Octolooks Scrapes Guide

Automatic WordPress Scraper and Content Crawler Plugin

Version 2.1.0

Table of Contents

Introduction

Thank you for purchasing Octolooks Scrapes. We have prepared this detailed documentation for your comfy usage however please do not hesitate to contact us for any issues that you may encounter.

How It Works

Octolooks Scrapes is a WordPress plugin which scrapes the content that you locate from the source you want and copy to your existing WordPress website. It consists of three task types: **“Single”** to scrape located areas from only one web page, **“Serial”** to scrape listing pages which have links to detail pages and **“Feed”** to scrape in any format feed links.

In order to create a scrape task it is enough to choose a task type, enter the source url, locate the content area you want and select schedule options basically. Newly created scrape task will run automatically in the frequency you have chosen.

In this documentation **“task”** represents the scrape created with the features you have chosen, **“run”** represents the actual code running of a task which has due time and **“process”** represents instant processes in task.

Requirements

Octolooks Scrapes is designed to give maximum performance even on a minimum system configuration however the hardware configuration of the server, connection speed and physical location are significant factors which are affecting the performance. The minimum requirements are listed below:

- PHP 5.2.4+
- dom, mbstring, iconv, json and simplexml extensions for PHP
- Configurable maximum execution time for PHP
- WordPress 3.5+
- IE Edge, Firefox, Safari, Chrome for admin panel

Installation

In order to install plugin to your server log in to your WordPress administration panel and click on **Plugins** » **Add New** from left navigation. In the opening screen click on **Upload Plugin** and select ol_scrapes.zip file from your computer, click on **Install Now** and wait for uploading. When it finishes click on **Activate Plugin** under Octolooks Scrapes in plugin listing page. You will see Scrapes tab on left navigation with Octolooks logo when activation process completed.

Validating your purchase code

In order to validate your purchase code log in to your WordPress administration panel and click on **Scrapes** » **Settings** from left navigation. If you have not validated your purchase code yet, you will be automatically redirected to this page when you click on **Scrapes** » **Add New**.

Your purchase code will be defined for a single domain name at first validation according to regular license terms. It will be valid only for first domain name that you entered including subdomain names and localhost for testing purposes.

For example, if you registered your purchase code for http://octolooks.com domain name, it will be valid for http://localhost, http://127.0.0.1, http://octolooks.com, http://www.octolooks.com, http://sub.octolooks.com (or any other subdomains) and it will be invalid for any other domains.

If you didn't decide your domain name yet, you can register it with http://localhost domain name to run the plugin on localhost for testing purposes and contact us for change after you decide your domain name. Also if you would like to change your domain name later, you can contact us anytime for change.

Regular license terms

Please visit url below for detailed information.
https://octolooks.com/terms/

Deactivation and deleting

In order to deactivate the plugin log in to your WordPress administration panel and click on **Plugins** » **Installed Plugins** from left navigation. Afterwards click on **Deactivate** link under Octolooks Scrapes plugin. You will see Scrapes tab with Octolooks logo is disappeared when the deactivation process completed. You can also click on **Delete** link and completely remove the plugin from your server.

Updating

When a new update is available you will receive an e-mail linked to your account with title **"Update available for Scrapes"**. If you don't have any currently running Scrapes tasks, deleting the old version and installing the newer version of the plugin is enough. But if you already have Scrapes tasks and want to preserve them, you should follow these steps below for a smooth update process.

1. Backup your existing tasks from **Tools** » **Export** menu, select **Scrapes** option and click on **Download Export File** button.
2. Deactivate **Octolooks Scrapes** from **Plugins** menu, and then delete.
3. From **Plugins** » **Add New** menu, click on **Upload Plugin** button and upload the new zip file. Then complete the installation by clicking on **Install Now** button.
4. From **Tools** » **Import** menu, click on **WordPress** » **Run Importer** link. If you don't have WordPress importer tool, you can install it from the same menu by clicking on **Install Now** link.
5. Click on **Choose File** button and select your backup xml file then complete the process by clicking on **Upload file and import** button.
6. Your old tasks will be on **Scrapes** » **All Scrapes** menu with the status **waiting**. In order to run them click on **Edit** button, check for the new fields that may appear for newer version and choose your options then finally click on **Edit** button to save changes.

Add New Task

The steps you will follow differ according to task type you choose from Single, Serial or Feed task types. All common settings for all task types are explained in detail below.

Main Options

Name

(Required) The field used to give a name and define task. It is required to give a name in order to show, edit or delete this task afterwards. You can change the name which is defined automatically.

Task type

(Required) The field used to set running type of the task. Remaining form fields are updated depending on the value of this field. The options are as follows.

Single: The option used to scrape a particular web page from a website.
Serial: The option used to scrape a particular listing pages from a website which have links to detail pages.
Feed: The option used to scrape Atom or RSS feed format web page.

Request Options

Cookies

The field used to set which cookie values to be sent to source url requests. When you click on Add new cookie link Name and Value fields are added to form. Remove button deletes cookie fields.

Source URL

(Required) The field used to set which source you want to scrape. For single task type it is the web address itself, for serial task type it is the listing page address, for feed task type it is the feed link which is entered to source url field.

Post Item

(Required, for serial task type) The field used to set which links redirect to detail pages. The XPATH information will be automatically entered after you first click on Select button and locate the link from the inspector screen and click on that link. Locating only one post item link is enough to set plugin scrape path. Select **“Exact match only”** to ignore links at the same level and only match the exact path entered.

Next Page

(For serial task type) The field used to set which link redirects to next list page. After you click on Select button and locate the next page link on inspector screen, the XPATH information of that item will be automatically entered.

Next page link will mostly be at the bottom of the listing page. In some websites when you choose the post item the next page field is added automatically, in this case you don’t have to relocate that link again.

Enter URL parameter: Page parameters starts with the value and increments, for example http://example/products?page=1 is the target website, if you set **name**: page, **value**:1, **increment**:1 to fields it will check http://example/products?page=**1** http://example/products?page=**2** http://example/products?page=**3** . You need to find the page parameter in the URL first, generally set value to 1 and increment to 1, so it will go like 1,2,3,4,5 Sometimes it may appear as item count like http://example/products?products=**20** http://example/products?products=**40** http://example/products?products=**60** then you need to set value to 20 and increment to 20, so it will go like 20,40,60.

Taxonomy Options

Post type

The field used to set the post type for posts which will be created automatically by the plugin in your WordPress site. All post types defined in your system by your theme and plugins are listed here. Categories and create categories fields are updated or hidden depending on your choice on this post type field.

Create categories

The field used to set the categories of automatically created posts to which newly created categories by the plugin in your WordPress site. In order to scrape the categories from source url choose a taxonomy, click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field.

If this area has multiple category values by setting a separator value you can make plugin to create multiple categories at once for the posts. This XPATH information and separator value will be enough for the plugin to detect other posts' category fields.

Enable find and replace rules: The option to enable find and replace rules for scraped value. When you click on Add new find and replace rule link, Find and Replace fields are added to form. It supports multiple rules and regular expressions (regex). Remove button deletes the fields.

Categories

The field used to set the categories of posts which will be created automatically by the plugin in your WordPress site. When this field left empty WordPress automatically assign your posts' categories to uncategorized.

Post Options

Title

The field used to set post titles which will be created automatically by the plugin in your WordPress site. Click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field. This XPATH information will be enough for the plugin to detect other posts' title automatically. The options are as follows.

Enable template: The option to set a template for value or not. When you choose this option, template field is added to form and you can combine custom text and [scrape_value], [scrape_date], [scrape_url] or [scrape_meta name="name"] shortcodes. By setting one template it will be enough for plugin to determine the template for other values in scraping process.

Enable find and replace rules: The option to enable find and replace rules for scraped value. When you click on Add new find and replace rule link, Find and Replace fields are added to form. It supports multiple rules and regular expressions (regex). Remove button deletes the fields.

Content

The field used to set post content which will be created automatically by the plugin in your WordPress site. The options are as follows.

Enable template: The option to set a template for value or not. When you choose this option, template field is added to form and you can combine custom text and [scrape_title], [scrape_content], [scrape_thumbnail], [scrape_gallery], [scrape_categories], [scrape_tags], [scrape_date], [embed][scrape_url][/embed], [scrape_meta name="name"] or [scrape_url] shortcodes. By setting one template it will be enough for plugin to determine the template for other values in scraping process.

Enable find and replace rules: The option to enable find and replace rules for scraped value. When you click on Add new find and replace rule link, Find and Replace fields are added to form. It supports multiple rules and regular expressions (regex). Remove button deletes the fields.

Detect automatically: The option to set automatically created posts' content from source url by using a special algorithm to locate the content.

Select from source: The option to set automatically created posts' content from the source url. When you choose this option, click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field. This XPATH information will be enough for the plugin to detect other posts' content automatically.

Allow HTML tags: The option to set whether it is allowed to use html tags in the content or not. When you choose this option the html tags in the content will not be cleaned.

Download images to media library: The option to set content image files will be retrieved from the remote source or will be downloaded first and retrieved from the local server. When you choose this option the images will be downloaded to media library and uses disk space from your server.

Excerpt

The field used to set post excerpt which will be created automatically by the plugin in your WordPress site. If this field is left empty WordPress will automatically create them. The options are as follows.

Enable template: The option to set a template for value or not. When you choose this option, template field is added to form and you can combine custom text and [scrape_value], [scrape_date], [scrape_url] or [scrape_meta name=“name"] shortcodes. By setting one template it will be enough for plugin to determine the template for other values in scraping process. This option is only available when Select from source option is selected.

Enable find and replace rules: The option to enable find and replace rules for scraped value. When you click on Add new find and replace rule link, Find and Replace fields are added to form. It supports multiple rules and regular expressions (regex). Remove button deletes the fields. This option is only available when Select from source option is selected.

Generate from content: The option to set automatically created posts’ excerpt from the post content.

Select from source: The option to set automatically created posts’ excerpt from the source url. When you choose this option, click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field. This XPATH information will be enough for the plugin to detect other posts’ excerpt automatically.

Tags

The field used to set the tags of posts which will be created automatically by the plugin in your WordPress site. The options are as follows.

Enable find and replace rules: The option to enable find and replace rules for scraped value. When you click on Add new find and replace rule link, Find and Replace fields are added to form. It supports multiple rules and regular expressions (regex). Remove button deletes the fields. This option is only available when Select from source option is selected.

Select from source: The field used to scrape the tags of automatically created posts from source url. When you choose this option click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field. If this area has multiple tag values by setting a separator value you can make plugin to create multiple tags at once for the posts. This XPATH information and separator value will be enough for the plugin to detect other posts' tags fields.

Enter custom: The option to set the post tags by manually entering tags with comma (,) separated value. When you choose this option and enter the value, all automatically created posts will have these (same) tags.

Featured image

The field used to set featured image for posts which will be created automatically by the plugin in your WordPress site. The options are as follows.

Detect from feed: (For feed task type) The option to set automatically created posts' featured image from the source url by using a special algorithm to detect.

Select from source: The option to set automatically created posts' featured image from the source url. When you choose this option, click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field. This XPATH information will be enough for the plugin to detect other posts' featured image automatically.

Select from media library: The option to set automatically created posts' featured images from the media library manually. When you choose this option and select an image from the media library, media id of the image will be automatically set and all posts created by this task will have this featured image (same) value.

Custom fields

The field used to set the custom meta fields of posts which will be created automatically by the plugin in your WordPress site.

When you click on Add new custom field link Name, Value and Attribute fields are added to form. Clicking on name field shows all available custom field names generated by other plugins and themes installed, you can select one or enter another. Remove button will delete the custom field.

Select button will open the inspector screen and clicking on related area for meta value will close the screen automatically and XPATH information will be entered automatically.

Attribute is the field to set the attribute name which used to retrieve attribute value from your chosen XPATH element, when leave empty default node value is set (e.g., **href** for hyperlinks, **src** for images).

Enable template: The option to set a template for value or not. When you choose this option, template field is added to form and you can combine custom text and [scrape_value], [scrape_date] or [scrape_url] shortcodes. Also you can use calc() shortcode for mathematical calculations. By setting one template it will be enough for plugin to determine the template for other values in scraping process. This option is only available when Select from source option is selected.

Enable find and replace rules: The option to enable find and replace rules for scraped value. When you click on Add new find and replace rule link, Find and Replace fields are added to form. It supports multiple rules and regular expressions (regex). Remove button deletes the fields. This option is only available when Select from source option is selected.

Allow HTML tags: The option to set whether it is allowed to use html tags in the content or not. When you choose this option the html tags in the content will not be cleaned.

Custom fields for WooCommerce

In order to create a WooCommerce product, required steps and main custom fields are listed below. *

1. Select **Product** as a Post type (Required).
2. Select **Product Categories** as a taxonomy (Optional, only if you need to create categories).
3. Select **simple** or **external** as product type (Required).
4. Set required **custom fields** listed below.

Name	Template	Description
_price	-	Required
_regular_price	-	Required
_sale_price	-	Discounted price, optional
_height	-	
_length	-	
_weight	-	
_product_image_gallery	-	Select image container which contains main image and thumbnails as a XPATH value.
_manage_stock	yes, no	
_stock	-	Required if manage stock is set yes
_stock_status	instock, outofstock	Required if manage stock is set no. Use **outstock** instead of **outofstock** for old versions.
_product_url	[scrape_url]	URL for external products
_button_text	text	Label for external product button (e.g., Visit)
_backorders	yes, no, notify	
total_sales	-	
_virtual	yes, no	
_visibility	visible, catalog, search, hidden	
_featured	yes, no	
_purchase_note	-	
_sku	-	

* Grouped, variable and downloadable product types and product attributes are not supported currently.

Customization Options

Translate fields

The field used to set the translation language for all fields of automatically created posts by the plugin in your WordPress site. Don’t forget, enabling the translation option may increase the total scraping time for each post.

Service: The option to set the translation service that translates the fields. Available services are Bing Microsoft Translator, DeepL Translator, Google Translate, Google Translate (Unofficial) and Yandex Translate.
API keys: The option to set the API keys that you get from the translation service provider. You can enter multiple API keys as a new line by pressing Enter key to use a different API key for each request.
Source: The option to set the language of source site that you are scraping the content.
Target: The option to set the target language that you want to translate the scraped content.

Spin fields

The field used to set a content spinning service that spins the automatically created posts to get a unique content. Spinner feature requires The Best Spinner account which you can create new one from http://thebestspinner.com. After that you can fill the **E-mail** and **Password** fields. You can also use other 3rd party spinner plugins for WordPress.

Publish Options

Author

The field used to set the publishing author for automatically created posts by the plugin in your WordPress site.

Status

The field used to set which post status for automatically created posts by the plugin in your WordPress site.

Date

The field used to set the publish date of posts which will be created automatically by the plugin in your WordPress site. The options are as follows.

Detect from feed (For feed task type): The option to set automatically created posts' publish date from source url by using an algorithm.

Process time: The option to set automatically created posts' publish date to post creation time.

Select from source: The option to set automatically created posts' publish date from the source url. When you choose this option, click on Select button and locate the related area from the inspector screen and click on it, the inspector screen will close automatically and XPATH information will be set to this field. This XPATH information will be enough for the plugin to detect other posts' publish date automatically.

Enter custom: The option to set the post publish date by manually entering value. When you choose this option and enter the value, all automatically created posts will have this (same) date.

Discussion

The field used to set whether user comments are allowed for automatically created posts by the plugin in your WordPress site or not.

Filter Options

Unique post check

The field to use whether posts should be unique or not. If you do not choose any options from this field plugin will continue to add new posts even the same post already exists in the WordPress site. On existing post and Complete run on existing fields are updated or hidden according to your choice. The options are as follows.

From title: The option to set unique post check will be made by post title. When you choose this option if the post in source url with the same title already exists in your WordPress site the status will be "on existing post" and plugin makes the action you choose.

From content: The option to set unique post check will be made by post content. When you choose this option if the post in source url with the same content already exists in your WordPress site the status will be "on existing post" and plugin makes the action you choose.

From source url: The option to set unique post check will be made by source url. When you choose this option if the source url is already scraped before, status will be "on existing post" and plugin makes the action you choose.

On existing post

The field to set the action when a post in source url already exists in the WordPress site. The options are as follows.

Complete process (For single task type): The option in the case that post already exists in the WordPress site to stop the process until next due time.

Skip to next process (For serial and feed post types): The option in the case that post already exists in the WordPress site to do nothing and pass the next process.

Update post: The option in the case that post already exists in the WordPress site to update the current post with the source url and pass the next process.

Complete run on loop

(For serial and feed task types) The field to set how many “on existing post” occurrence is needed to stop the task until next run time in order to save system resources.

Filters

The field to set filters for posts before they are being created. They are checking the scraped value, with operator to some other arbitrary value you enter. For example;

price , greater than , 5 . When you set this, the products which their price greater than 5 will be filtered, not inserted as a post or product to your website.

Product 1, price 10, it is going to be filtered -
Product 2, price 3, it is not going to be filtered +

Another example for contains operator; **title , contains, "some word”** . This rule will filter titles which contains the "some word" and they will not be entered to your website.

This is a blog title with some word , this going to be filtered -
This is another blog title , this one is not going to be filtered +

For technical information, the filtering mechanism is like PHP if condition mechanism, when condition holds (returns true) it filters out and skips next post if (scrape value, operator, arbitrary value you enter manually) is true then filter out, else update or insert the post according to your uniqueness section preference.

Schedule Options

Cron type

The field used to set the method of calling the task when it is due time in the plugin in your WordPress site. The options are as follows.

System: The option to add a system command to trigger WordPress cron scheduling system as a unix cron job. This option does not require any additional trigger and task runs at time in the background. If your server is not eligible for adding system crons, it will automatically be converted to WordPress cron type.

WordPress: The option to call run process as a WordPress scheduled job. The task is triggered when your WordPress website is visited by someone and runs in the background.

Total Posts

(Required, for serial and feed task types) The field used to set how many posts will be created in the task in each run. When the task reaches this amount of post it stops the execution until next run. The options are as follows.

Unlimited: The option to set the total number of posts to unlimited value. When you choose this option total posts field will be hidden.

Total Runs

(Required) The field to set how many times the task will run. When the task reaches the total run value it stops running. The options are as follows.

Unlimited: The option to set the total number of runs to unlimited. When you choose this option total runs field will be hidden.

First run

The field the set first running time of task.

Run frequency

The field to set the time interval of each task run.

On uncompleted run

(For serial and feed task types) The field to use what action will be taken when a process is not finished and according to run frequency field another task should start. In order not to encounter a situation like this and keep tasks processing in their normal time please give realistic run frequency values. The options are as follows.

Terminate previous run: The option to stop the previous running and uncompleted task when next task should start according to chosen run frequency field.

Wait until previous run is completed: The option to wait for the previous running and uncompleted task when next task should start according to chosen run frequency field.

Run type

(For serial task type) The field to set where the task should start from for every run.

Other Options

Timeout for processes

The field to set maximum time the task will wait for a reply for http requests. If a response does not come from source url the process will be at “on error” status.

Wait next processes

(For serial task type) The field to set how much time to wait between processes. If this value is too short plugin will make a lot of http requests to source url server(s) in a short time period and may cause a temporary or permanent IP ban to your server.

On error

(For serial task type) The field to set what action will be taken if task encounters an error during scrape process.

Skip to next process: The option to set the task skips to next process without finishing the current process in error state.
Complete run: The option to set the task finishes the current job completely and wait until next process due time in error state.

Create / Edit

The button to save (Create) or update (Edit) the options. If any of the options is invalid or missing, button will be passive and not clickable. When all options are valid you can click on this button to save the changes.

First run will start immediately after you click on Create button. Also editing will abort the task if running then restart it immediately. When save operation is completed you will be redirected to "All Scrapes" screen.

View all tasks

You can reach your tasks in the left navigation by clicking Scrapes tab menu after you log in to your WordPress administrator panel.

Columns

Name

The field that shows the name of the task and its id at creation time. Same name can be used in multiple tasks but id is unique for each task and is created automatically.

Status

Preparing: Shows that the task is preparing for starting or stopping.
Running: Shows that the task is already processing and running
Complete: Shows that the task reaches the maximum run count and won't start again.

Waiting next run: Shows that the task will start automatically when it is due time.

Deactivated: Shows that the task is in trash, it is only visible for trash status tasks.

Schedules

Last run: Shows that the last time when it is started to process

Last complete: Shows that the last time when it is finished to process.

Last scrape: Shows that the last time when it is scraped a post.

Next run: Shows the next time when it is going to process again.

Total run: Shows that the number of successfully completed processes and remaining number of task runs if any.

Actions

Run: The button which is used to start the task manually. It is visible when in complete or waiting next status. In waiting next run status starts immediately without waiting the next due time.

Pause: The button which is used to stop the task manually. It is visible when in running status. Stops the task until its next run time.

Edit: Redirects to task edit page. When you edit a task if there is any running task currently it stops and resets the task and activated with the new settings.

Copy: Helps to create a copy of the current with the same name, settings with a different ID. The copied task is in preparing status and activates when you edit and save the task.

Trash: Moves task to trash, stops the currently running task if any, and prevents future runs.

Restore: Moves your task to task lists back, when you restore a task it is in preparing status and editing activates it again.

Support and contact

Since we only give product support to purchased users, please contact us via **Contact Form** at https://octolooks.com/contact/ . Once you sent an email to us, we are going to do our best to reply your questions as soon as possible.