

Sitemap

Component for Joomla! 3, 4, 5 & 6

This manual documents **version 41.x** of the Joomla! extension.

https://www.aimy-extensions.com/joomla/sitemap.html

1 Introduction **3**

2 Supported Joomla! Versions **3**

3 Technical Requirements to Use Aimy Sitemap Generator **3**

4 Installing the Sitemap Extension **3**

5 Workflow **4**

6 Configuring the Sitemap Extension **4**
6.1 Aimy Sitemap: Preferences . . . 5
6.1.1 Tab: Aimy Sitemap . . . 5
6.1.2 Tab: Default Values . . . 6
6.1.3 Tab: Crawler . . . 6
6.1.4 Tab: Link Check (PRO Feature) . . . 7
6.1.5 Tab: Notifier . . . 7
6.1.6 Tab: Periodic Crawl (PRO Feature) . . . 8
6.1.7 Tab: Permissions . . . 8

7 The robots.txt File **9**
7.1 Editing . . . 9
7.2 Adding a Sitemap field . . . 9
7.3 How To Set Rules For Aimy Sitemap's Crawler . . . 10
7.4 Syntax Validation (PRO Feature) . . . 11
7.5 Further Information . . . 11

8 Crawling Your Joomla! Website **12**

9 Managing Your Index and Sitemap **14**
9.1 Setting Attributes . . . 14
9.2 Searching, Filtering and Sorting . . . 16
9.3 When to Re-Crawl Your Website? . . . 17

10 Notifying Search Engines **17**

11 Link Check: Handling Broken Links (PRO Feature) **18**
11.1 Exporting Broken Link Data . . . 20

12 Generating an HTML Sitemap **21**
12.1 Options . . . 21

13 Periodic Crawling (PRO Feature) **24**
13.1 Modern Console Plugin: Aimy Sitemap CLI . . . 24
13.2 Obsolete Approach: `crawl-website.php` . . . 24
13.2.1 Scheduling Automatic Crawls . . . 24
13.3 Crawling Information . . . 26

14 Special Cases **29**
14.1 Crawling HTTPS-only Websites . . . 29
14.2 Joomla! Installation in Subdirectory . . . 29
14.2.1 Rewrite Rule . . . 29
14.2.2 Sitemap Index File . . . 30
14.2.3 Robots.txt Directive . . . 30
14.3 Using a Canonical URL (PRO Feature) . . . 30
14.4 Using Aimy Sitemap on Large Websites . . . 30

15 Notes and Limitations **32**

16 Debugging and Errors **33**

17 Copyright & Trademark Notice **33**

1 Introduction

The Joomla! extension Aimy Sitemap generates an XML sitemap of your website for search engines. You may generate an HTML sitemap for your visitors as well. You can customize the sitemap with different options.

Aimy Sitemap generator includes a crawler that visits every page of your website, analyzes its content and queues its URL for inclusion in your sitemap.

This manual guides you through all steps necessary to install, configure and use the extension to enrich your website and increase your SEO results.

2 Supported Joomla! Versions

Aimy Sitemap supports Joomla! 3.9 and up, 4.0 and up, 5.0 and up and 6.0 and up.

Please note that the periodic crawling script included in the PRO version does not support Joomla! 6.0 and up - you can and should use the (free) Aimy Sitemap CLI console plugin on Joomla! 4.0 and up instead.

3 Technical Requirements to Use Aimy Sitemap Generator

Aimy Sitemap has no dependencies besides those of Joomla!, neither cURL nor url-fopen are required.

Aimy Sitemap supports **MySQL** and, in the PRO version, **PostgreSQL** databases and requires **PHP version 5.3.10 or higher** (like Joomla! 3 itself).

In order to crawl HTTPS websites, your PHP installation has to support SSL.

To automatically crawl your Joomla! website periodically, a system service like **cron** or **Task Scheduler** is required. Most hosters provide an interface to such services as well.

Writing an additional, compressed XML sitemap file (*sitemap.xml.gz*) requires your PHP installation to support Zlib compression, which is usually enabled by default.

4 Installing the Sitemap Extension

The installation of the sitemap extension follows the common Joomla! procedures.

In case you are not familiar with these procedures proceed as follows:

1. Download the extension's ZIP archive
2. Log into your Joomla! backend as "Super User"
3. From the menu, choose "*Extensions*" → "*Manage*" → "*Install*"
4. Click on the "*Or browse for file*" button and select the ZIP archive

The extension's archive will be uploaded and installed afterwards.

For further information, please have a look at the Joomla! documentation "Installing an Extension".

5 Workflow

The common workflow for generating a sitemap is as follows:

1. Configure Aimy Sitemap
2. Edit your robots.txt file (optional)
3. Crawl your Joomla! website
4. Fix broken links found on your Joomla! website (optional)
5. View and customize the index of your sitemap
6. Write the sitemap to your webspace
7. Notify search engines (optional)

Whenever you want to update your sitemap repeat steps 3 to 7.

If you use Aimy Sitemap PRO you may update your sitemap automatically with *Periodic Crawls*, explained later on in this manual.

6 Configuring the Sitemap Extension

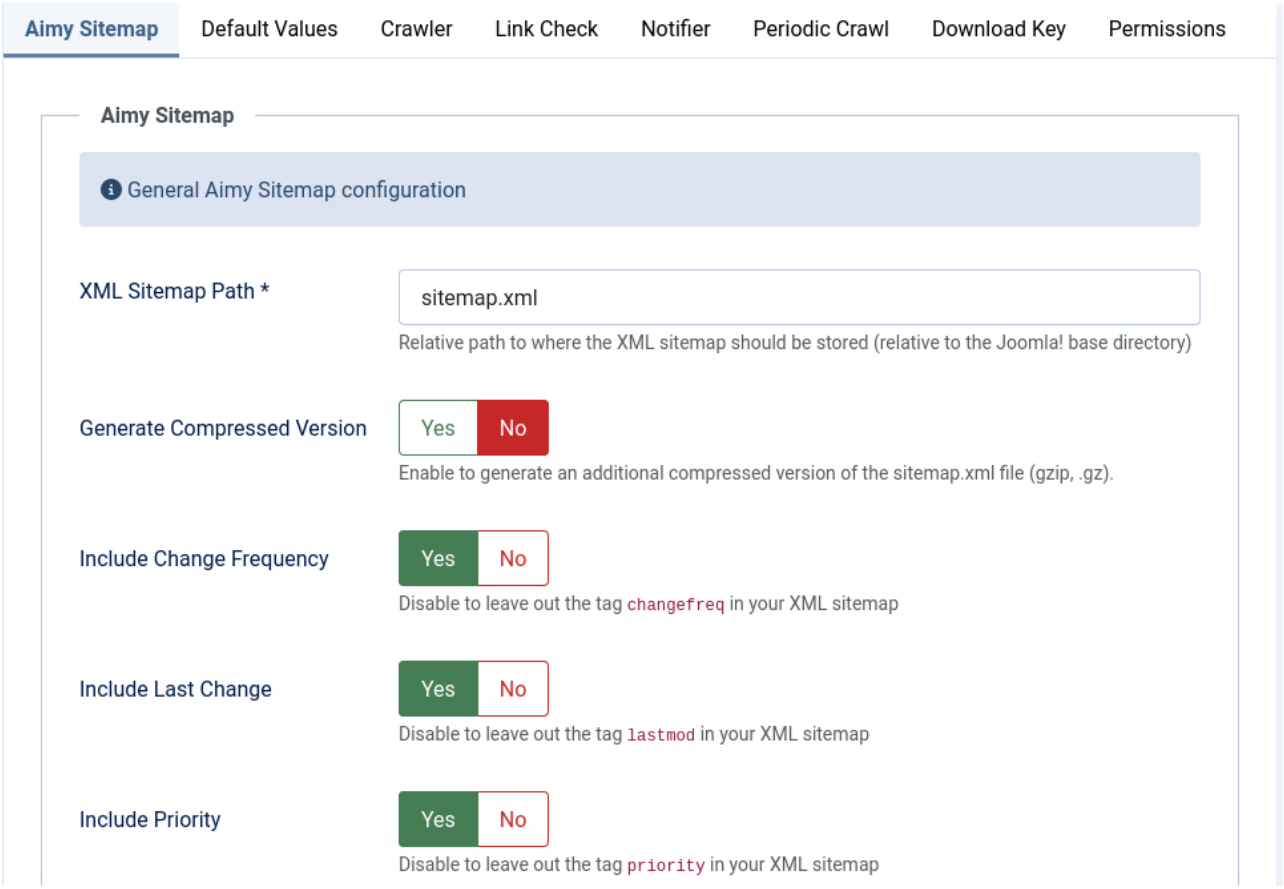
After a fresh installation, click on the "*Go to dashboard*" button on the installation report page.

At any time, you can reach Aimy Sitemap's configuration using the "*Options*" button available on Aimy Sitemap's dashboard or the "*Options*" button in the right corner of each view's toolbar.

You can customize the sitemap itself, the crawler, the notification of search engines and access permissions in a couple of ways. Let's have a look at the different tabs.

6.1 Aimy Sitemap: Preferences

6.1.1 Tab: Aimy Sitemap



- **XML Path**: Specify a custom path for your XML sitemap in case you do not want to use the common */sitemap.xml*.

  Any path is relative to the root directory of your Joomla! installation.

- **Generate Compressed Version** (PRO Feature): If enabled, an additional (gzip) compressed version of your *sitemap.xml* file is generated and stored to the path set in **XML Path** with an ".gz" suffix appended.

  Example: When **XML Path** is set to */sitemap.xml*, the compressed version of your sitemap is stored as */sitemap.xml.gz* on your webspace.

  Note: This feature requires your PHP installation to support *Zlib* compression, which is usually enabled by default.

- **Include Change Frequency** (PRO Feature): If enabled, a `changefreq` tag will be included for all URLs in your XML sitemap.

- **Include Last Change** (PRO Feature): If enabled, a `lastmod` tag will be included for all URLs in your XML sitemap.

- **Include Priority** (PRO Feature): If enabled, a `priority` tag will be included for all URLs in your XML sitemap.

- **Enable Debugging Mode**: If enabled, a debugging log file will be created. Some actions will not be taken (i.e. your website's index will not be updated after a crawl finishes). Have a look at **"Debugging and Errors"** for details.

6.1.2 Tab: Default Values

- **Priority**: Set a default value for the priority of your documents and items. With the priority you can give search engines a hint on the importance of your pages. Valid values range from 0.0 to 1.0.
- **Change Frequency**: Set a default value for how often the documents of your website change.
- **State**: Define whether *new* documents should be added to the sitemap by default. You may deactivate single items individually, see *"Updating your Sitemap"*.

6.1.3 Tab: Crawler

- **Include Images**: Choose how images should be handled by the crawler and whether they should be included in your website's index:
  - **Yes**

    If selected, all images found will be crawled and added to your website's index (unless excluded by an *robots.txt* entry or by an exclude pattern).
  - **Ignore HTML image elements**

    If selected, the crawler will ignore all images that are referenced by the HTML tag `img`. Images referenced otherwise (i.e. as a link) will be crawled and added to your website's index (unless excluded otherwise).
  - **No images at all**

    If selected, no images will be crawled or added to your website's index at all.

  Hint: If you like images to be crawled for link checking purposes but not be added to your website's index, choose "*Yes*" and add appropriate *Exclude Patterns* (see below).
- **Delay**: The crawler goes through your Joomla! website one resource after another. You may set a delay in seconds. The crawler will wait between each request so your server is not put under too high load.
- **Timeout**: You can define a timeout in seconds that sets the time, the crawler will at maximum wait for your webserver to respond to a request. If the timeout is reached, the crawler will abort.

  If your webserver is rather slow, you may want to increase this value to at least 30 seconds.
- **Retry Limit**: In case a resource is not available, you can define how often the crawler should retry to fetch the resource again.
  - **Retry Delay**: How many seconds should the crawler wait before retrying? This value is added to the *Delay* you may have defined in these options above.
- **Exclude-Pattern**: You may specify wildcard patterns, one on each line, that should be *excluded* from your sitemap. These patterns are case-sensitive. Regular expressions are not supported.

  Examples:
  - `*.gif` will exclude any file having the extension "`.gif`".
  - `*/sampledata/*` excludes any files that contain "`/sampledata/`" in their URL, like "`/images/sampledata/apple.jpg`".

- **Disguise as Browser**: Enable this option if your website delivers different content based on the User-Agent sent with a request. Aimy Sitemap's crawler will then crawl your website disguised as a *Firefox* browser.
- **Protocol** (PRO Feature): Select which protocol should be used to crawl your website.

  If "*Automatic*" is chosen, Aimy Sitemap's crawler will use the currently active protocol. That is the protocol you used to log into your Joomla! backend.
- **Use Browser Notifications** (PRO Feature): Enable this feature to get a notification on your desktop by your browser after a successful crawl.
- **Check Canonical** (PRO Feature): If enabled, the canonical links of your website's pages will be checked during a crawl:
  - **Yes, if link set**: if a page contains a canonical link, it is compared with the page's URL. If path and parameters do not match, the URL will not be added to the index.
  - **Yes, link required**: each page is required to contain a matching canonical link (path and parameters). Pages with a mismatch or no canonical link at all will not be added to the index.

Note: If you update the crawler's configuration at any later time, please crawl the website again to apply your changes.

Besides that you can enable support for **HTTP Authentication**:

- **Use HTTP Authentication** (PRO Feature): Enable this option if you want Aimy Sitemap's crawler to use HTTP Authentication (according to RFC 7617, "Basic HTTP Authentication Scheme") - if so, be sure to enter both username and password as well.

  WARNING: This feature is meant to be used during development of a (not yet public) website - take the displayed hints serious and only enable this feature if appropriate.
  - **Username**: The username that should be used for HTTP Authentication.
  - **Password**: The password that should be used for HTTP Authentication.

6.1.4 Tab: Link Check (PRO Feature)

- **Enable Link Check** (PRO Feature): Enable this feature to let Aimy Sitemap check for broken links on your website during crawls.

  Broken link information is then stored along with the URLs the links have been found on. The collected information can be viewed at any time by choosing "*Link Check*" from Aimy Sitemap's dashboard or "*Aimy Sitemap*" → "*Link Check*" from Joomla!'s components menu.

Note: If you update the link check configuration, please crawl the website again to let your changes take effect.

6.1.5 Tab: Notifier

Choose if search engines should be notified by Aimy Sitemap.

Note: As of January 2024, **Google** does no longer provide its XML sitemap ping API but recommends to submit a website's XML sitemap either *once* using Google Search Console or using a *Sitemap* field in `robots.txt`. The latter can easily be achieved by using the "*Add Sitemap field*" feature of Aimy Sitemap's `robots.txt` editor.

Note: As of August 2017, **Yandex** does no longer support ping notifications for sitemap changes. **Bing** shut down its support for XML sitemap pings in May 2022. However, both search engines support the new **IndexNow** protocol which can be used by Aimy Sitemap since v31.0 if **Aimy IndexNow PRO** is installed on your website (and support enabled in the options).

For more information on Aimy IndexNow please have a look at `https://www.aimy-extensions.com/joomla/indexnow.html`.

Note: The notification will not be done automatically. If you have generated or updated your sitemap, go to "*Components*" → "*Aimy Sitemap*" → "*Notify Search Engines*" and click the "*Start Notifying*" button to start the notification from the Joomla! backend.

6.1.6 Tab: Periodic Crawl (PRO Feature)

In the PRO version of Aimy Sitemap you can setup periodic crawls (see Chapter *Periodic Crawling*). In the options you can customize actions that are triggered whenever a periodic crawl finishes and new or updated content has been found on your website.

- **Write Sitemap**: Write an updated version of your `sitemap.xml` file(s) to disk.

  Note: If **Generate Compressed Version** is enabled in the *Aimy Sitemap* tab, a compressed version will be generated as well if **Write Sitemap** is enabled.
- **Notify Search Engines**: Notify all enabled search engines.

Note: Each action uses the options specified for the respective task so no additional configuration is necessary.

Note: These settings are only evaluated by the included `crawl-website.php` script, not by the Aimy Sitemap CLI console plugin.

6.1.7 Tab: Permissions

Manage the permission settings for different user groups.

Aimy Sitemap allows to set permissions for the following actions:

- **Access Administration Interface** allows users to view Aimy Sitemap's administration interface.
- **Configure** allows users to view and change Aimy Sitemap's configuration.
- **Edit** allows users to edit URLs on the "Manage URLs" page and change their state.
- **Crawl Website** allows users to use Aimy Sitemap's crawler on your website and, as a result, update the set of URLs.
- **Notify Search Engines** allows users to notify the configured search engines.
- **Write Files** allows users to initially write or update the configured **sitemap file** or the **robots.txt file** on your webspace.

For details, have a look at the official Joomla! Access Control List Tutorial:

`https://docs.joomla.org/J3.x:Access_Control_List_Tutorial`.

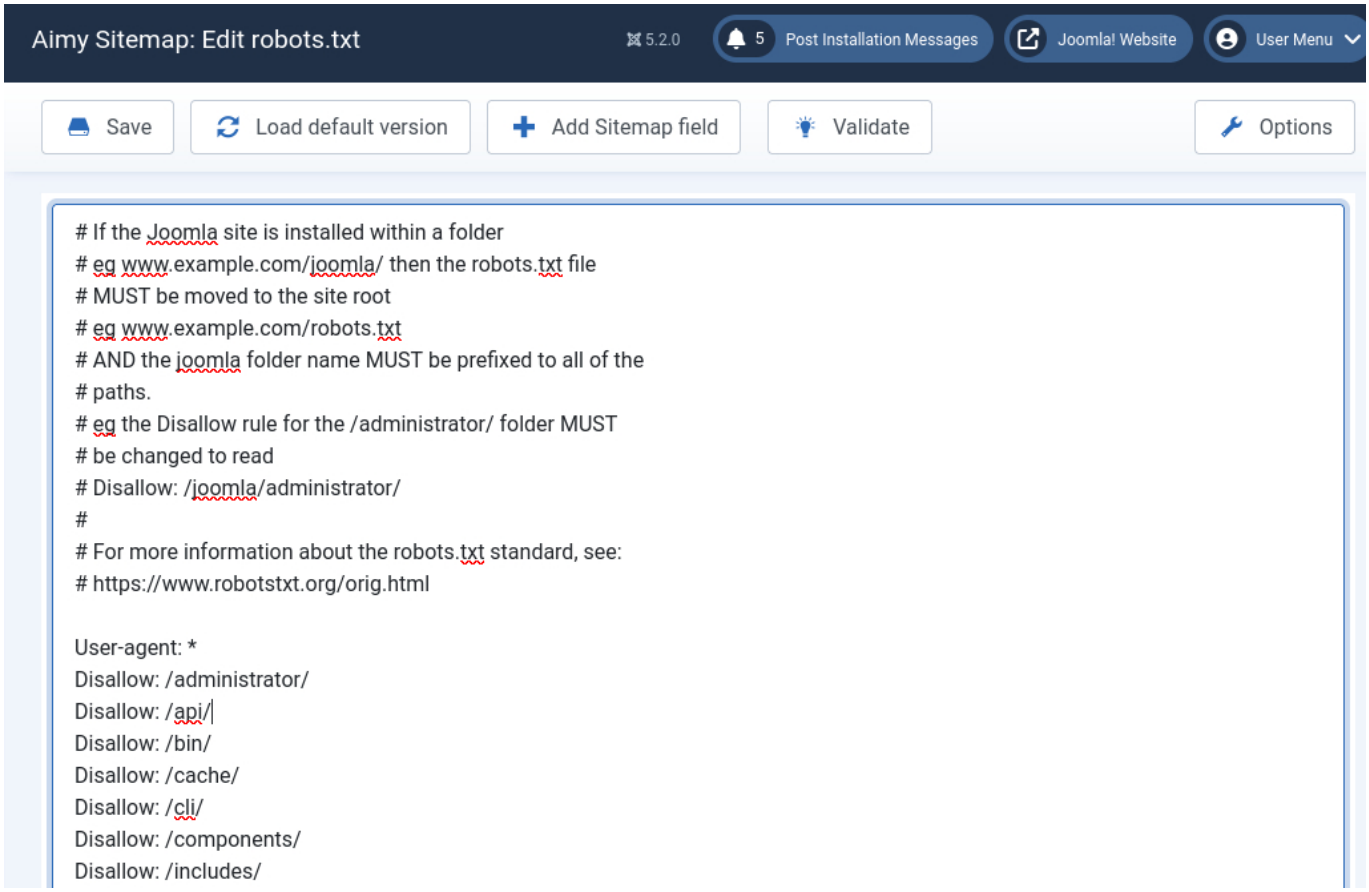
7 The robots.txt File

Aimy Sitemap comes with a simple `robots.txt` editor that allows you to set up your ruleset directly from Joomla!'s backend. You may use those rules to give instructions to Aimy Sitemap's crawler as well.

7.1 Editing

To edit your `robots.txt`, choose "*Edit robots.txt*" from Aimy Sitemap's menu. Your current file will automatically be loaded if present in the root directory of your Joomla! installation - otherwise a default version will be loaded. To write your changes to disk, click "*Save*".

At any time, you can click "*Load default version*" to load the default `robots.txt` file that comes with Aimy Sitemap.



7.2 Adding a Sitemap field

To automatically append a "*Sitemap*" field for your website, just click the "*Add Sitemap field*" button. It will automatically append a line like to following to your website's `robots.txt` file:

```
Sitemap:  https://www.YOUR-DOMAIN.com/sitemap.xml
```

This way visiting search engine bots will know where to find your website's XML sitemap and visit it.

7.3 How To Set Rules For Aimy Sitemap's Crawler

Aimy Sitemap will not crawl URLs that are either forbidden for any bot ("`User-agent: *`") or especially for Aimy Sitemap's crawler, called `AimySitemapCrawler`.

Use the following syntax in that case:

```
User-agent: AimySitemapCrawler
Disallow: /images/my-document.pdf
Disallow: /images/my-image.jpg
Disallow: /images/*.gif$
Disallow: /calendar/
```

This may be useful if your website uses extensions which generate a very large or even infinite set of onpage URLs (like some calendar extensions).

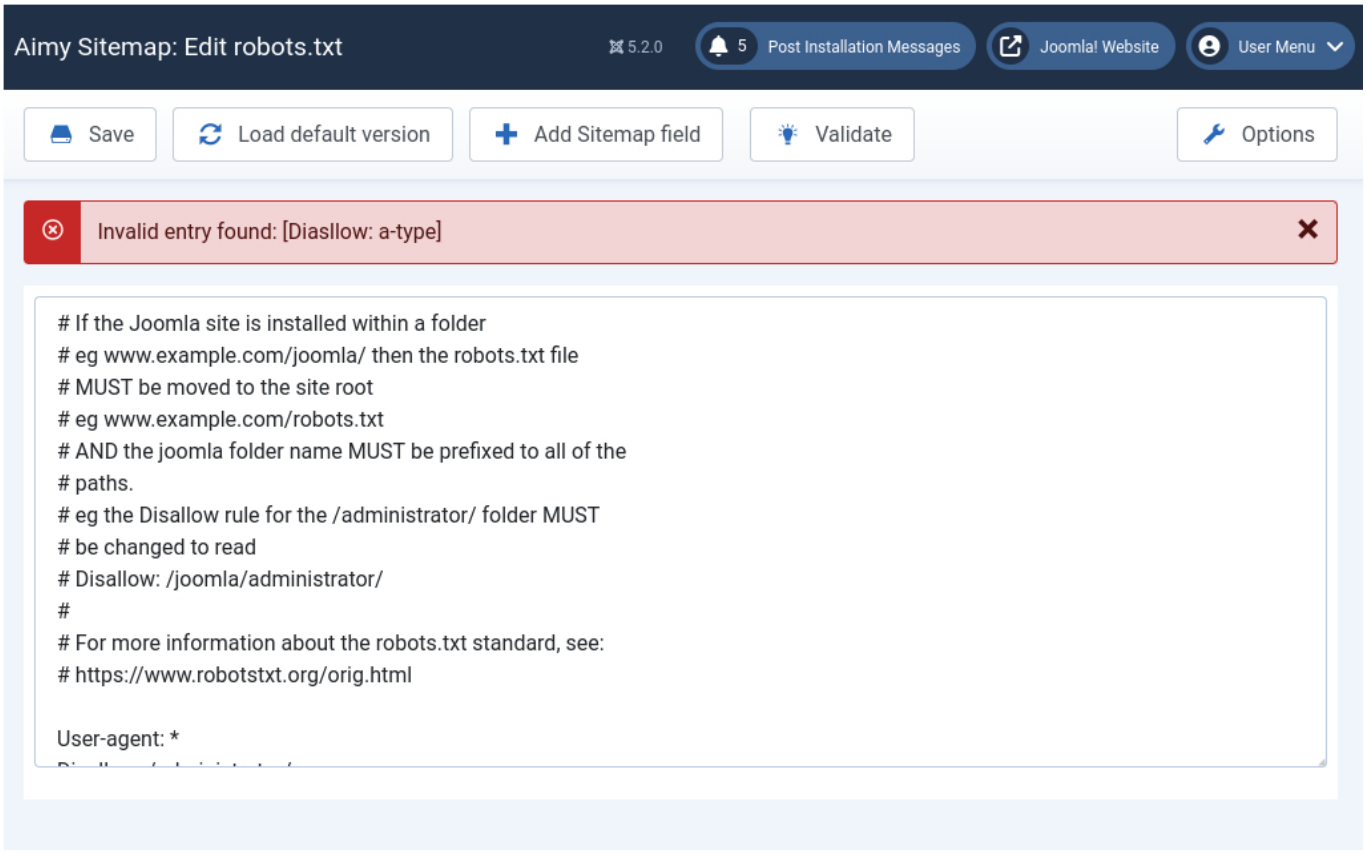
Aimy Sitemap supports patterns for Allow and Disallow directives as handled by Google's bot, that is, the special characters "*" and "$" are supported:

- ***** matches any character, zero or more times
- **$** denotes the end-of-string position

7.4 Syntax Validation (PRO Feature)

Aimy Sitemap PRO allows to easily validate the syntax of your website's `robots.txt` file. To start the validation, click the "*Save*" button if you did any changes to the file and click the "*Validate*" button afterwards.

If an error is spotted during validation, the syntactically invalid line will be shown in an error message:

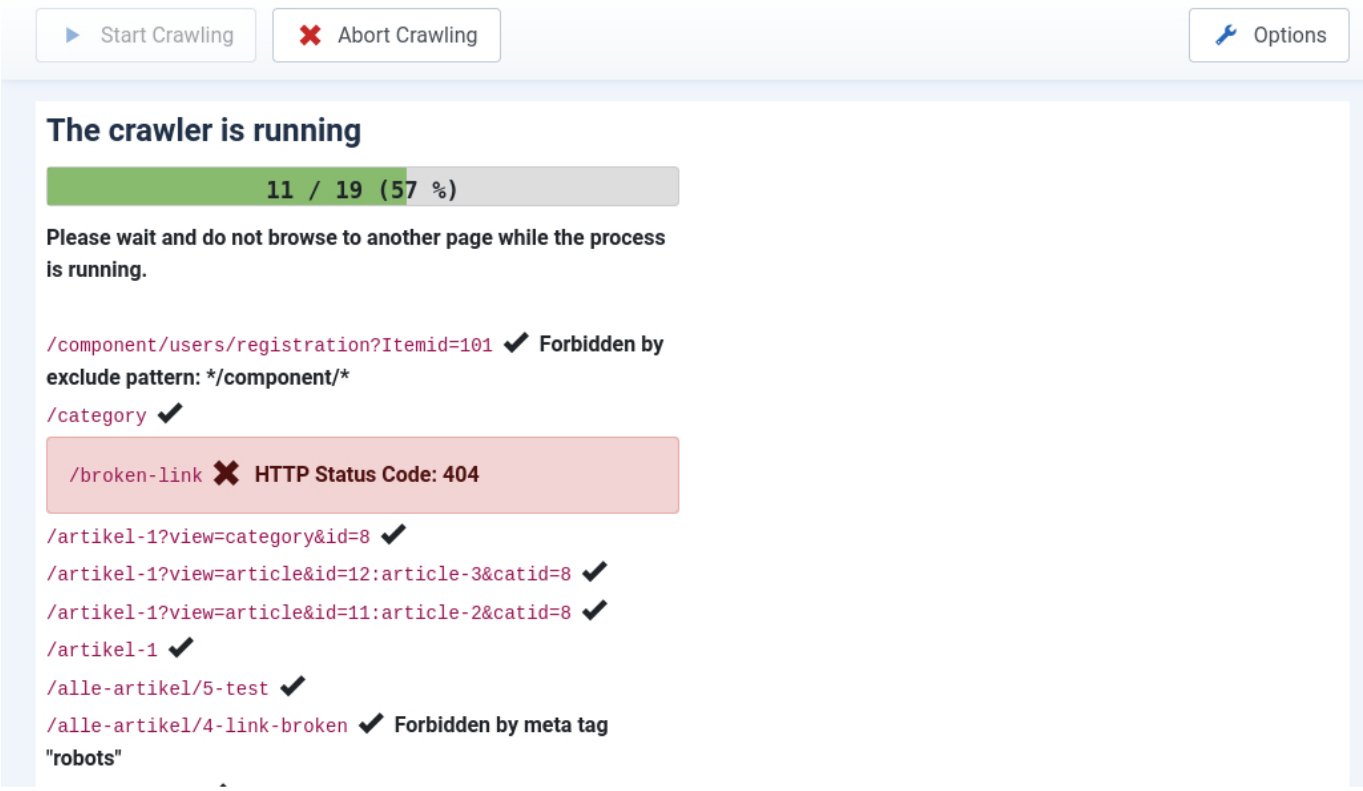


7.5 Further Information

For further information on the `robots.txt` file, visit robotstxt.org. For more details and examples on the supported pattern syntax, have a look at Google's documentation.

8 Crawling Your Joomla! Website

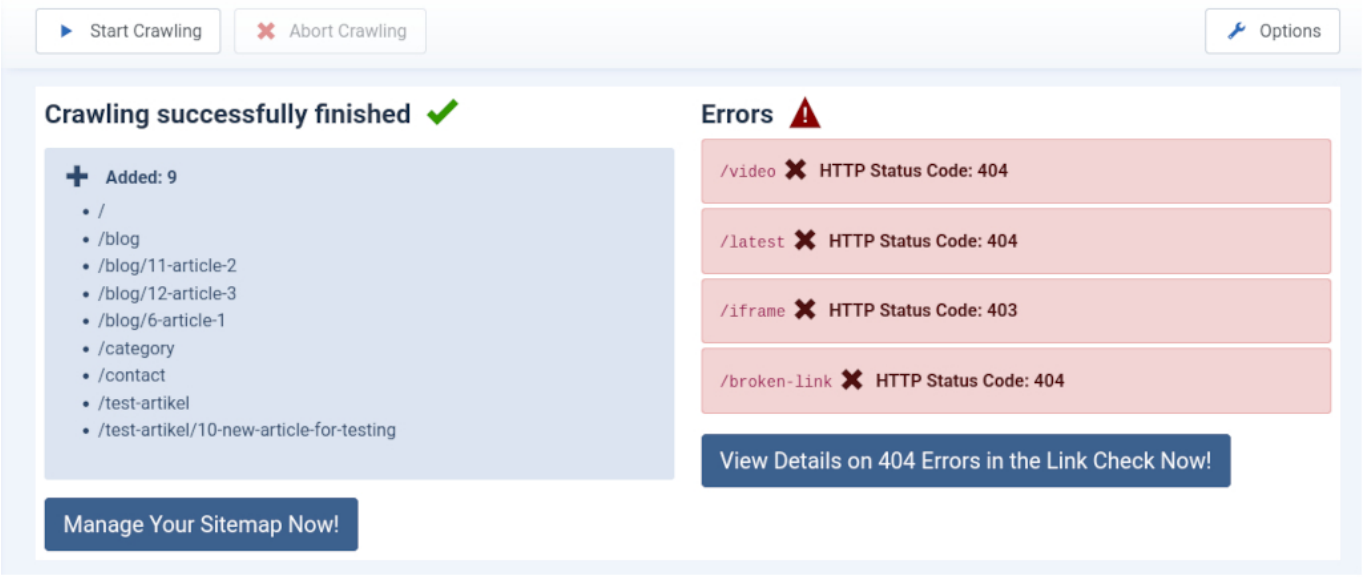
If you have just installed Aimy Sitemap on your Joomla! website, you will not see any entries in the **Manage URLs** view yet. You have to crawl your website first. To do so, choose "*Components*" → "*Aimy Sitemap*" → "*Crawl Website*" and click the "*Start Crawling*" button in the toolbar.



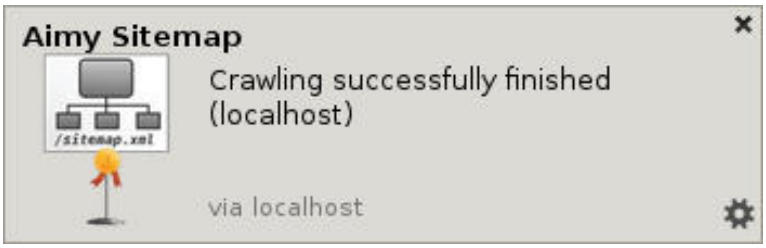
Note: A **progress bar** (as shown in the screenshot above) is included in the PRO version of Aimy Sitemap - the free of charge version of Aimy Sitemap comes with an animated image that indicates crawling activity only.

Note: If you **abort a crawl** you can easily resume it later on by clicking the "*Start Crawling*" button once more - you will be asked whether to **resume the previous crawl** or start a new one.

After crawling has finished, you can view all URLs included in your index by clicking "*Manage Your Sitemap Now*" or by choosing "*Manage URLs*" from Aimy Sitemap's menu or dashboard. All errors detected during a crawl are listed on the right side – all HTTP 404 ("*not found*") errors can be viewed in detail in the "*Link Check*" view (see chapter "*Link Check: Handling Broken Links*").



If you enabled "*Use Browser Notifications*" in Aimy Sitemap's "*Options*", a browser notification will be sent after a successful crawl and displayed on your desktop (PRO version).



If you have finished looking through the list and configuring your individual URLs settings, click "*Write Sitemap*" to write the XML sitemap to your webserver's disk.

9 Managing Your Index and Sitemap

The "*Manage URLs*" view allows you to have a look at the index generated by Aimy Sitemap's crawler and manage your sitemap by setting attributes of URLs and by selecting which URLs should be included in your final sitemap.

If you discover items you do not want to display in the sitemap after crawling, **disable** the items in the list. Click the button in the "*State*" column for any single file or use the checkboxes and change the state of multiple files at once.

Next time you click the "*Write Sitemap*" button, all active URLs will be written to the sitemap file.

Good to know: If you upgrade from Aimy Sitemap to Aimy Sitemap PRO all your settings will be kept.

9.1 Setting Attributes

You may set the following attributes individually per document:

- State / Include in Sitemap

  Set whether the document should included in your sitemap.

- Priority

  Set the priority of the document for your website's content, ranging from 0.1 (less important) to 1.0 (very important).

- Change Frequency

  Set how often the document is regularly changed.

- Document Language (PRO Feature)

  This attribute affects the HTML view of the sitemap for multi-language websites. See *Generating an HTML Sitemap*.

- Title (PRO Feature)

  The title is relevant for the HTML view of the sitemap. If you do not like the automatic title, you may set it manually.

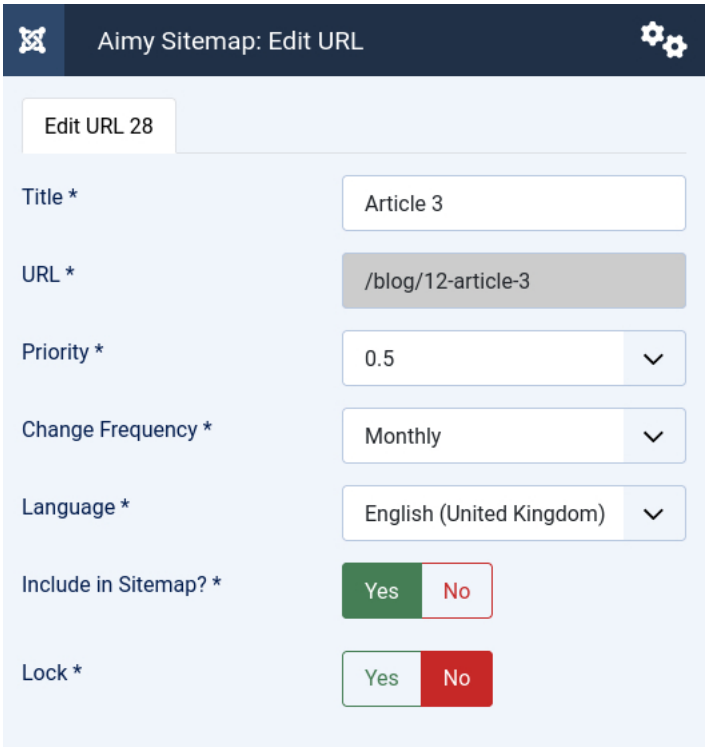
- Lock

  If you manually changed a document's language or title, lock the data set. Otherwise it will be updated with the next crawl.

Note: If you have manually set either language or title, be sure to set the lock attribute to "*Yes*" in order to disable automatic updates.

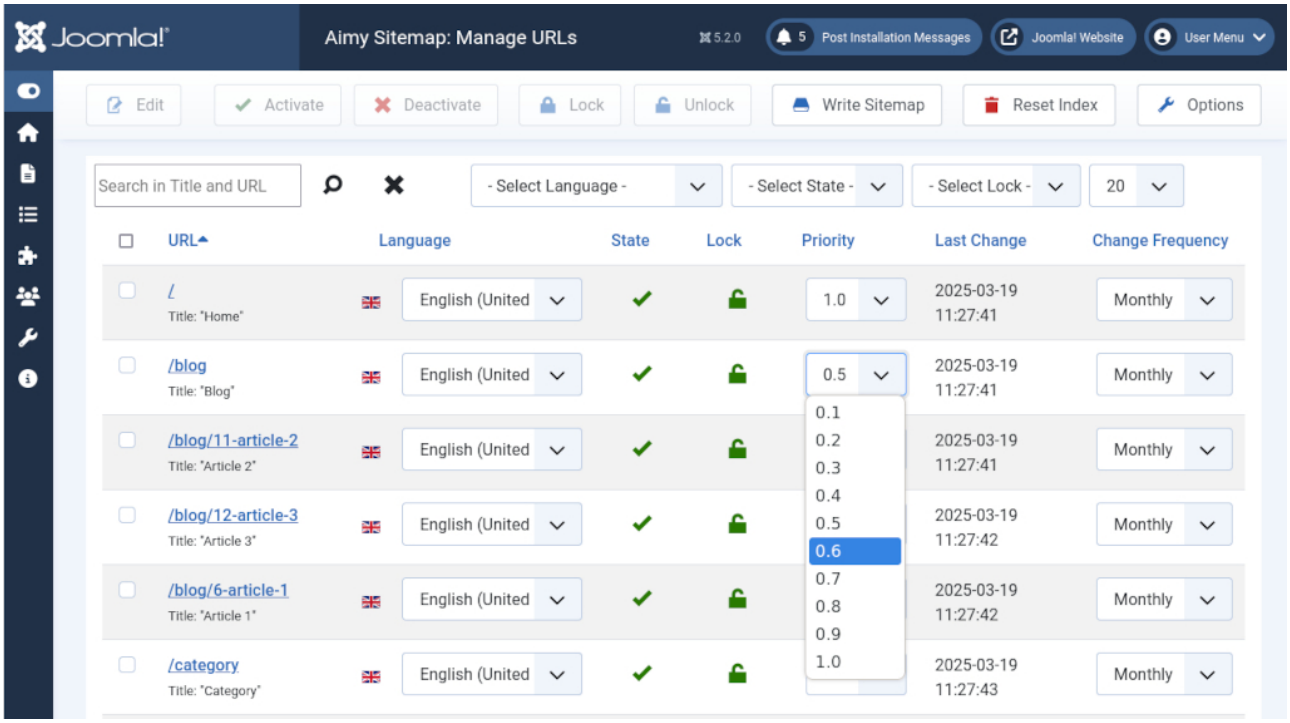
The settings can be done either in

1. a **dedicated edit view** for each entry reached by clicking on the URL or selecting the UR'Ls checkbox and clicking the *Edit* button in the toolbar or in



2. the **quick edit mode** that allows you to set your URLs' attributes directly from the URL list.

   After you changed a value it will be saved automatically in the background. Failure is indicated by a red background color of the changed field.



9.2 Searching, Filtering and Sorting

Aimy Sitemap allows you to search and filter your URLs to select just the set you want to work on. The following tools are available and combinable:

Search in Title and URL

Type your term into the **search field** and either press ENTER or click the magnifier button next to the field to start the search. The set of shown URLs will be updated accordingly.

To reset your search, click the "X" button next to the **search field**.

Filter by Language (PRO Feature)

This filter allows you to select URLs from their document's language. Your selection will take effect automatically.

To reset your selection, select "-Select Language-" from the drop-down list.

Filter by State

This filter allows you to select whether to show "activated", "deactivated" or "all" URLs. Your selection will take effect automatically.

To reset your selection, select "All" from the drop-down list.

Filter by Lock (PRO Feature)

This filter allows you to select whether to show "locked", "unlocked" or "all" URLs. Your selection will take effect automatically.

To reset your selection, select "All" from the drop-down list.

Sort by Column

To sort your URLs by the value of one of the columns, click the column's heading. By clicking the heading once, the URLs will be sorted ascending, a second click sorts descending. A small caret will be displayed right to the column's name indicating the sorting direction.

The "*URL*" and "*Change Frequency*" columns are sorted alphabetically, "*Priority*" is sorted numerically and "*Last Change*" is sorted by date.

Pagination

For convenience, the URLs of your index will be displayed grouped by virtual pages. You can customize the amount of URLs per page. Navigate through the pages using a toolbar below the list of URLs.

You can set a custom amount of URLs per page by selecting your desired number from the drop-down box in the upper right just above the list of URLs.

9.3 When to Re-Crawl Your Website?

If you have

- changed your content (renamed, added or deleted articles / fixed broken links) or
- changed the crawler configuration (e.g. exclude patterns)

crawl your Joomla! website again, view the result and click "*Write Sitemap*".

10 Notifying Search Engines

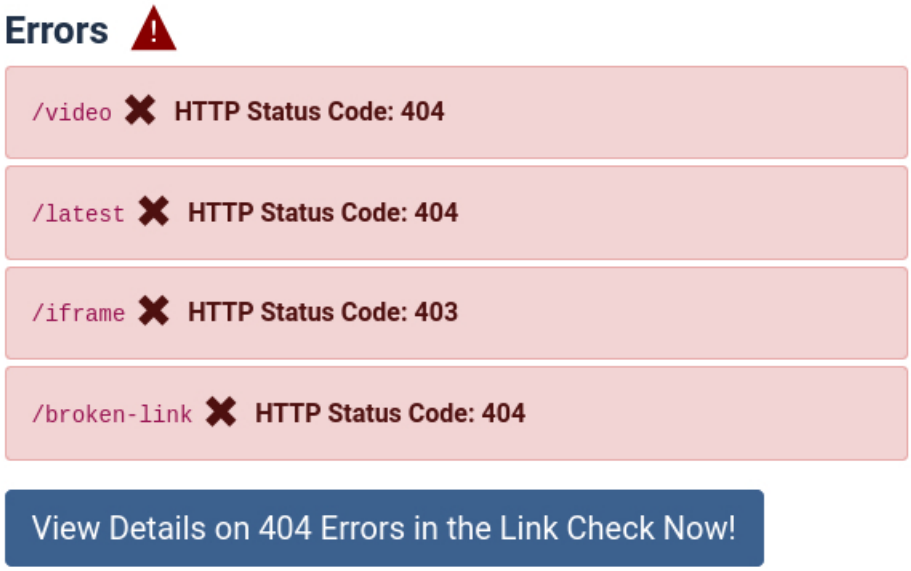
Note: As of January 2024, **Google** does no longer provide its XML sitemap ping API but recommends to submit a website's XML sitemap either *once* using Google Search Console or using a *Sitemap* field in `robots.txt`. The latter can easily be achieved by using the "*Add Sitemap field*" feature of Aimy Sitemap's `robots.txt` editor.

Note: As of August 2017, **Yandex** does no longer support ping notifications for sitemap changes. **Bing** shut down its support for XML sitemap pings in May 2022. However, both search engines support the new **IndexNow** protocol which can be used by Aimy Sitemap since v31.0 if **Aimy IndexNow PRO** is installed on your website (and support enabled in the options).

So you can still notify search engines about new or updated URLs. To do so, choose "*Notify Search Engines*" from the menu and click "*Start Notifying*" on the toolbar.

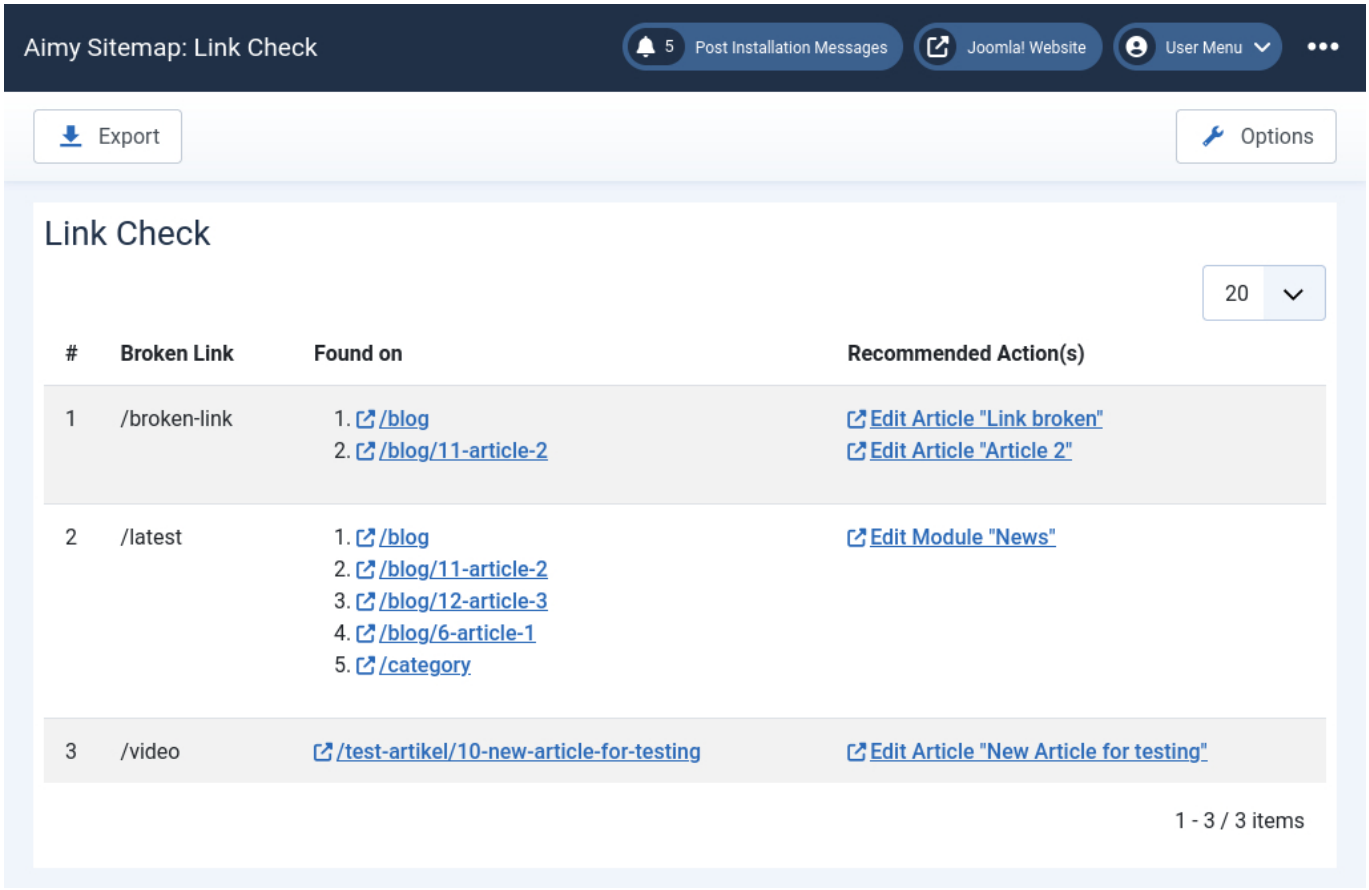
11 Link Check: Handling Broken Links (PRO Feature)

During a crawl Aimy Sitemap PRO collects information on broken links found on your website if the "*Enable Link Check*" option is set. Please note, that only **internal links** are checked that return **status code 404** (not found). After a crawl the crawling report shows a button which leads you to the report on broken links if broken links have been found:



You can view the report at any time by choosing "*Components*" → "*Aimy Sitemap*" → "*Link Check*" from Joomla!'s components menu or from Aimy Sitemap's dashboard.

The link check report lists all broken links found on your website along with the page or pages **the link has been found on** and a set of **recommended actions**. This way, finding and fixing those links is quite simple to do even on larger websites.



To help fixing the broken links, Aimy Sitemap tries to locate each broken link's resource and provides a link that allows to edit those resources easily. Currently, the following Joomla! resources types are supported:

1. Articles (`com_content`)
2. Categories (`com_content` / `com_categories`)
3. Modules (`com_modules`)

Note: To maximize the likeliness a broken link can be automatically detected in a supported resource type, we recommend to enable SEF for your website.

Note: Recrawling your website after you fixed broken links is optional, but recommended to make sure no broken links are left.

11.1 Exporting Broken Link Data

Aimy Sitemap allows to export all broken link data as a file in **Comma-Separated Values** format (CSV).

The set of fields contained in a broken link export consists of:

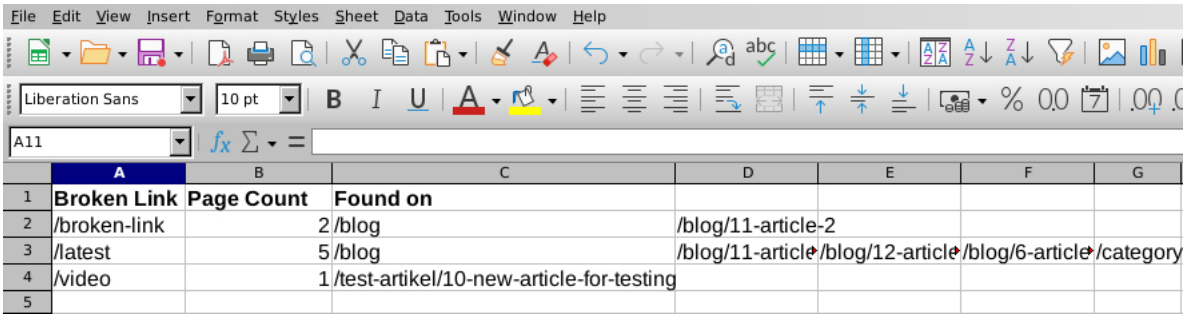
1. the link ("*Broken Link*")
2. the count of pages the link has been found on ("*Page Count*")
3. (4., 5., ...) the list of pages ("*Found on*")

The semicolon character ("`;`") is used as a separator. The first line provides headings for the first three fields.

The following shows the content of an example export:

```
Broken Link;Page Count;Found on
/images/missing-image.jpg;2;/about-us.html;/pictures.html
/missing-page.html;3;/;/about-us.html;/services.html
```

You can easily use this file to import the broken link data in your favorite **spreadsheet** application (i.e. Microsoft Excel or LibreOffice Calc):



	A	B	C	D	E	F	G
1	**Broken Link**	**Page Count**	**Found on**				
2	/broken-link	2	/blog	/blog/11-article-2			
3	/latest	5	/blog	/blog/11-article	/blog/12-article	/blog/6-article	/category
4	/video	1	/test-artikel/10-new-article-for-testing				
5							

12 Generating an HTML Sitemap

Aimy Sitemap provides different HTML views of your sitemap. If you want to display an HTML sitemap on your website, add it as a menu item by clicking:

"*Menus*" → "*YOUR MENU*" → "*Add New Menu Item*"

As *Menu Item Type* choose "*Aimy Sitemap*" → "*HTML-Sitemap*".

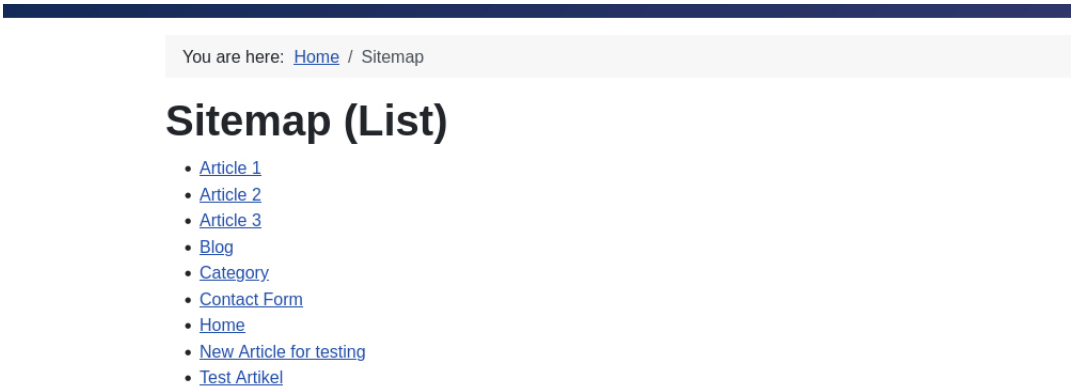
12.1 Options

Aimy Sitemap's HTML views allow you to set the following options beside the standard Joomla! options for a menu item:

Variant

List

All URLs are displayed as a simple top-to-bottom list, sorted alphabetical by their title.



Index

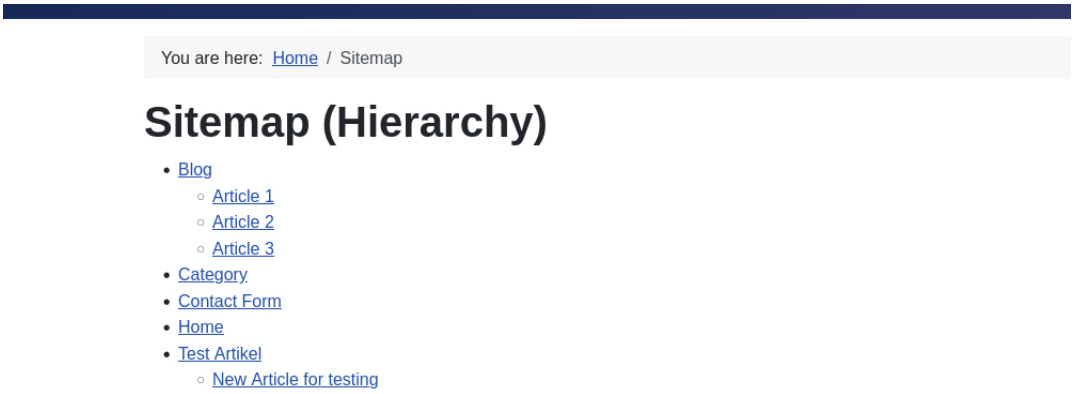
All URLs are displayed in blocks named after their first letter. Both blocks and the contained links are sorted alphabetical by title.



Hierarchy (PRO Feature)

URLs are displayed hierarchically, with each collection being sorted alphabetically.

Note: This feature requires that SEF URLs are enabled.



Notes On Styling

You can use CSS to style the list elements.

Levels

Each `ul` element provides a CSS class that reflects its level. The class is named `aimysitemap-lvl-$L`, with `$L` being the numeric level, i.e. `aimysitemap-lvl-2`.

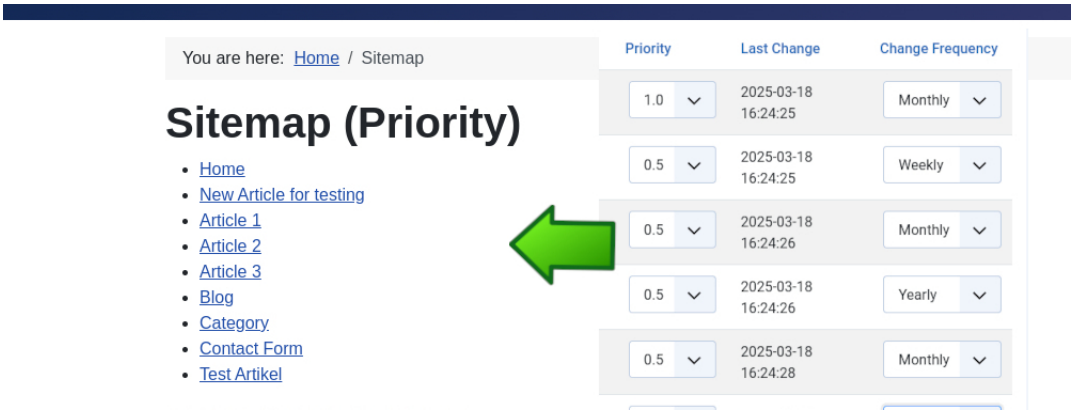
The top-level `ul` element has a level of 1 (hierarchy root).

Languages

Each link that has an assigned language does provide this information as a CSS class as well. The class is named `aimysitemap-lang-$L`, with `$L` being the language code in lowercase letters, i.e. `aimysitemap-lang-de-de` for a link to a German document and `aimysitemap-lang-en-gb` for an English one.

Priority (PRO Feature)

URLs are displayed in descending order of their assigned priority. URLs containing the same priority are sorted alphabetically by title.



Prevent Duplicate Titles

If this option is set to *Yes*, all URLs will be merged by their title so that no title can be present on your HTML sitemap more than once.

Note: If this option is on, the *shortest URL* of the set of URLs sharing the same title will be selected.

BACKGROUND: In some cases Joomla! generates multiple URLs for a resource. As a result, Aimy Sitemap's crawler adds these resources to the index multiple times as well. That's technically correct. As these URLs will likely have the same title, it may seem wrong

to your human visitors to show them more than once in your HTML sitemap.

Alternatively change the titles manually in the *Manage URLs* view (PRO Feature).

Remove Emojis (PRO Feature)

If this option is set to *Yes, Leading* leading emojis and symbols will be removed from linked page titles. *Yes, All* will remove all emojis and symbols. *No* will keep the titles untouched.

Alternatively change the titles manually in the *Manage URLs* view (PRO Feature).

Container Style

Use this option to set the style of container classes used to match those used by your template: choose between *Bootstrap 2* and *Bootstrap 3/4*.

Filter by Language (PRO Feature)

If this option is set to *Yes*, Aimy Sitemap will only show entries that match the frontend language. You see - and may change - the language of the entries in *Manage URLs*.

Exclude Images (PRO Feature)

If this option is set to *Yes*, entries that reference a file with a common image extension (i.e. `jpg`, `png`, ...) will be excluded from the HTML sitemap.

Exclude Documents (PRO Feature)

If this option is set to *Yes*, entries that reference a file with a common document extension (i.e. `pdf`, `rtf`, `txt`, `doc`, `xls`, `ppt`, ...) will be excluded from the HTML sitemap.

Exclude Multimedia Files (PRO Feature)

If this option is set to *Yes*, entries that reference an file with a common multimedia extension (i.e. `mp3`, `mp4`, `ogg`, ...) will be excluded from the HTML sitemap.

Show Credits (PRO Feature)

If this option is set to *Yes*, a short credits paragraph ("*Generated by Aimy Sitemap for Joomla!*") containing a link to Aimy Sitemap's website will be shown below the HTML sitemap. If set to *No* the paragraph will be omitted.

The titles for the sitemap are extracted from the page's title tag automatically. In the PRO version of Aimy Sitemap you may set the title manually as well (see *Managing your Index and Sitemap* → *Setting Attributes*).

13 Periodic Crawling (PRO Feature)

Aimy Sitemap PRO supports crawling your website automatically. After such a **periodic crawl** your *Index* is updated. Additionally the `sitemap.xml` file may be updated and search engines notified if new or updated content is found during the crawl. These actions depend on your settings in the *options*.

Information about a currently running periodic crawl can be viewed in Joomla!'s administration backend. Once a periodic crawl finishes, it stores detailed information about its results, changes and found errors which is then available for review in the backend as well ("Components" → "Aimy Sitemap" → "Periodic Crawl").

There are two option available to crawl your website from the command line: a modern Joomla! console plugin and an (obsolete) script.

13.1 Modern Console Plugin: Aimy Sitemap CLI

Aimy Sitemap CLI is a free Joomla! console plugin that allows to crawl your website from the command line. It is not bundled with Aimy Sitemap so you need to download and install it seperately:

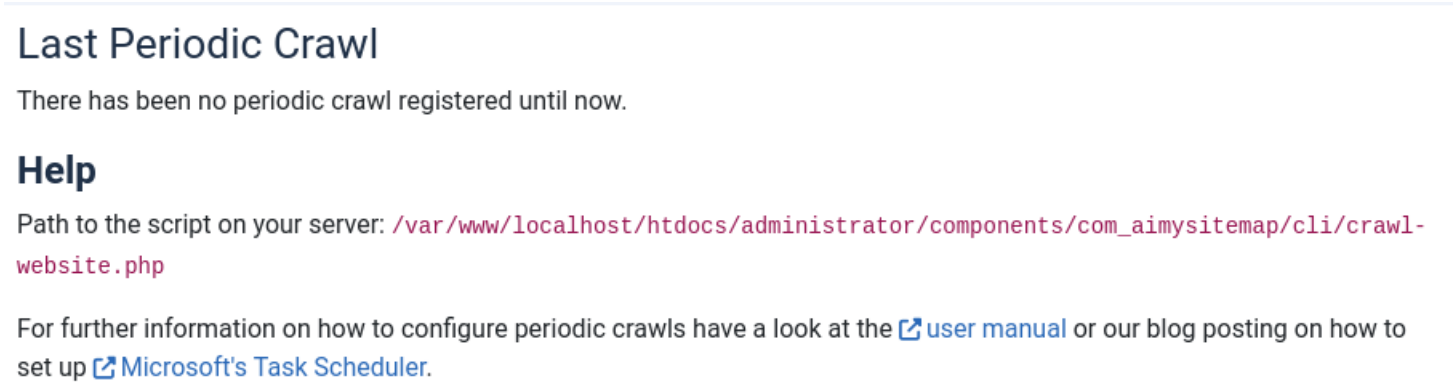
`https://www.aimy-extensions.com/joomla/sitemap-cli.html`

The modern CLI plugin requires at least Joomla! 4 and comes with its own user manual.

13.2 Obsolete Approach: `crawl-website.php`

The Aimy Sitemap extension includes a command line script named `crawl-website.php` that allows periodical crawling. It is still supported and runs on Joomla! 3 to 5. In order to execute this script, you first have to know its absolute path.

To look the path up, choose "Components" → "Aimy Sitemap" → "Periodic Crawl" in your Joomla! administration backend. The absolute path on your server is shown under "*Help*":



13.2.1 Scheduling Automatic Crawls

In order to schedule an automatic crawl of the sitemap generator, a **system service** is required. On a Unix or Unix-like system (i.e. Linux, FreeBSD, Mac OS X, Solaris) **cron** is usually used for

such tasks. On a Microsoft Windows server **Task Scheduler** is available. Most hosters provide an interface to schedule jobs for periodical execution as well.

Whatever way you choose to start the periodic crawl, be sure to always select the **Command Line Interface** (CLI) version of the **PHP interpreter** to run the script. For security reasons, the script will refuse to run if the Common Gateway Interface (CGI) version is used as an interpreter.

If in doubt, ask your webhoster, webmaster or administrator for the path of the command line interface PHP interpreter.

Setting Up a Cron-Job (Unix) If you have a Unix or Unix-like server with shell access, log into your server using the user account that should be running the automatic crawl periodically. This may be the same user your Joomla! website is running as.

First, determine the path of your PHP **CLI** interpreter. If you do not already know where the binary is located, run the following command to find out:

```
which php
```

Then, add a new cron-job running the following command:

```
crontab -e
```

Your default editor will open up, allowing you to enter a new cron-job. Each cron-job consists of the following six required fields: *minute*, *hour*, *day of month*, *month*, *day of week*, *command*. The special value of "*" can be used as placeholder for any valid value in one of the first five fields, i.e. any hour or any day of the week.

Set up your cron-job to be run at your desired time and be sure to use the full path to your PHP binary followed by the full script path as a *command*.

To suppress all non-error output during execution, you can pass either `-q` or `--quiet` as an argument.

Save your changes and exit your editor afterwards.

Examples:

To run your periodic crawls each night at 23:55 you would enter:

```
55 23 * * * $full_php_path $full_command_path
```

Note: Replace `$full_php_path` with the full path of your PHP interpreter (i.e. `/usr/bin/php`) and `$full_command_path` with the path you previously looked up in Aimy Sitemap's "*Periodic Crawl*" view.

To run your periodic crawls each Sunday at noon, suppressing all non-error output, you would enter:

```
00 12 * * sun $full_php_path $full_command_path --quiet
```

Setting Up a New Task (Windows) On a Microsoft Windows server, **Task Scheduler** can be used to schedule automatic website crawls periodically. To start the utility, click "Start" → "Accessories" → "System Tools" → "Task Scheduler".

Create a new "Basic Task" and enter your desired configuration.

When asked what *Action* should be performed, choose "*Start a program*". Choose your PHP interpreter as a program and enter the full path to the `crawl-website.php` script in the "*Add arguments*" field below.

To suppress all non-error output, add `--quiet` as an argument (after the script's path) in the "*Add arguments*" field.

Verify your settings in the next mask and save the task by clicking "*Finish*".

For an in-depth step-by-step howto on how to set up a task for Aimy Sitemap's periodic crawling script using Microsoft's Task Scheduler have a look at our website.

Using Your Hosters Interface Most hosters provide an interface to the periodic command execution system services in their management frontends. First, verify that your hoster provides such an interface for your website and have a look at your hoster's documentation.

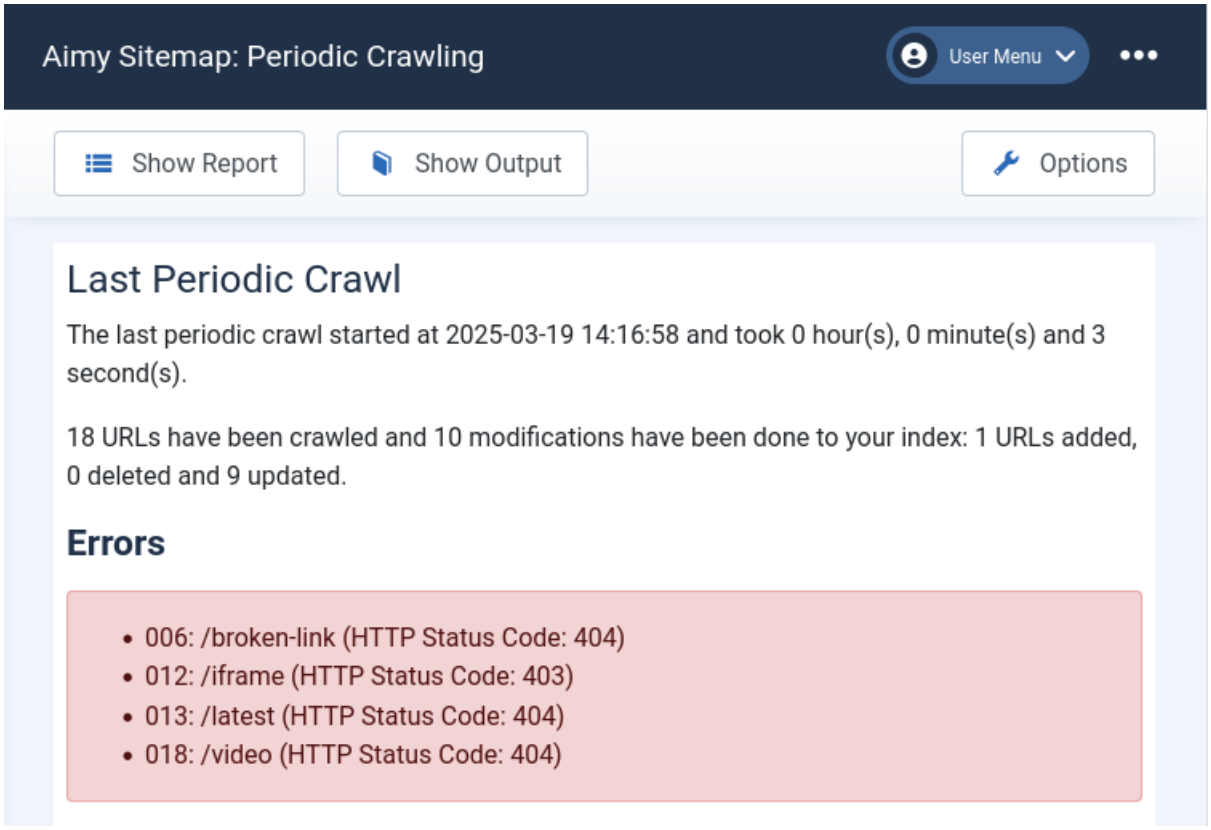
When asked for it, choose the highest version of PHP available (i.e. 5.6) as an interpreter and use the full path of the `crawl-website.php` script as periodic task.

By default, PHP processes started from the command line do not have a time limit set. However, if your hoster has overwritten this default setting (`max_execution_time`), be sure to either restore the default (0) or at least set the highest limit available for the directive (>60) to decrease the probability that the `crawl-website.php` script will get aborted by your system.

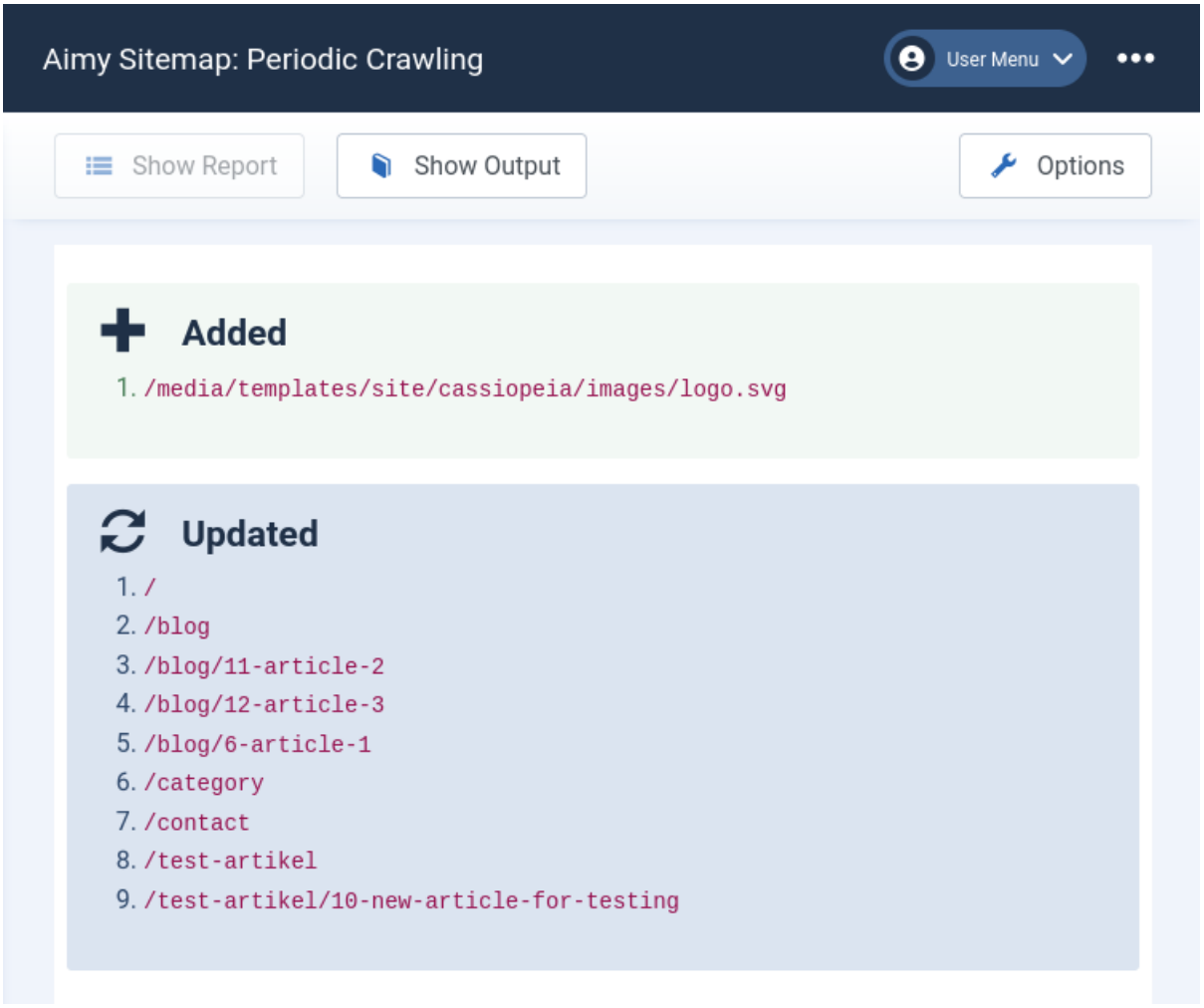
13.3 Crawling Information

Information about the current and - once it finishes - the last periodic crawl, including a detailed report of changes detected and the whole command line output, can be found within your website's Joomla! administration backend: "Components" → "Aimy Sitemap" → "Periodic Crawl".

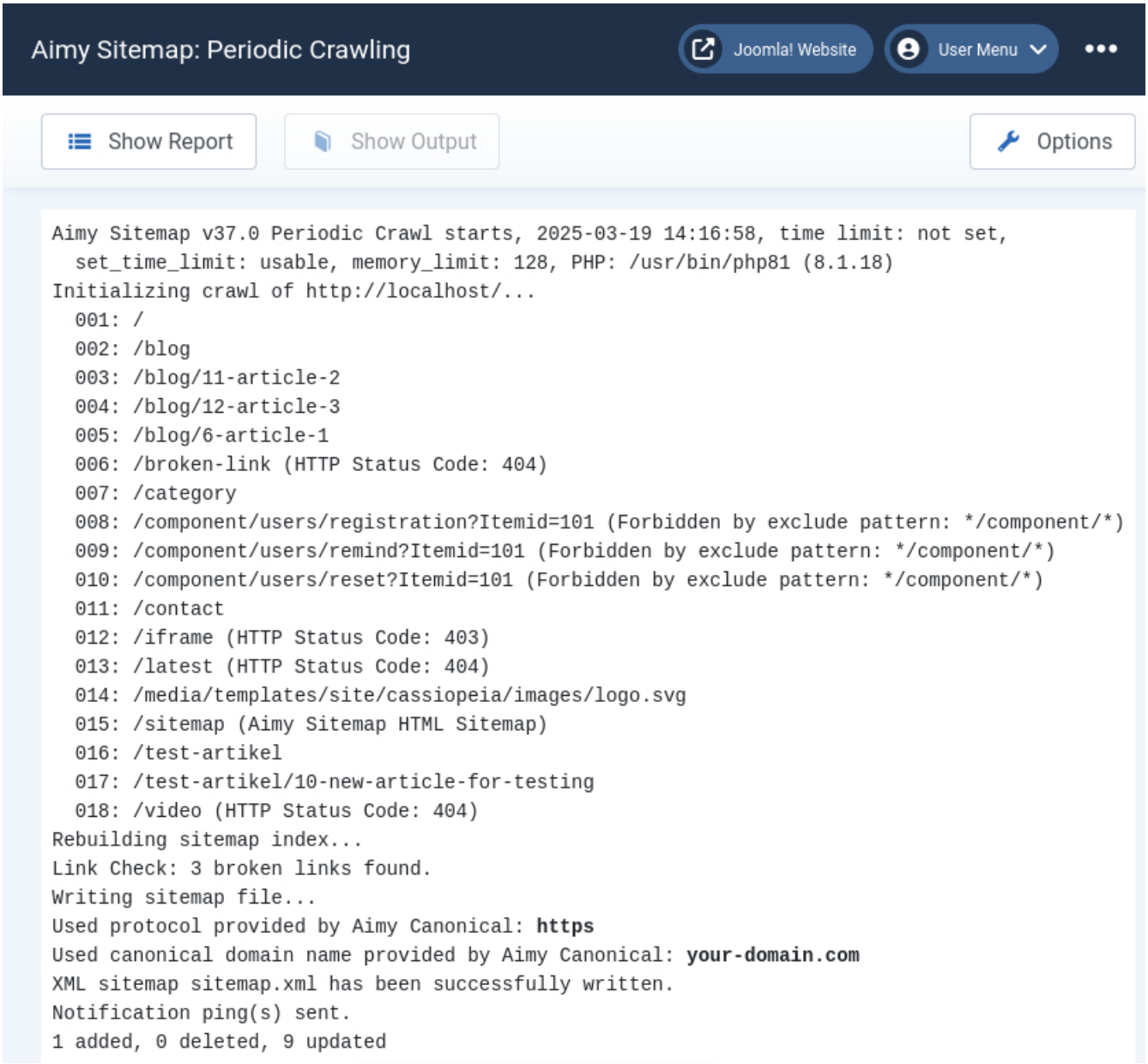
The default view provides a short report on the last crawl, along with all errors that may have occurred during the crawl and a short help:



Once a periodic crawl finishes, you can click on the "*Show Report*" button to view a detailed report on all changes done to your index:



To view the full output of the last or (if running) the current periodic crawl, click the "*Show Output*" button:



Note: If you passed "`--quiet`" as an argument to the "`crawl-website.php`" script, no non-error output is sent to the command line. For convenience, the output that would have been generated is stored and provided for review nevertheless.

While a **periodic crawl is running**, information on its progress is regularly updated along with its command line output. However, the "*Show Report*" button is deactivated until the crawl finishes. You can **reload the view** to refresh the information shown.

14 Special Cases

14.1 Crawling HTTPS-only Websites

Aimy Sitemap PRO fully supports crawling of HTTPS-only websites (websites that cannot be accessed unencrypted using HTTP), while the free of charge version of Aimy Sitemap limits the amount of entries that are written to the website's XML sitemap to 50 entries with this setup.

In case you already crawled and managed your URLs and do an upgrade to Aimy Sitemap PRO afterwards, all settings will be kept.

The only requirement for an HTTPS crawl is that your **PHP installation** supports SSL.

To find out whether your PHP installation supports SSL, log into your Joomla! backend and choose "*System*" → "*System Information*" and click the "*PHP Information*" tab. In the first table, look for "*Registered Stream Socket Transports*".

If the "**SSL**" transport is listed in this set of stream transports, Aimy Sitemap's crawler will be able to crawl your website using HTTPS.

14.2 Joomla! Installation in Subdirectory

If you have installed Joomla! to a **subdirectory** and want to use the `sitemap.xml` file Aimy Sitemap generates as a top-level sitemap nevertheless, you cannot directly set the path in Aimy Sitemap's options. That is, because for security reasons Aimy Sitemap will non write a sitemap to a file outside of Joomla!'s root directory.

However, there are a couple of ways to use the sitemap located within the Joomla! subdirectory as a top-level sitemap so that search engines will find and use it:

14.2.1 Rewrite Rule

You can set a rewrite rule for your webserver so that it "virtually" serves `/sitemap.xml` using the one created in your subdirectory. This approach is quite common, for example Joomla! itself uses rewrite rules for search engine friendly URLs.

The following rule is specific for *Apache*. If you use a different webserver for your site, have a look at the software's manual on how to write a similar rule.

Add the following lines to your **top-level** `.htaccess` file or create a new one if it does not already exists:

```
RewriteEngine on
RewriteCond %{REQUEST_FILENAME} !-f
RewriteRule ^sitemap\.xml$ /SUBDIR/sitemap.xml [L]
```

Just replace `SUBDIR` with the name of your subdirectory.

14.2.2 Sitemap Index File

Another way to put your subdirectory sitemap to use is to create a *Sitemap Index*. That's a simple XML file telling a search engine where to look for further sitemaps.

Here's an example for a top-level `sitemap.xml` file that references your real sitemap in your subdirectory:

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex xmlns="https://www.sitemaps.org/schemas/sitemap/0.9">
  <sitemap>
    <loc>https://YOURDOMAIN/SUBDIR/sitemap.xml</loc>
  </sitemap>
</sitemapindex>
```

Just replace YOURDOMAIN and SUBDIR with the appropriate values for your setup.

For more information have a look at `https://www.sitemaps.org/protocol.html#index`.

14.2.3 Robots.txt Directive

You may use a `robots.txt` directive to give search engines a hint on where to look for your sitemap as well. Please note that this directive is an extension of the standard robots.txt specification that may not be supported by all search engines.

Add the following line to your `robots.txt` file:

```
Sitemap: /SUBDIR/sitemap.xml
```

14.3 Using a Canonical URL (PRO Feature)

If your Joomla! website can be addressed using more than one domain name, one way to prevent search engines interpreting it as having *duplicate content* is to set a *canonical link tag* on your pages.

Our free plugin **Aimy Canonical** allows you to set a *unique domain name* and your preferred protocol (`http` or `https`) that is used to create the *canonical link tag* for all of your website's pages.

If Aimy Canonical is installed and enabled, Aimy Sitemap will use its settings and keep your sitemap's URLs in sync with it.

Feel free to download the Aimy Canonical Joomla! plugin on our website:

`https://www.aimy-extensions.com/joomla/canonical.html`

14.4 Using Aimy Sitemap on Large Websites

Although we recommend the usage of Aimy Sitemap on small to medium websites containing a couple of hundred pages only, it may be used on larger websites as well. However, here are a few hints on how to **optimize your setup**:

Use a professional hosting package

...that provides enough memory, handles a couple of simultaneous HTTP connections and allows you to set up cron jobs (preferably without time limit).

Enable periodic crawling

...and set it up to crawl your website in a reasonable interval, i.e. once per week. On very large websites (i.e. 100,000 pages) a crawling time of a couple of days is not uncommon, depending on the resources available to the host system.

Review your website's structure

...and optimize your website's **robots.txt** file accordingly: Add *Disallow* directives for all groupable resources that should not be added to your sitemap anyway. If in doubt, set the rules for Aimy Sitemap's crawler only:

```
User-agent: AimySitemapCrawler
Disallow: *.m4v
Disallow: *.gif
Disallow: /archive/
```

This will not only speed up the crawling process but will save you time reviewing and managing your website's index as well.

Enable server-side caching

...in Joomla!'s "Global Configuration" and set it up correctly for your use case (if possible).

Use a current version of PHP

PHP 7 comes with a huge performance benefit (compared to PHP 5). Prefer the most current stable version that is supported by all extensions used on your website.

Additional **notes** on things to keep in mind:

Multiple XML sitemaps & XML sitemap index

If the amount of URLs that should be added to your website's XML sitemap exceeds the limit of 50,000 URLs, Aimy Sitemap will split the XML sitemap into appropriate parts and generate a sitemap index file for you, which points to the splitted XML sitemaps.

Example: If you set an *XML Path* of "`sitemap.xml`" in Aimy Sitemap's options and the amount of active URLs is 80,000, Aimy Sitemap will generate "`sitemap-1.xml`" (containing 50,000 URLs) and "`sitemap-2.xml`" (30,000 URLs). The "`sitemap.xml`" file will contain the XML sitemap index.

15 Notes and Limitations

- The "Manage URLs" backend view may not represent the current **sitemap on disk**: keep in mind to click "*Write Sitemap*" to apply changes like updating frequency, priority or state.
- If you changed your **preferences for the crawler**, for example the inclusion of images, crawl again to update your index.
- Any change of the **default values** will --by definition-- apply for new entries only.

  If you like new default values to apply, click the "*Reset Index*" button in "Manage URLs" and re-crawl your website. Your new default values will then be applied to all found URLs.
- The **sitemap file** will be deleted when Aimy Sitemap is uninstalled as well.
- Only URLs of up to 767 bytes in length are indexed (URL encoded).
- Websites which are put into **offline mode** cannot be crawled as expected. That's because in offline mode there are no links to your content that Aimy Sitemap's crawler could extract and analyze.
- Documents which require **HTTP authentication** ("HTTP Auth") cannot be crawled. Including those in your sitemap wouldn't be useful, because search engines won't be able to index them as well.
- Command line output stored for **periodic crawls** is limited to 1,000 lines for performance and database compatibility reasons.
- The webserver that serves your website is required to support the **HTTP HEAD** request method - most webservers do so by default.
- A somewhat modern browser is required in order to use **browser notifications**. The following browsers (and versions) are known to provide support:
  - Firefox (version 22 or above)
  - Chrome (version 22 or above)
  - Safari (version 6 or above)
  - Opera (version 26 or above)
  - Edge (version 14 or above)

  Currently, no version of Internet Explorer supports browser notifications.
- The free of charge version of Aimy Sitemap limits the amount of entries written to an **XML sitemap** to 50 if the website is reachable using **HTTPS only**.
- If you are using **Microsoft's Internet Explorer** to use Aimy Sitemap in your website's Joomla! backend, make sure to use **version 9 or above**.
- HTTP Authentication support of Aimy Sitemap's crawler is limited to the **Basic HTTP Authentication Scheme** as described in RFC 7617.

16 Debugging and Errors

If you have problems crawling your Joomla! website, this may have a couple of reasons. If none of our hints in the backend or manual helps to solve the problem, you can turn on **debugging mode** to get some more details.

Debugging mode can be activated by either...

1. enabling the *Enable Debugging Mode* option in Aimy Sitemap's options or
2. by setting *Debug System* to "Yes" in Joomla!'s configuration (*System* → *Global Configuration* → Tab *System*).

If you start a crawl afterwards it will be logged in a file named "`aimysitemap.php`" in the **Log Folder** set in Joomla!'s *Global Configuration* (*System* → *Path to Log Folder*). Go through the file to find hints on the issue or send the file to Aimy Extensions' support team for review and help.

Note: If debugging mode is enabled, the index of your website will not be updated.

Additionally there may be common problems depending on your setup. In the FAQ on our website you find some hints:

`https://www.aimy-extensions.com/joomla/sitemap.html#tab-faq`

17 Copyright & Trademark Notice

The Joomla!® name and logo are trademarks of Open Source Matters, Inc. in the United States and other countries.

Mentioned hard- and software as well as companies may be trademarks of their respective owners. Use of a term in this manual should not be regarded as affecting the validity of any trademark or service mark. A missing annotation of the trademark may not lead to the assumption that no trademark is claimed and may thus be used freely.