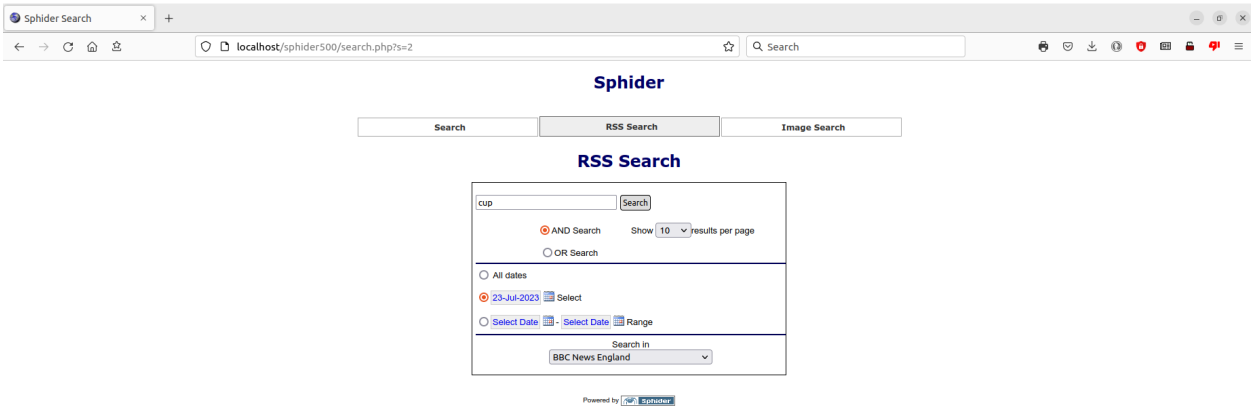


Figure 32: RSS Search screen with criteria entered

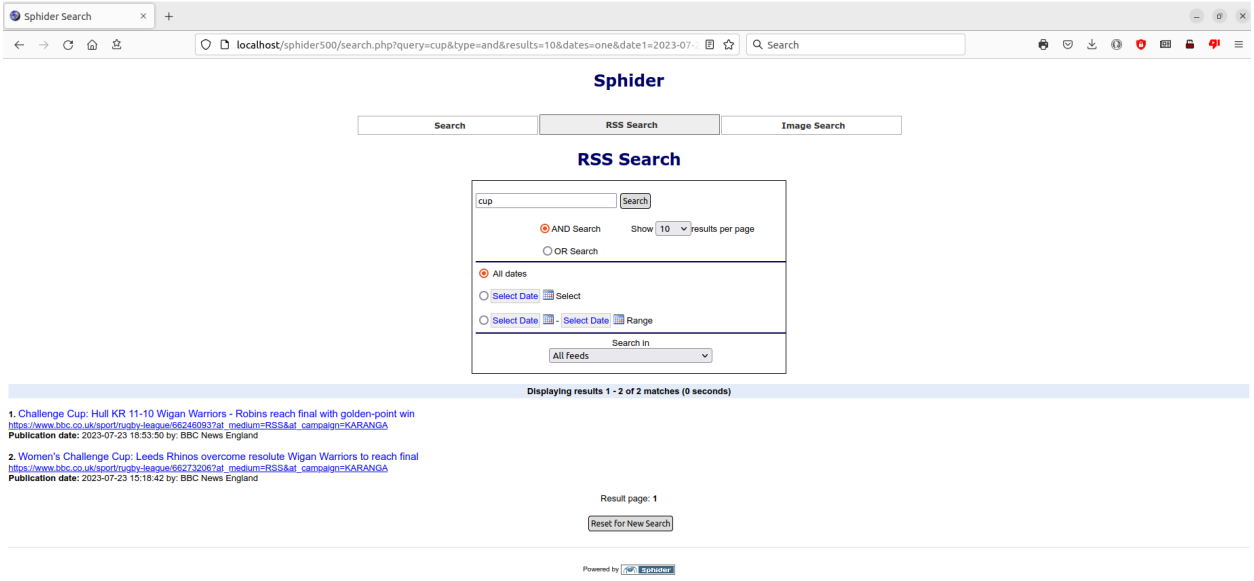


Figure 33: RSS Search results

41
This search returned only two items. As with the default search, results can run multiple pages.

The number of results per page may also be changed, either on the page or in Settings.

There is another type of search available, and that is the Image Search. (Not in SphiderLite)

Using the Image Search, you may use a single string of character to narrow the search and search in the image name, in the images' 'alt" tag, or in the images' URL. The search can also be narrowed by search a specific site, or search All Sites. The number of results per page may also be specified. As with a RSS Search, entering an '*' (wildcard) in the query box will return all images for the site chosen.

Illustration 31 shows The Image Search screen with results.

In the example displayed, the PHP installation includes the Imagick module. If Imagick is not available, the results will be the same, except the thumbnail preview on the left will be absent. The Search feature automatically will detect whether or not Imagick is installed and adjust the results accordingly. If you do not have direct control over PHP, ask your hosting company if Imagick might be installed. It is well worth it.

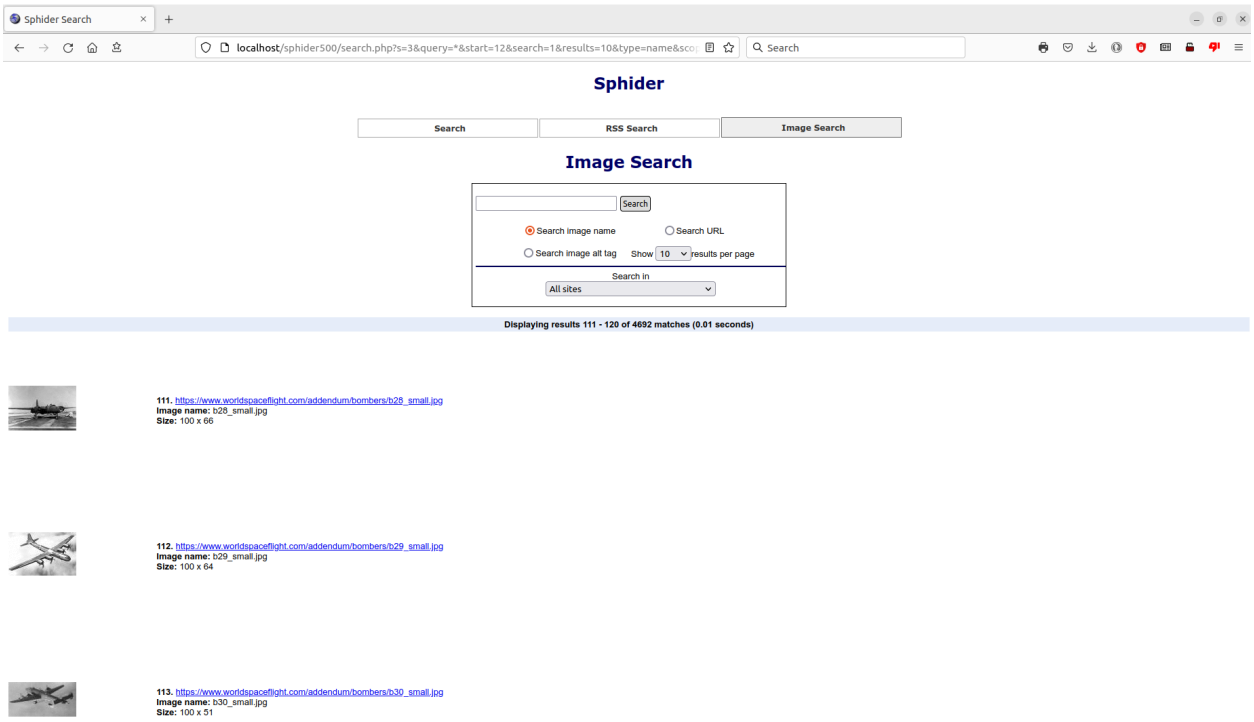


Figure 34: Image Search results

As with any of the search results (legacy, RSS, or Image), clicking on the underlined links will cause that link to open in a new tab.

An image preview will not be present when a mobile browser is used.

Spidering from the command prompt

In addition to indexing (or re-indexing) a web site from the Admin control panel, sites may also be spidered from the command prompt. To do so, first do a cd (change directory) to [path_to_sphider] /admin. The command prompt usage is as follows:

Usage: php spider.php <options>

Options:

-all	Re-index everything in the database
-u <url>	Set url to index
-f	Set indexing depth to full (unlimited depth)
-d <num>	Set indexing depth to <num>\n";
-s	Crawl using a sitemap, if available
-c	Create a links report
-i	Ignore robots.txt for indexing images ()Not in SphiderLite)
-l	Allow spider to leave the initial domain
-k	Allow Sphider to index referenced images not native to the domain (Not in SphiderLite)
-r	Set spider to re-index a site
-L <lang>	Specify the common text language
-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)

An example of how to use the command indexing is given:

php spider.php -u http://www.mysite.com -f -r -L es -n /mysearch\\n/docs

The first part, "php", allows you to execute php files.
"spider.php" is the spider function itself.
"-u http://www.mysite.com" tells spider to only index mysite.com.
The "-f" says to index to an unlimited depth.
The "-r" indicates that this is a re-index.
The “-L es” says that Spanish should be the common text language.
The "-n /mysearch\\n/docs" tells spider.php not to look in www.mysite.com/mysearch or in www.mysite.com/docs.

RSS Feeds may also be spidered from the command prompt in the MB version. This can be very useful when setting up cron jobs to keep rapidly changing feeds updated with the laster entries.

Usage: php rss_spider.php <options>

Options:

-all	Re-index everything in the database
-u	Set url to indexing
-r S	et spider to reindex a site

An example of how to use the command indexing is given:

```
php rss_spider.php -all
```

This will cause all RSS Feeds in your database to be rescanned and any new items indexed. This command may be run as a cron job or as a scheduled task in Windows. Pretty simple, eh?

Database.php

This file provides the connection to your database. It ships with default settings which must be changed before it can be used.

```
<?php
        $database="sphider";
        $mysql_user = "root";
        $mysql_password = "";
        $mysql_host = "localhost";
        $mysql_table_prefix = "";

        $db = new mysqli("p:".$mysql_host,$mysql_user,$mysql_password,$database);
      if ($db->connect_errno) {
        trigger_error("Database connection failed: ".htmlententies($db->connect_errno),
E_USER_ERROR);
      }

?>
```

$database="*sphider*"; Change *sphider* to the name of the database you have created and intend to use for your Sphider tables.

$mysql_user = "*root*"; Change *root* to your database user id.

$mysql_password = ""; Set your database password. NEVER HAVE A BLANK PASSWORD TO YOUR DATABASE!

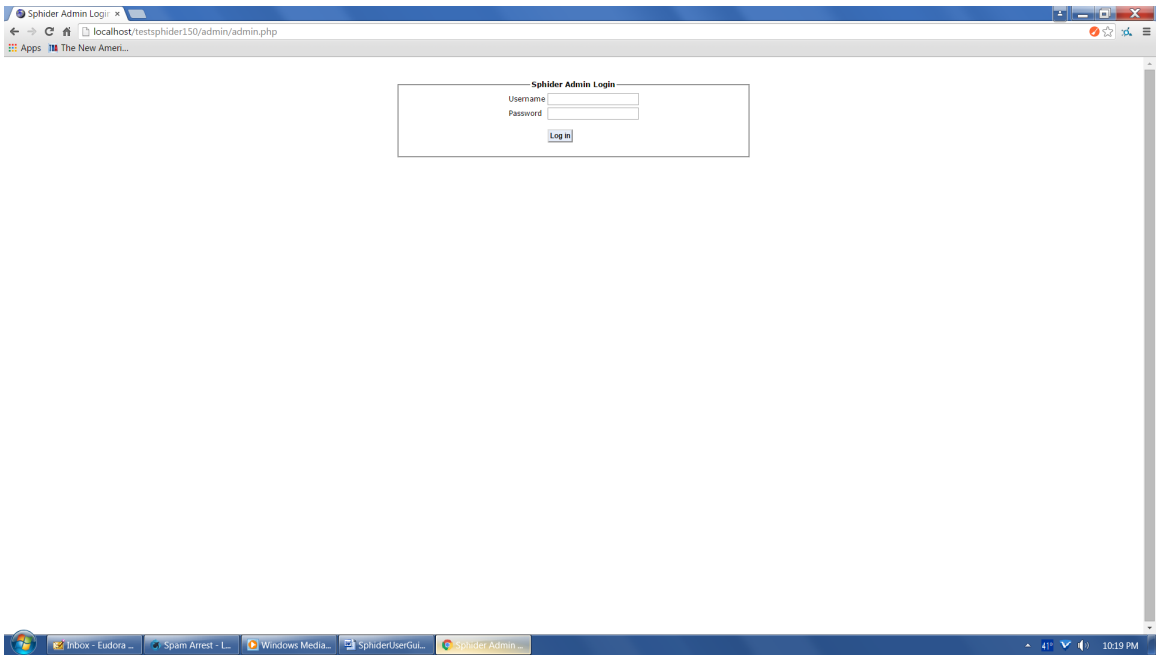
$mysql_host = "*localhost*"; Change *localhost* to your mysql host name, **if needed**. There are many cases when you will not need to change this.

$mysql_table_prefix = ""; A table prefix is optional. If used, the prefix will become part of the database table names. Be sure you set this BEFORE you create your tables or Sphider will not work. An example of when you would want to set a prefix would be if you have an existing database for your site and you do not wish to create another database, but just expand the existing one. To prevent any naming conflicts between Sphider tables and existing tables, you might want to create a prefix like "sph_500_". When you run the install script, your tables will have names like "sph_500keywords" and "sph_500_settings".

My.cnf

This file allows for efficient backup and restore of the database. The “host”, “user”, and “password” values should match data in **database.php**.

Auth.php



The auth.php scripts controls access to the admin panel. The default user and password are both set to "admin". YOU ARE HIGHLY ENCOURAGED TO CHANGE THESE!

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

These items are at the top of auth.php, lines 3 and 4 to be precise.

Auth.php is located in the [path_to_spider]/admin directory. Changing the user id and password are important to securing your Sphider installation. However, this in and of itself is insufficient. The ENTIRE [path_to_spider]/admin directory should be password secured.

To do so, cd (change directory) to [path_to_spider]/admin.
At the command promtp, type: htpasswd -c .htpasswd user_name (change user_name to who should have access to admin).
Hit ENTER. You will be prompted for a password. You will then be asked to re-enter the password.
Next, at the prompt, type: pwd <ENTER>

Record the result. It will be something like "/home/webuser/public_html/mysearch/admin".
Now open .htaccess for editing. Create it if id doesn't exist.

In .htaccess, put insert the following lines:

```
AuthType Basic
AuthUserFile "/the/complete/path/you/recorded/from/the/pwd/step"
AuthName "Admin Area"
require valid-user
```

Save and exit. The admin directory is now password secured.[2]

[NOTE: Some host providers do not permit this method of securing a directory, but will provide a way to do so through their Control Panel.]

There is still the risk that when you enter the user id's and passwords to first the directory, then to auth.php, that this data can be intercepted. Normal http access is not encrypted. If you have SSL for your site, You should add one additional line to .htaccess:

```
SSLRequireSSL
```

This will force https, and thus encryption, on your user ids and passwords. If you do not have SSL but can get SSL, do so. Even a free, self signed certificate will do. You probably won't want to use a self signed certificate for merchant activities, but it will secure your admin directory.

2 If you are using an Apache server (2.4 or later), htaccess may not work. You will need to edit apache2.conf, like this:

```
<Directory /var/www/html>
  Options Indexes FollowSymLinks
  AllowOverride All
  Require all granted
</Directory>
```

Creating your own templates

A number of templates are provided. The most important are “standard” and “mobile”. Regardless of template set in the configuration, the “mobile” template will be used if a search is run from a mobile device. If the appearance is not to your liking, the search.css file in [path-to-sphider]/templates/mobile can be edited to your liking.

The search.css file found for each template controls the look of your search pages and can easily modified. You can alter the layout, the font size, colors, background image if desired, or just a plain background. Borders may be changed or eliminated entirely.

If you are not satisfied with any of the pre-made templates to use on the search pages, it is easy to create your own. When doing so, using the provided “standard” template will serve as a guide.

In the [path_to_spider]/templates directory, create a new sub-directory. Because of the way Sphider is written, this sub-directory should contain ONLY lower-case alpha characters. This is the name of your new template. From the standard sub-directory, copy search.css and m_search.css to your new sub-directory. The search.css files are where you restyle your template. You can change backgrounds, font colors, sizes, and type. A working knowledge of CSS is needed to successfully make these changes. The m_search.css files contain the CSS used on mobile devices.

Preventing Sphider from indexing a page or parts of a page

Method 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.

Method 2 - Must include / must not include string list

A powerful option Sphider supports is defining a must include / must not include string list for a site (click on Advanced options in Index screen for this). Any url containing a string in the 'must not include' list is ignored. Any url that does not contain any string in the 'must include' list is likewise ignored. All strings in the string list should be separated by a newline (enter). For example, to prevent a forum in your site from being indexed, you might add www.yoursite.com/forum to the "must not include" list. This means that all urls containing the string will be ignored and wont be indexed. Using Perl style regular expressions instead of literal strings is also supported. Every string starting with a '*' in front is considered as a regular expression, so that '*/[a]+/' denotes a string with one or more a's in it.

Method 3 - Ignoring links

Sphider respect rel="nofollow" attribute in <a href..> tags in web pages, so for example the link foo.html in <a href="foo.html" rel="nofollow> is ignored.

Method 4 - Ignoring parts of a page

Sphider includes an option to exclude parts of pages from being indexed. This can, for example, be used to prevent search result flooding when certain keywords appear on certain part in most pages (like a header, footer or a menu). Any part of a page between **<!--sphider_noindex-->** and **<!--/sphider_noindex-->** tags is not indexed, however links in it are followed.

Indexing Tips

Sometimes indexing a site presents some messy issues you would like to avoid.

Lets say there is a page, http://www.yoursite.com/someinfo.htm, which you DO want indexed. However, you then discover that you are also indexing http://www.yoursite.com/someinfo.htm?option=this&option2=that. How do you stop this from happening? Simple. Edit the affected site, and in the URL must not include list, enter this line:
**/htm\?/*
If the page extension is .aspx instead of .htm, do this:
**/aspx\?/*

What if you have a situation where you have *http://www.yoursite.com/folder/index.htm*. You find that there is an entry for BOTH *http://www.yoursite.com/folder/* and *http://www.yoursite.com/folder/index.htm*. These would essentially be duplicates since *.../folder/* implies *.../folder/index.htm*. You can prevent this from happening by entering this line:
**#$\/$#*
in the URL must not include list.
One word of caution if you do this. This will exclude *http://www.yoursite.com/* as well! Set up your sites to always include the index.html (or .php, or .asx, or ...) at the end, thus, *http://www.yoursite.com/index.html*.

Often it assumed that EVERY directory has an "index.html". The truth is, most don't, so when an address like http://somesite.com/subdirectory/ is encountered, either a directory listing (not desirable) or a non-existent page is entered into the index. Many hosts provide an option NOT to display directory contents, but some don't. So how do you stop this? Another rule in the URL must not include box can fix this.
**#\/$#*
What this does is say, do ignore any url that ends with a "/". There IS a downside to this, and that is that "http://somesite.com/" will also be ignored! You can fix this by editing the starting address for your site to "http://somesite.com/index.html" (or index.php or index.aspx or whatever the homepage actually is named).

When clearing or deleting a site which has been indexed, the pending and all of the link-keyword tables are purged. If the site is being deleted, the images table is purged as well. However, the keywords table is NOT purged! Why? Because a keyword just may also be referenced in another site! It is advisable to go to the “Clean tables” tab and clean the keywords table of keywords with no associated site. It is also a good idea to clean the temp table, UNLESS you have an site in an “Unfinished” state.

About robots.txt

Sphider follows commands in a robots.txt file. There are things you need to know about how robots.txt files are constructed and the method Sphider uses to obey robots.txt.

By current standards, URL's in robots.txt are case sensitive. Sphider follows that standard. An example:

'disallow: /Images' and
'disallow: /images' are NOT the same thing.

Directives are also somewhat case sensitive. Permitted are:

'User-agent' or 'user-agent'
'Allow' or 'allow'
'Disallow' or 'disallow'
'Sitemap' or 'sitemap'

Sphider accepts the above, plus it even accepts 'User-Agent'. If the person who wrote you robots.tx is a caps happy Neanderthal and you have 'USER-AGENT', 'ALLOW', or 'DISALLOW', Sphider will have no idea what you are talking about and ignore the directive. Google might be more forgiving, but Sphider isn't.

Sphider does not, at this time, recognize the use of '?' or '$'.

Sphider DOES recognize the use of '*' (wildcard), but ONLY in 'disallow' directives. A wildcard in 'allow' opens up a whole new can of worms!

Sphider only recognizes TWO user agents: The user-agent name specified on the Settings tab, and the * user agent.

An 'allow: /' in the Sphider agent section overrides every single 'disallow:' in the * agent section.

General method of Sphider determinations:
1) Sphider-agent permits vs Sphider-agent denys:
Exact matches and we drop the deny (more permissive)
2) Star-agent permits vs Star-agent denys:
Exact matches and we drop the deny (more permissive)
3) Sphider-agent permits vs Star-agent denys:
Exact matches and we drop the Star-agent deny (more specific)
Special case: Sphider-agent "Allow: /" negates ALL Star-agent denys!
4) Sphider-agent denys vs Star-agent permits:
Exact matcfhes and we drop the Star-agent permit (more specific)
Special case: Sphider-agent "Disallow /" negates ALL star-agent permits!

About common text languages

Sphider will use common text language files to determine what words should NOT be indexed.

These files are located in 'include/common' and have names such as 'en_common.txt'. The user may edit these files as he/she feels the need. When editing these files, be sure to use a UTF-8 capable editor. There needs to be one, and only one, word per line. Be careful when editing in Windows. A utility like Notepad++ can be useful for specifying UTF-8 and UNIX line endings. Windows line ending introduce a lot of unnecessary garbage into the text file.

When spidering from the command prompt and using the -L option, only certain strings will be accepted as valid languages. Here is a list of language strings to use:

Language string	Language	Language string	Language
en	English	el	Greek
sq	Albanian	hi	Hindi
am	Amharic	hu	Hungarian
ar	Arabic	it	Italian
bn	Bengali	ja	Japanese
bg	Bulgarian	lv	Latvian
ca	Catalan	no	Norwegian
zh-cn	Chinese-Simplified	pl	Polish
zh-tw	Chinese-Traditional	pt	Portuguese
hr	Croatian	ro	Romanian
cs	Czech	ru	Russian
da	Danish	sr	Serbiasn
nl	Dutch	sk	Slovak
et	Estonian	sl	Slovenian
fa	Farsi (Persian)	es	Spanish
fi	Finnish	sw	Swahili
fr	French	sv	Swedish
de	German	tr	Turkish

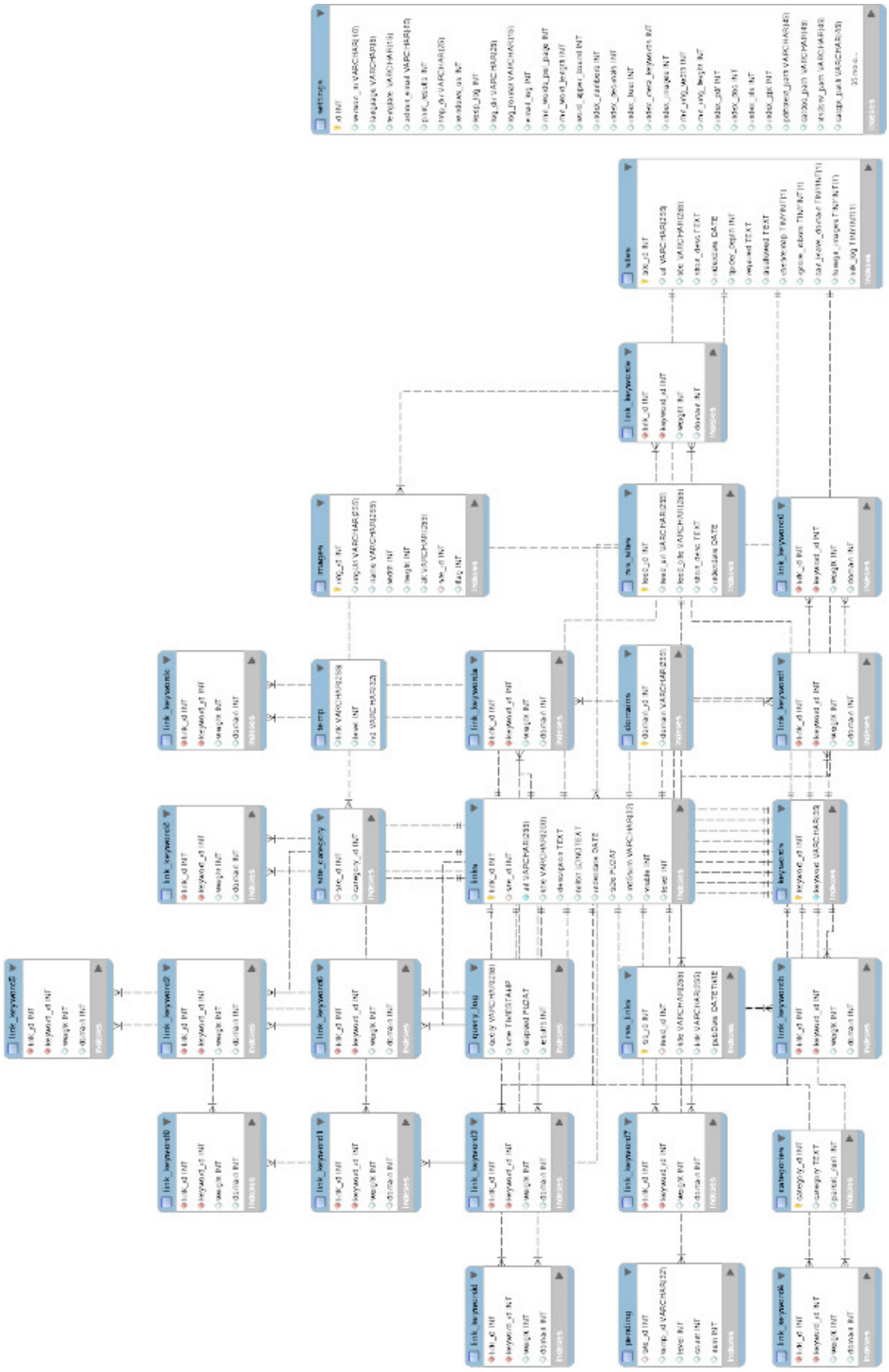
If no common text language is specified, English is the default.

Sphider can detect a language specified in the <html> tag, if such tag exists and a language specified. If that language is in the above list and is different than the user designated common text language, the page language will override the common text language for that page only.

An example of a language being designated on the page:

```
<html lang='de'>
```

ER DIAGRAM



LITE ER DIAGRAM