What are Product Crawlers?

Product crawler is used to populate last level of category with products.
It is separately configured for each website. It means that it will crawl only a single website for which it is configured in order to fetch products, images and its description.
You can add multiple configurations to configure separate websites at a time. Each configuration is good for a single website.

You will have to enter set of configuration parameters properly for successful crawling. In this tutorial we will learn how to successfully configure and crawl for a website.

Suppose you want to fetch products from www.bizrate.com then you will set parameters of the crawler based on the html tags of this website (you need to know basic html to set parameters). These settings will working for only this website and will fetch products from this website. Once you are done with this website, you can reconfigure for any other website. Usually one or two websites are enough to extract products.

* Now question is how each product will be inserted into a particular category?
While setting parameters for the crawler you will also specify under which category it should insert extracted products.

* Another question arises that the pages crawled will have products other than we want to insert, how can we prevent them from getting inserted into our database?
Again at the time of specifing parameters for crawler you will also specify set of keywords and ignore keywords. Auto crawler reads the chunk of webpage and matches the information extracted with the keywords and ignore keywords. In order to add product extracted into a category it must match the keywords and must not match ignore keywords. If match is complete, product name, its discription is saved under that category (and image is copied as well -*optional*-).

Example: -
-> Computers
---> CPU
------> AMD (Keywords: AMD. Ignore keywords: Windows XP)
---------> Product added can be **AMD Athlon 2800+** or **AMD Athlon XP**

* Another question may come to mind is why dont we talk about importing prices when configuring product crawler?

There are two separate crawlers, one is product crawler and another is price crawler. Product crawler only crawl for products, their images and description and add only those information to the website database. However price crawler crawl merchant website for price information for products.

Both are separate so that both can be configured to run at different intervals. Further it helps in splitting task thus splitting server load.

Product crawler (unlike price crawler) does not have to run at regular intervals. Once you have extracted and inserted all products in the database (which you wanted to insert) its task finishes. On the other hand, price crawler should run regularly to keep prices updated.

You can setup both crawlers as separate cronjobs using the ssh calling scripts.

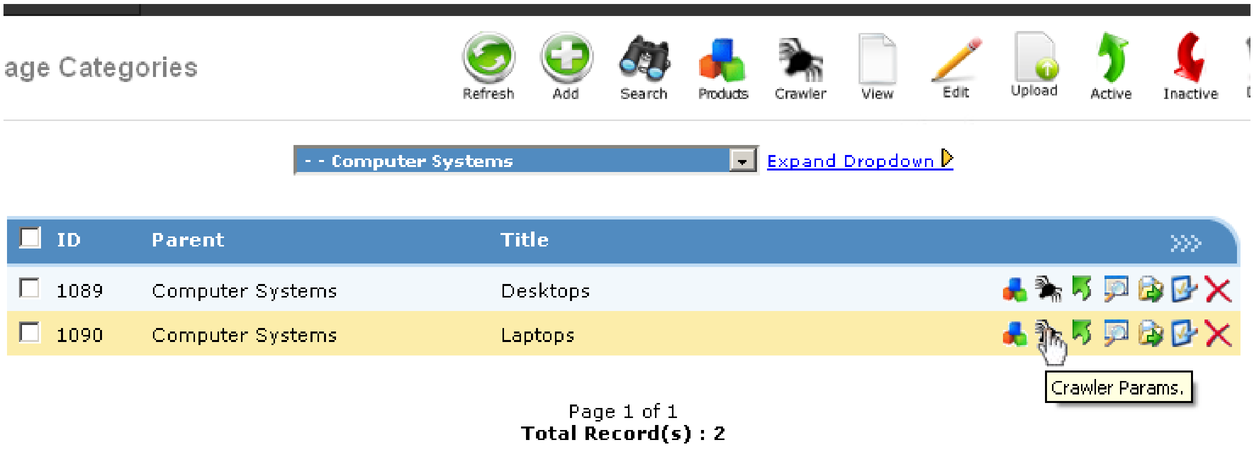
Use the product ssh file to call products crawler once in a while, like once in a week:
php /path/to/your/script/admin/cronjob_product_ssh.php

Similarly you can call the price ssh file to call prices crawler as often as you think the prices would change, like may be once every day:
php /path/to/your/script/admin/cronjob_price_ssh.php

Note: Crawling is resource hungry task and requires time, cpu and other resources of the server. So you must setup these above cronjobs wisely and carefully.

Setting up your Product Crawlers?

Suppose we have following set of main, sub and micro categories. **Computers > Computer Systems > Laptops**
Click on **Content > Categories > Manage Categories** (Alternatively you can also search a category).
In the category listing click on the Product Crawler icon (See screenshot below)



- - Computer Systems

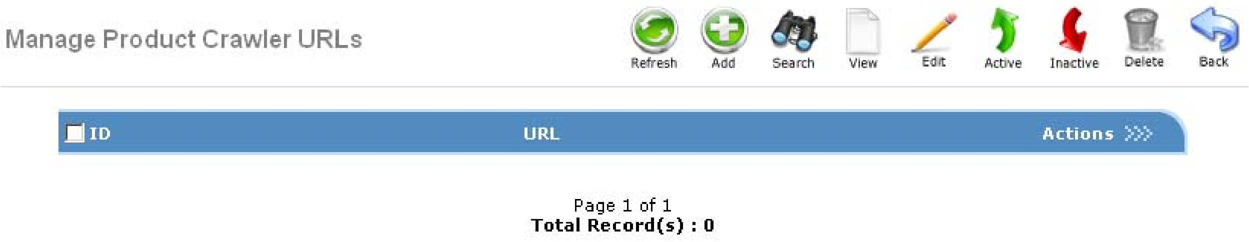
Expand Dropdown

□ ID	Parent	Title
□ 1089	Computer Systems	Desktops
□ 1090	Computer Systems	Laptops

Page 1 of 1
Total Record(s) : 2

The new page will show listing of the urls to be spidered for the products under that

category (see image below).
Click Add icon (or update current url) to insert configuration for the product crawler for that category.



Manage Product Crawler URLs

□ ID	URL	Actions ⋙

Page 1 of 1
Total Record(s) : 0

The new page which will have number of parameters for you to customize.

Load / Save Settings:
You can save the crawler settings with a name using this feature so that to create new similar crawlers, you just need to load a setting and then edit it to create a new one.

Product Classification:
1.Category:
This is the category in which the products will be added.

2.Keywords:
These are the words that should be present (all keywords should be present) within the chunk to add it to that category. In our example, it is laptop

3.Ignore Keywords
These are the keywords if found in the chunk (if anyone is found), then that chunk should be ignored. In our example we leave it blank.

4. URL
It is the url from where the crawler will start crawling for the products of this category. We put http://www.bizrate.com/laptopcomputers/index__start--1.html inorder to spider bizrate for laptops.

5. Spider URLs containing
It should contain a phrase which should be present in the subsequent urls to be spidered. Meaning, when you will go to the above url, you will notice there are links to the other categories and products on that page as well as links to about us and privacy policy. You dont want your crawler to waste time in unnecessary pages so you instruct it to spider only those urls which contains that particular phrase.

We put laptopcomputers/index__start as it is present in all the listing pages for laptops and compareprices.html as it is present in all laptop product details page (this second entry will also be used as Process URLs in step 3).

This will ensure that crawler will only go into pages where product info will be present. Therefore saving time and resources. If left blank it will go into all links like about us and others.

6. Process URLs containing

It should contain a phrase which is present in the urls of the product detail page. So that only those pages are processed for product extraction. Example can be, product_detail or product_info

In our case, it is compareprices.html as this is present in all the product detail pages on that URL.

Note: Process URLs should be the sub-set of the Spider URLs. It must contain one (or more) entries from Spider URLs only.
In other words, a page spidered for URLs can only be processed for product extraction BUT it is not neccessary that all pages spidered should be processed.

7. Chunk Start

A chunk is the block of html tags/codes in which a product with its complete information is present. If in a website, products are listed row-wise then chunk will be within <tr> and </tr>. Chunk Start is the start of html tags in which single product's title and detail is enclosed in.

In our case, we can pick <table cellpadding="0" cellspacing="0" border="0"> as start. You need to know about HTML very well to put these settings.
We chose above HTML tag because it is unique single line HTML tag that occurs before the occurance of product name and images.

8. Chunk End

It is the end of html tags in which single product's title and detail is enclosed in.

In our case, we can pick <div style="clear:both;padding:0 0 8px 0;"></div> as end.
We chose above HTML tag for the reason that it is unique single line HTML tag which only occurs after the product name, detail and description.

9. Title Start

It is the start of html tags in which product title is enclosed in.

In our case, we can pick <h1> as start.

10. Title End

It is the end of html tags in which product title is enclosed in.

In our case, we can pick </h1> as end.

10. Manufacturer Start

It is the start of html tags in which product manufacturer is enclosed in.

11. Manufacturer End

It is the end of html tags in which product manufacturer is enclosed in.

12. Manufacturer Code Start

It is the start of html tags in which product manufacturer code is enclosed in.

13. Manufacturer Code End

It is the end of html tags in which product manufacturer code is enclosed in.

14. Details Start

It is the start of html tags in which product details are enclosed in.

In our case, we can pick <a href="javascript:showDetails(false)" >Hide Details as start.

15. Details End

It is the end of html tags in which product details are enclosed in.

In our case, we can pick </div> as end.

16. Image URL

It is the location where image is stored on the remote server. Remember each image on the server is differentiated by the product id or some other unique text, enter [IMAGEID] in place of that. Your image url will look like
http://image.bizrate.com/resize?sq=400&uid=[IMAGEID] or
http://images.bizrate.com/462/p_[IMAGEID].jpg.

In our case it is http://image.bizrate.com/resize?sq=400&uid=[IMAGEID]

Now you have to also tell the crawler how to extract the **[IMAGEID]** and replace it in the url you have just provided in order to copy image from remote server.

You can also find this information in the HTML source of the same page. Locate the unique image id on the page's html source. Provide its prefix and suffix as the parameter so that [IMAGEID] can be easily extracted.

In our case we will enter --pid as prefix and / as suffix. The reason we will enter these tags as prefix and suffix is Product ID is surrounded by these two tags. (Example : **--pid**337038385**/"**)

Can/Cannot add new products

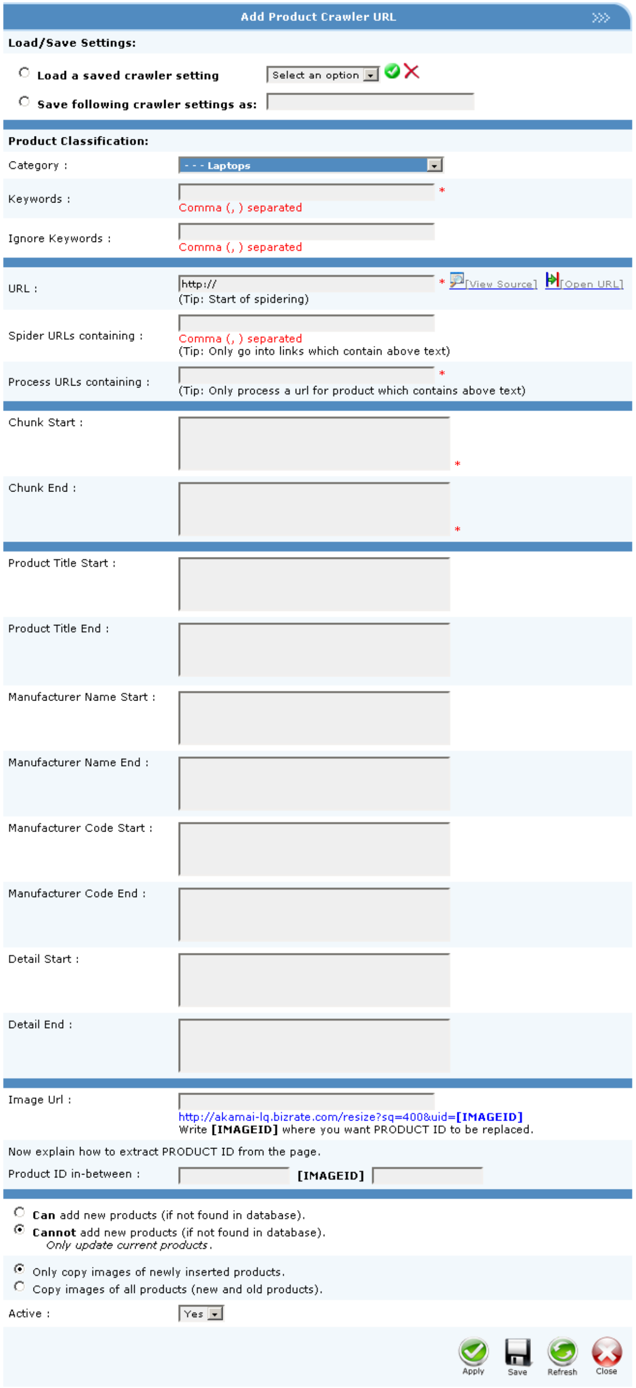
This option is to tell the script that if the product is not found in our database then ignore it or add a new product.

Copy images of newly inserted products only

Using this option, you can tell the script that whether you want to copy images for all products (new and old) or only for newly inserted products.

Once you have saved these settings. Click on Options in left menu and Run Product Spider. If you have entered everything correctly. You should have all the products with title and description in your database.

You can view log of the actions perform once task is finished by the spider.



Add Product Crawler URL

Load/Save Settings:

○ **Load a saved crawler setting** Select an option

○ **Save following crawler settings as:**

Product Classification:

Category : - - - Laptops

Keywords : *
Comma (,) separated

Ignore Keywords :
Comma (,) separated

URL : http:// * [View Source] [Open URL]
(Tip: Start of spidering)

Spider URLs containing :
Comma (,) separated
(Tip: Only go into links which contain above text)

Process URLs containing : *
(Tip: Only process a url for product which contains above text)

Chunk Start : *

Chunk End : *

Product Title Start :

Product Title End :

Manufacturer Name Start :

Manufacturer Name End :

Manufacturer Code Start :

Manufacturer Code End :

Detail Start :

Detail End :

Image Url :
http://akamai-lq.bizrate.com/resize?sq=400&uid=**[IMAGEID]**
Write **[IMAGEID]** where you want PRODUCT ID to be replaced.

Now explain how to extract PRODUCT ID from the page.

Product ID in-between : **[IMAGEID]**

○ **Can** add new products (if not found in database).
● **Cannot** add new products (if not found in database).
Only update current products.

● Only copy images of newly inserted products.
○ Copy images of all products (new and old products).

Active : Yes

This configuration completely depends upon the HTML format of the bizrate website. If bizrate will change its HTML format this tutorial will not work. Parameters will have to be updated again to make it work.

Another Example HTML parameters:

As a second example we will crawl mobile phones category of www.e7.ro. The url is http://www.e7.ro/telefoane-gsm-price-up-c-65-p-1.html

1. URL

For this example we put http://www.e7.ro/telefoane-gsm-price-up-c-65-p-1.html inorder to spider for mobile phones.

2. Spider URLs

We put -c-65-p- as it will be present in all mobile phone listing pages and we put c-65-p-1-pr- as it is present in all products` details page for mobile phones.

This will ensure that crawler will only go into pages where product info will be present. Therefore saving time and resources. If left blank it will go into all links like about us and others categories.

3. Process URLs

It should contain a phrase which is present in the urls of the product detail page. So that only those pages are processed for product extraction.

In this case, it is c-65-p-1-pr- as this is present in all the product detail pages on that URL.

4. Keywords

These are the words that should be present (all keywords should be present) within the chunk to add it to that category. In our example, it is mobil

5. Ignore Keywords

These are the keywords if found in the chunk (if anyone is found), then that chunk should be ignored. In our example we leave it blank.

6. Chunk Start

A chunk is the block of html tags/codes in which a product with its complete information is present.

In this case, we can pick </table><!-- border --> as start.
We chose it because is a unique single line HTML tag that occur before product name and image. So we choose this tag.

7. Chunk End

In this case, we can pick <span id="product_alt_price" style="white-space:

nowrap;"></span></font> as end.

8. Title Start
In this case, we can pick <td class="DialogTitle"> as start.

9. Title End
In this case, we can pick </td> as end.
Product Title is enclosed in TITLE START and TITLE END.
10. Details Start
In this case, we can pick <table width="100%" cellspacing="0" cellpadding="0"> as start.

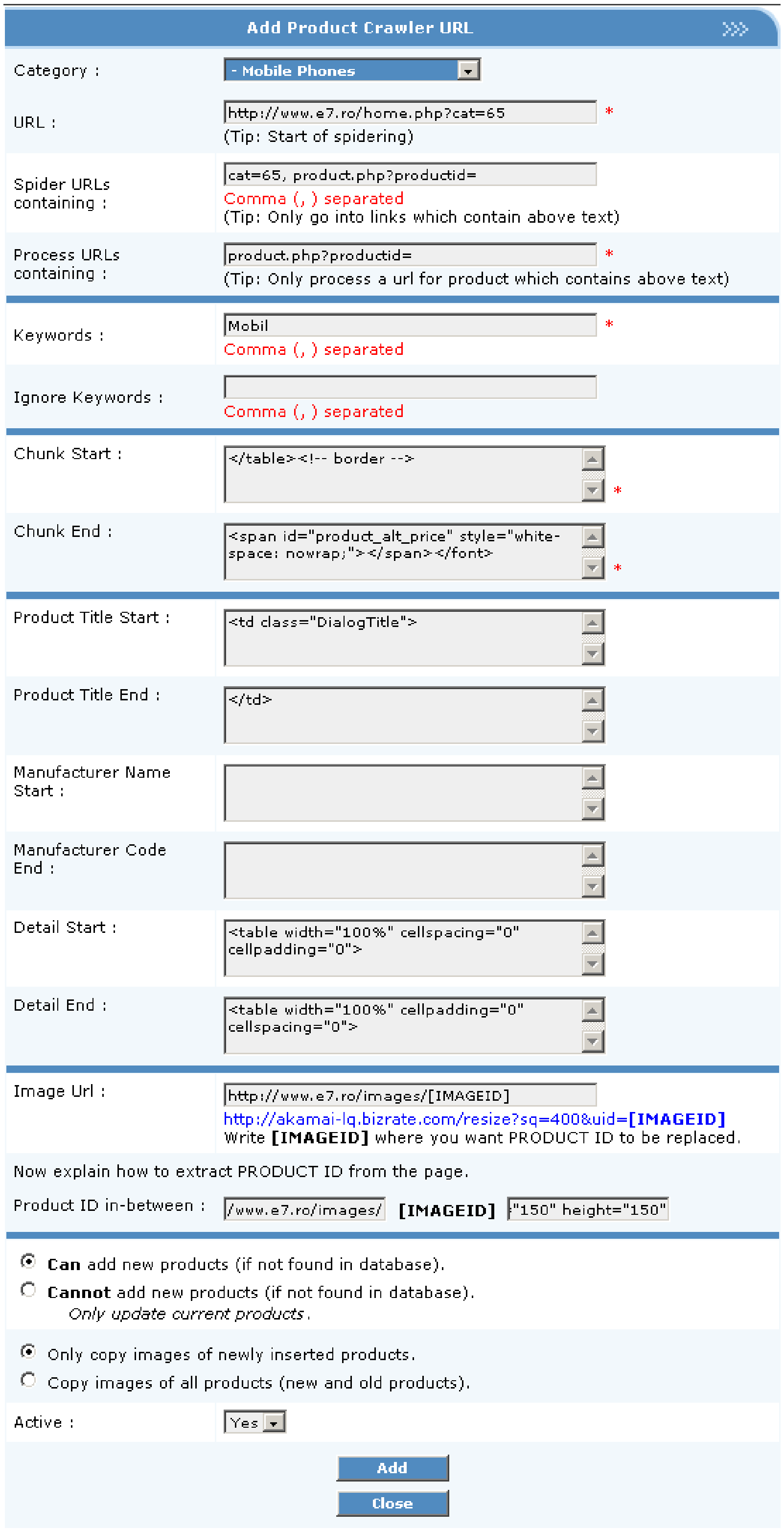
11. Details End
In this case, we can pick <tr><td class="ProductPriceConverting" valign="top"> as end.

11. Image URL
The images on this site are in format: http://www.e7.ro/images/**P/xl_1890.jpg** where **P/xl_1890.jpg** changes with image to image.

In this case add http://www.e7.ro/images/[IMAGEID]

Now you have to also tell the crawler what to replace with **[IMAGEID]**. The image on the webpage is displayed using this html line <img id="product_thumbnail" src="http://www.e7.ro/images/P/xl_1890.jpg" width="150" height="150"> where **P/xl_1890.jpg** is the actual unique id for a product image.

In this case we will enter id="product_thumbnail" src="http://www.e7.ro/images/ as prefix and " width="150" height="150" as suffix.
The reason we entered these tags as prefix and suffix is because **P/xl_1890.jpg** is surrounded by these text.



Add Product Crawler URL

Category :	- Mobile Phones
URL :	http://www.e7.ro/home.php?cat=65 * (Tip: Start of spidering)
Spider URLs containing :	cat=65, product.php?productid= Comma (,) separated (Tip: Only go into links which contain above text)
Process URLs containing :	product.php?productid= * (Tip: Only process a url for product which contains above text)
Keywords :	Mobil * Comma (,) separated
Ignore Keywords :	 Comma (,) separated
Chunk Start :	</table><!-- border --> *
Chunk End :	<span id="product_alt_price" style="white-space: nowrap;"></span></font> *
Product Title Start :	<td class="DialogTitle">
Product Title End :	</td>
Manufacturer Name Start :	
Manufacturer Code End :	
Detail Start :	<table width="100%" cellspacing="0" cellpadding="0">
Detail End :	<table width="100%" cellpadding="0" cellspacing="0">
Image Url :	http://www.e7.ro/images/[IMAGEID] http://akamai-lq.bizrate.com/resize?sq=400&uid=**[IMAGEID]** Write **[IMAGEID]** where you want PRODUCT ID to be replaced.

Now explain how to extract PRODUCT ID from the page.

Product ID in-between : /www.e7.ro/images/ **[IMAGEID]** "150" height="150"

● **Can** add new products (if not found in database).

○ **Cannot** add new products (if not found in database). *Only update current products.*

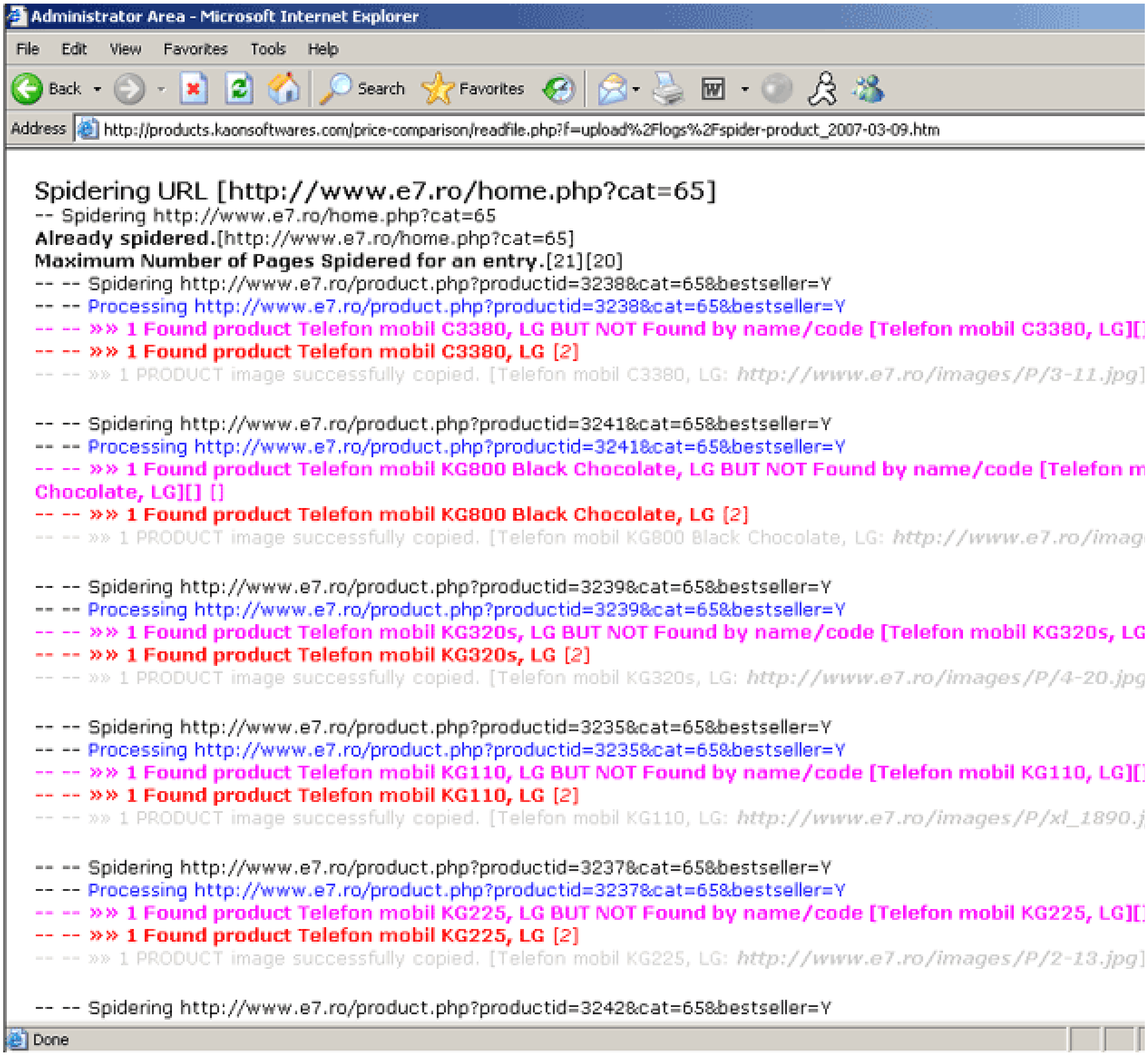
● Only copy images of newly inserted products.

○ Copy images of all products (new and old products).

Active : Yes

Once you have saved these settings. Click on Options in left menu and Run Product Spider. If you have entered everything correctly. You should have all the products with title and description in your database.

You can view log of the actions perform once task is finished by the spider. A similar log will be created as shown below:



```
http://products.kaonsoftwares.com/price-comparison/readfile.php?f=upload%2Flogs%2Fspider-product_2007-03-09.htm
```

```
Spidering URL [http://www.e7.ro/home.php?cat=65]
-- Spidering http://www.e7.ro/home.php?cat=65
Already spidered.[http://www.e7.ro/home.php?cat=65]
Maximum Number of Pages Spidered for an entry.[21][20]
-- -- Spidering http://www.e7.ro/product.php?productid=3238&cat=65&bestseller=Y
-- -- Processing http://www.e7.ro/product.php?productid=3238&cat=65&bestseller=Y
-- -- »» 1 Found product Telefon mobil C3380, LG BUT NOT Found by name/code [Telefon mobil C3380, LG][
-- -- »» 1 Found product Telefon mobil C3380, LG [2]
-- -- »» 1 PRODUCT image successfully copied. [Telefon mobil C3380, LG: http://www.e7.ro/images/P/3-11.jpg]

-- -- Spidering http://www.e7.ro/product.php?productid=3241&cat=65&bestseller=Y
-- -- Processing http://www.e7.ro/product.php?productid=3241&cat=65&bestseller=Y
-- -- »» 1 Found product Telefon mobil KG800 Black Chocolate, LG BUT NOT Found by name/code [Telefon m
Chocolate, LG][] []
-- -- »» 1 Found product Telefon mobil KG800 Black Chocolate, LG [2]
-- -- »» 1 PRODUCT image successfully copied. [Telefon mobil KG800 Black Chocolate, LG: http://www.e7.ro/imag

-- -- Spidering http://www.e7.ro/product.php?productid=3239&cat=65&bestseller=Y
-- -- Processing http://www.e7.ro/product.php?productid=3239&cat=65&bestseller=Y
-- -- »» 1 Found product Telefon mobil KG320s, LG BUT NOT Found by name/code [Telefon mobil KG320s, LG
-- -- »» 1 Found product Telefon mobil KG320s, LG [2]
-- -- »» 1 PRODUCT image successfully copied. [Telefon mobil KG320s, LG: http://www.e7.ro/images/P/4-20.jpg

-- -- Spidering http://www.e7.ro/product.php?productid=3235&cat=65&bestseller=Y
-- -- Processing http://www.e7.ro/product.php?productid=3235&cat=65&bestseller=Y
-- -- »» 1 Found product Telefon mobil KG110, LG BUT NOT Found by name/code [Telefon mobil KG110, LG][
-- -- »» 1 Found product Telefon mobil KG110, LG [2]
-- -- »» 1 PRODUCT image successfully copied. [Telefon mobil KG110, LG: http://www.e7.ro/images/P/xl_1890.j

-- -- Spidering http://www.e7.ro/product.php?productid=3237&cat=65&bestseller=Y
-- -- Processing http://www.e7.ro/product.php?productid=3237&cat=65&bestseller=Y
-- -- »» 1 Found product Telefon mobil KG225, LG BUT NOT Found by name/code [Telefon mobil KG225, LG][
-- -- »» 1 Found product Telefon mobil KG225, LG [2]
-- -- »» 1 PRODUCT image successfully copied. [Telefon mobil KG225, LG: http://www.e7.ro/images/P/2-13.jpg]

-- -- Spidering http://www.e7.ro/product.php?productid=3242&cat=65&bestseller=Y
```

Last Updated: 11th AUG 2014