

**MANUAL FOR USING AND MANAGING  
DATA EXPLOITATION SYSTEM  
VIETSPIDER**

Ha Noi, 2011

<b>1 Overview about VietSpider.....</b>	<b>4</b>
1.1 Introduction.....	4
1.1.1 Preamble .....	4
1.1.2 Presentation form of this book.....	4
1.1.3 Questions and feedback .....	4
1.1.4 What is VietSpider data exploitation system?.....	4
1.1.5 Advantages of VietSpider.....	5
1.1.6 Document author.....	5
1.1.7 Working flow chart of VietSpider.....	6
1.2 Install program.....	6
1.3 How to start and login to the system.....	8
1.4 How to logout the system .....	10
1.5 The main components.....	11
1.5.1 Titlte bar.....	11
1.5.2 Left menu .....	12
1.5.3 Browasing content .....	12
1.5.4 Tool.....	13
1.5.5 Expansion functions.....	14
<b>2 Data exploitation.....</b>	<b>19</b>
2.1 Exploitation of data on the user interface .....	23
2.1.1 Interface .....	23
2.2 Exploitation of content on the browser interface.....	25
2.2.1 Interface .....	25
2.2.2 Login by account on browser.....	26
2.2.3 Logout system.....	26
2.2.4 Search.....	26
<b>3 Channel Management.....</b>	<b>27</b>
3.1 Create new channel.....	27
3.2 Make new category.....	28
3.3 Remove a category or channel.....	29
3.4 View current or removed categories .....	31

3.5 Restore deleted categories or data flows.....	32
3.6 Copy channel .....	34
3.7 Cut channel .....	34
3.8 Paste channel.....	35
3.9 Send channel or category to other computer .....	37
3.10 Rename .....	38
3.11 Crawl deep .....	40
3.12 Set crawl frequency .....	41
3.13 Configure poperties.....	42
3.14 Reload. ....	47
3.15 Apply to other channels. ....	48
3.16 Export.....	49
3.17 Import.....	51
3.18 Create a new channel .....	53
3.19 Crawl data .....	71
3.19.1 Interface .....	71
3.19.2 Crawl content.....	73
3.19.3 Pause crawling .....	73
3.19.4 Add crawling channel .....	73
3.19.5 Remove channel from the list of scanned channels .....	74
3.19.6 Right-click menu.....	75
3.20 User management .....	76
3.20.1 User account .....	77
3.20.2 Manage catagories in user group .....	80
3.21 Configuration .....	84
3.21.1 Configuration menu. ....	84

## 1. Overview about data exploitation system VietSpider

### 1.1 Introduction

#### 1.1.1 Preamble

- **Data exploitation system VietSpider** is the solution in exploiting, collecting and categorizing data from the internet serving specific purposes.
- This documentation will provide you with an overview of how best to work with the program. A lots of specific examples as well as detailed instructions will be given to help you know how to use and manage the program easily.

#### 1.1.2 Presentation form of this book

- The contents of the document is presented in detail and complete

#### 1.1.3 Questions and feedback

- VietSpider always welcome user's feedback. We hope that this document is the best one in helping the user working with the program. We always check for program development in order to further improve it in our ability. So, if you have questions and feedback you can contact to our email address: [nhudinhthuan@yahoo.com](mailto:nhudinhthuan@yahoo.com)

#### 1.1.4 What is the data exploitation system VietSpider?

- The main function of the system is exploiting or collecting data from websites then analyze and categorize for specific purposes. The system focuses on three flowing main components:
  - Data exploitation and separation.
  - Data aggregation, indexing and semantic analysis.
  - Data publishing and searching on browser.

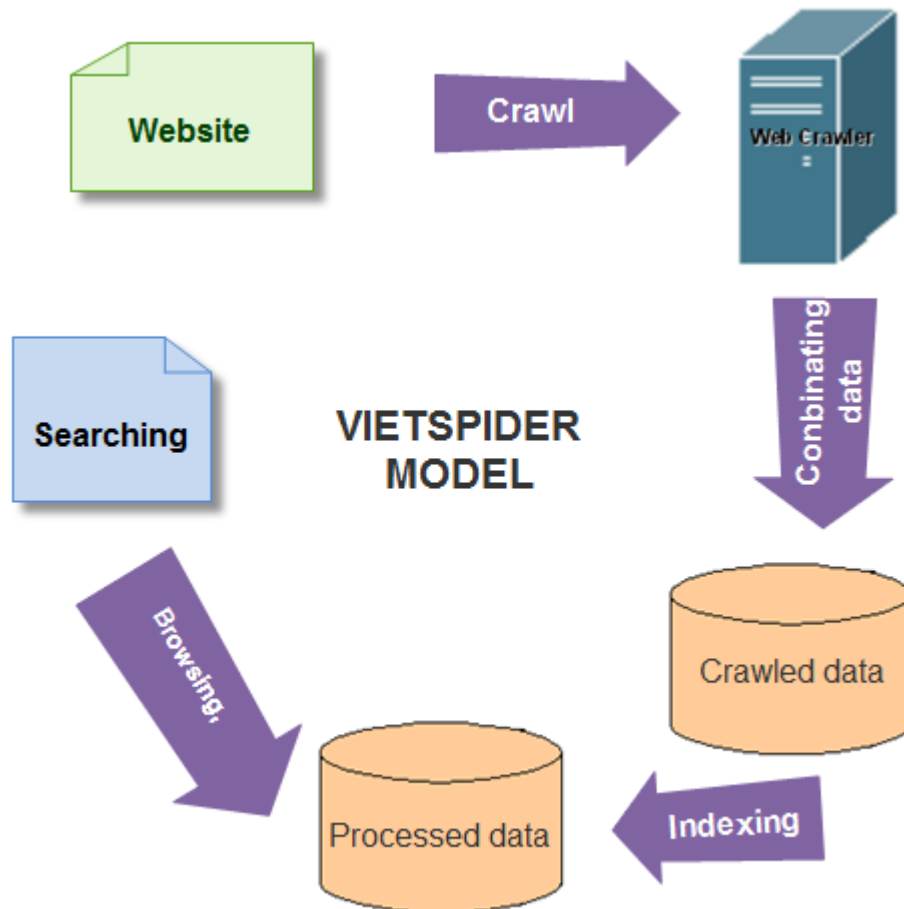
## 1.1.5 Advantages of Vietspider

- Vietspider was developed on Java platform so the program can run on multiple platforms of personal computer or server with different operating systems like Windows, Linux, Solaris, Mac OS ...
- Indexing solution combined with database management in data storage.
- Semiautomatic solution in analyzing and refining data allows administrators can access remotely to manage applications on a graphical user interface.
- Web interface solutions emphasize aspects of search and quick browse data.
- Semantic analysis system for automatic classification and linking of related content.

## 1.1.6 Document author

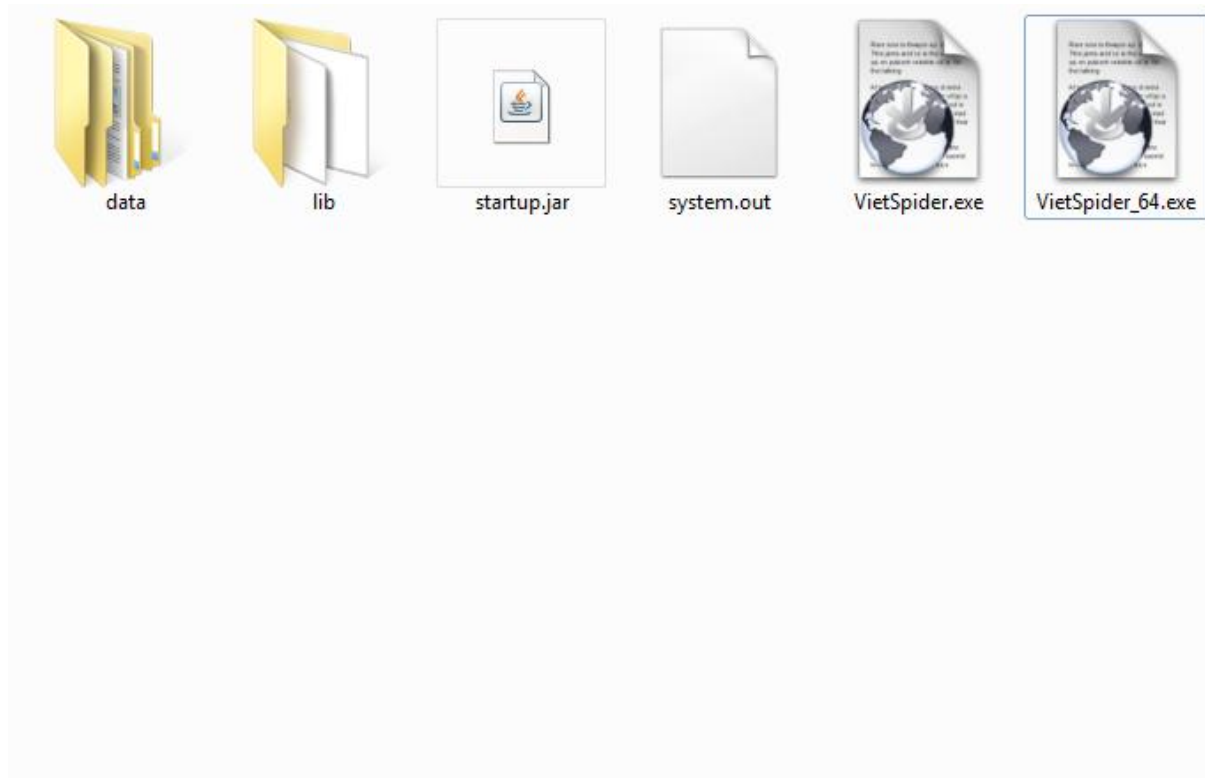
- Nguyen Quang Truong, email: [truongnq@mails.hanu.edu.vn](mailto:truongnq@mails.hanu.edu.vn), Gtalk: truongnq2306;
- Do Ngoc Tuan, email: [tuandn@ncvietnam.com](mailto:tuandn@ncvietnam.com)

### 1.1.7 Working flow chart of Vietspider

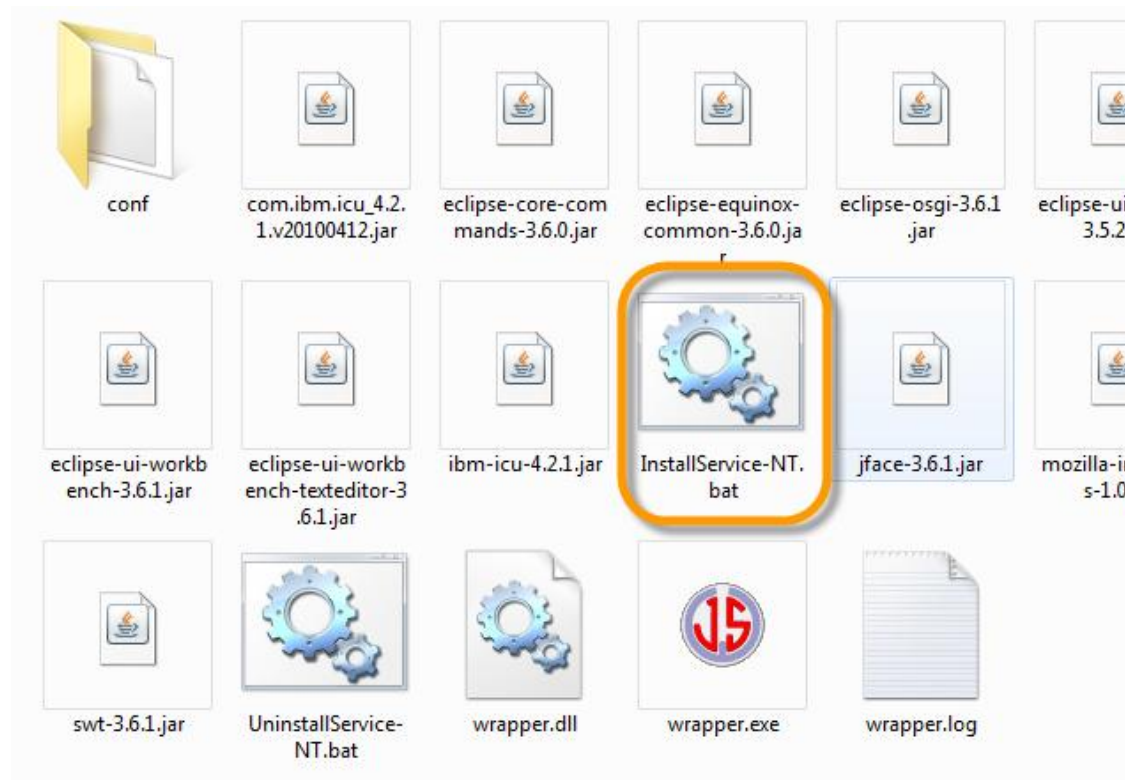


## 1.2. Install program

- In order to run Vietspider you must make sure that your computer has java installed (JDK or JRE) version 1.6 or higher, otherwise please download and install [here](#). Some Vietspider versions have a built-in JRE (you will find JRE directory in root folder after unpacking Vietspider) so it does not need to install java anymore. Vietspider is software that does not need to install, just unzip downloaded package and run the program. The following image displaying all files and folder after unzipping Vietspider package.



- You can choose one of the following files to run VietSpider:
  - **VietSpider.exe:** Run completely VietSpider that includes VietSpider Server (server program) and VietSpider Client (other program to control the system). New user can run this file VietSpider.exe and begin to learn how to use the program.
  - **VietSpiderServer\_XX\_XXX.exe:** VietSpider server programs run independently and launch those files you already start VietSpider Server program on your own computer. (For 64 bit OS, you should run file VietSpiderServer\_64\_2048.exe, and other files: VietSpiderServer.exe for 32 bit machine using 512mb RAM, VietSpiderServer\_32\_1024.exe for 32 bit machine using 1gb RAM, VietSpiderServer\_32\_2048.exe for 32 bit machine using 2gb RAM.
- **VietSpiderClient.exe:** After you started VietSpider Server, you can start VietSpider client to use and control VietSpider on the user interface.
- **Installing Window Service for VietSpider:**
  - Open the directory VietSpider -> lib -> windows\_32



- Run file InstallService-NT.bat.
- Then open folder Control Panel -> Administrative Tools -> Services. Check for exist of a service named VietSpider.
- Click start to run VietSpider Server.
- Then you can use VietSpiderClient.exe connecting to server and work normally with the program.
- In order to remove service, just go to folder VietSpider find and run file UninstallService-NT.bat.
- This feature only comes with the enterprise version of VietSpider including from professional to higher one.

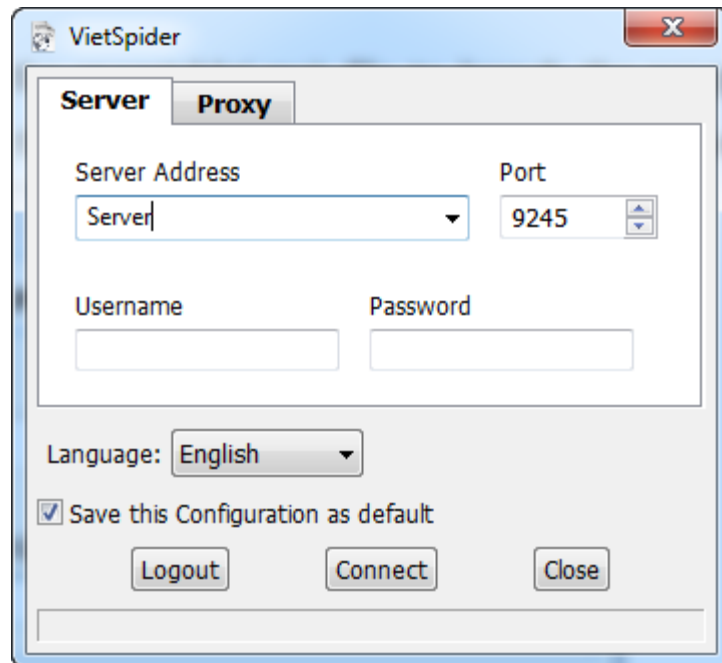
## 1.3 Start and login to the system.

### Follow steps:

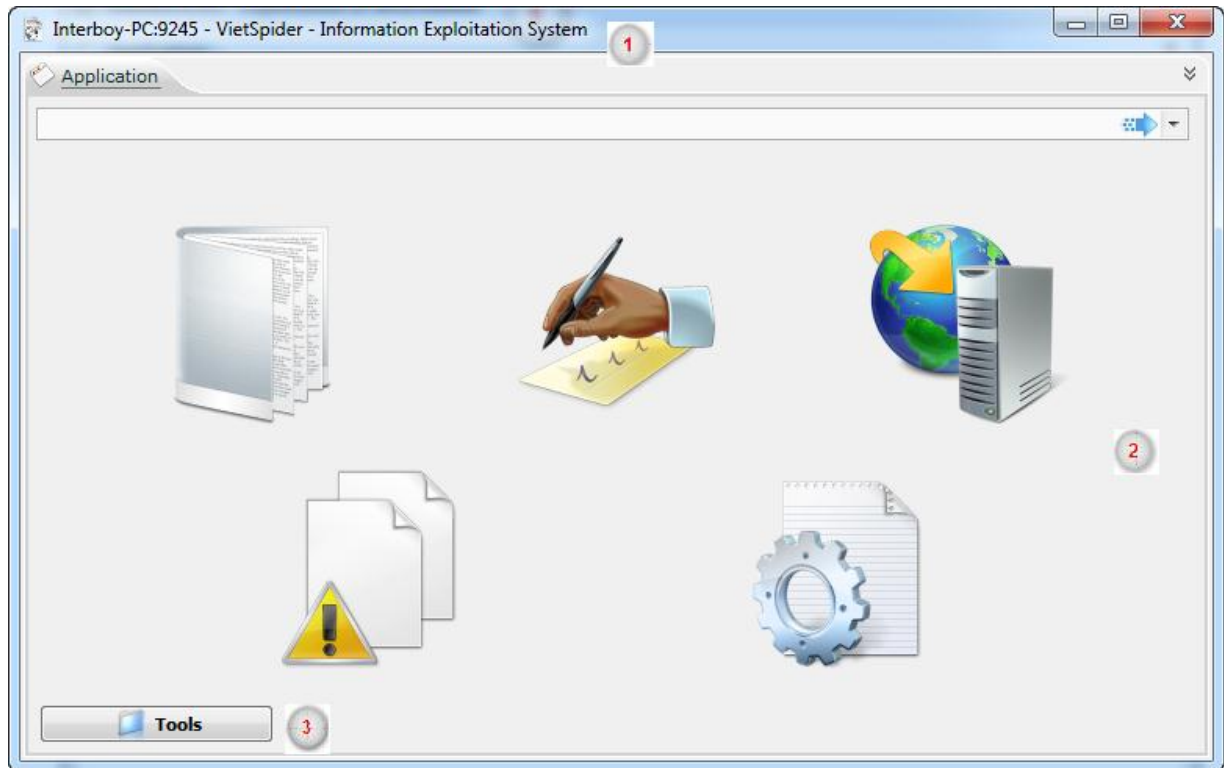
1. Login process to the system helps to check the legality of user accessing to work with the program. Sign in parallelly with confirming access rights also issued operation restrictions on the system.



2. Click on the executable(.exe) file to launch the program. A window screen appears displaying the requirement notification of the information to login to the sytem.



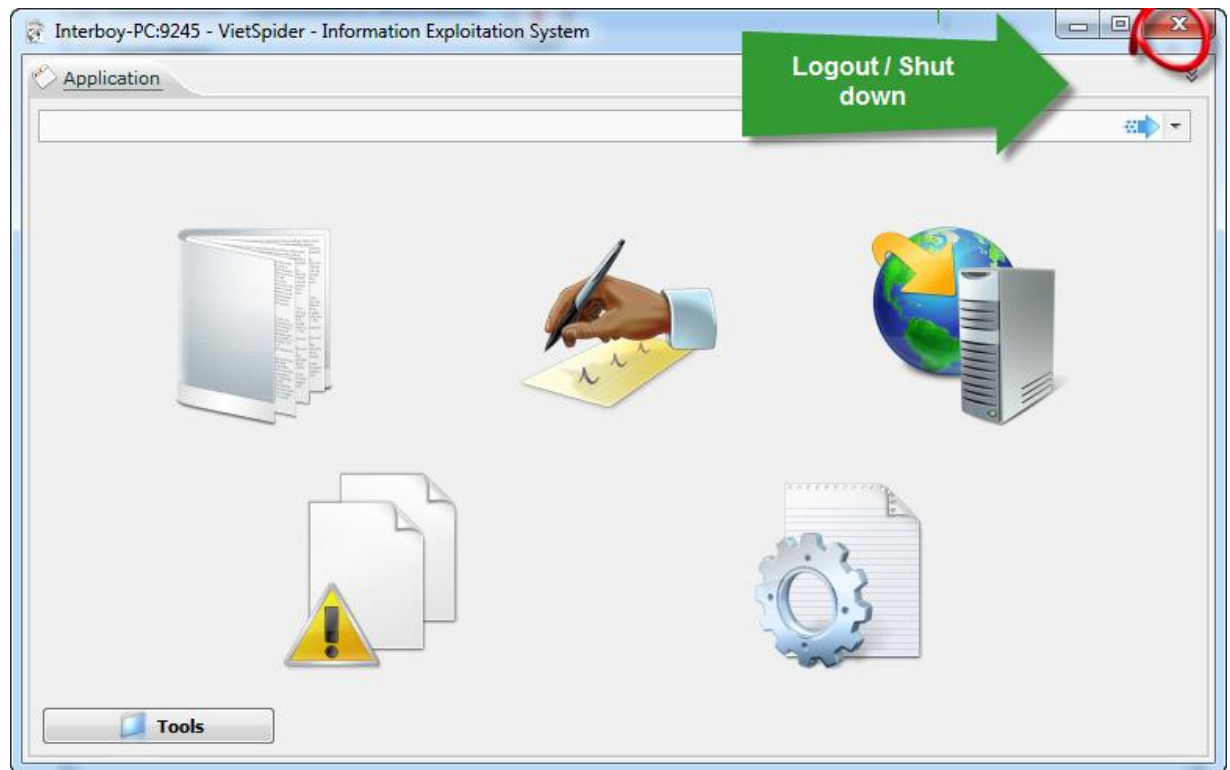
3. Fill in form the following information: **Server address, port, account name, password.**
4. Click on 'Access' button or press 'Enter' to access the program.  
*Note:* If one of entered values is incorrect or invalid, the program will display an error message and require user re-entry until all needed information is correct. If you do not want to login to the system again, then click on **Close** button to exit the program.
5. After successfully login into the system, the program will have the following interface:



Title	Content
1	Address of the server system is accessing.
2	Icons present program's functions
3	Some additional features of the program.

#### 1.4 How to logout the system

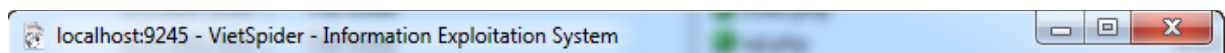
1. Logout process will help user exit the system. User can click on the closing button of window or click on the text **More plugins...** on the bottom left corner of the program, then select **Logout** ( *as image below* ) to exit the program.



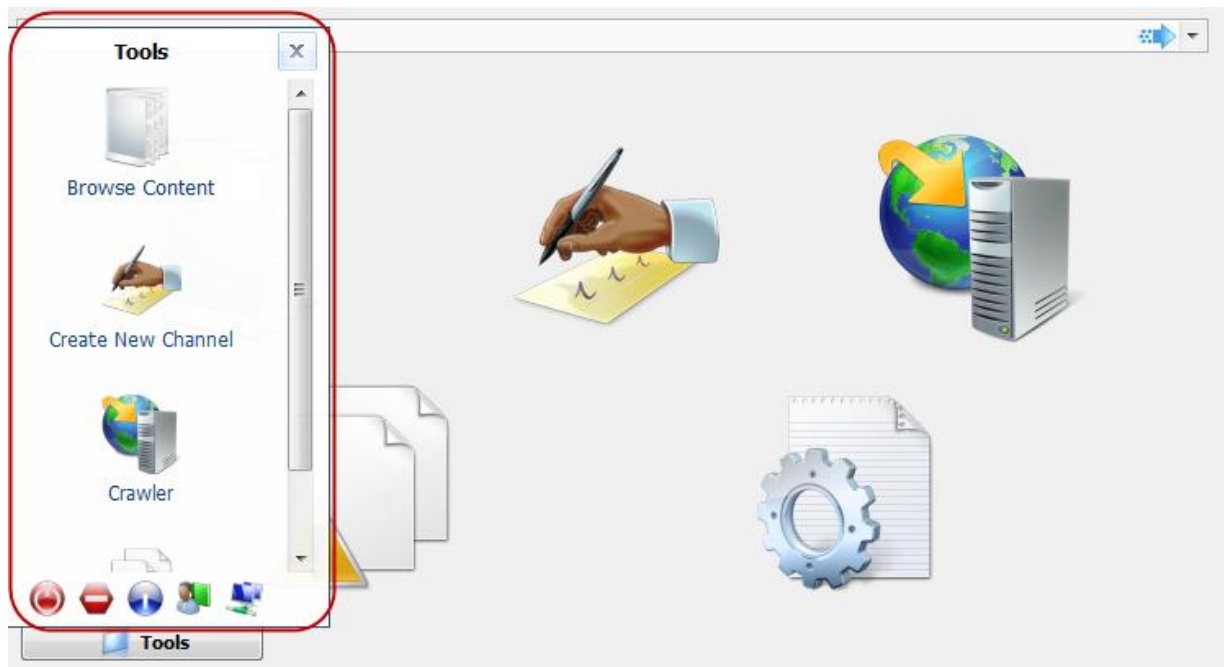
## 1.5 The main components

### 1.5.1 Title bar

- On the left side of title bar, the information about the server address accessing displayed. The following picture illustrates server address is **localhost** working on port 9245.

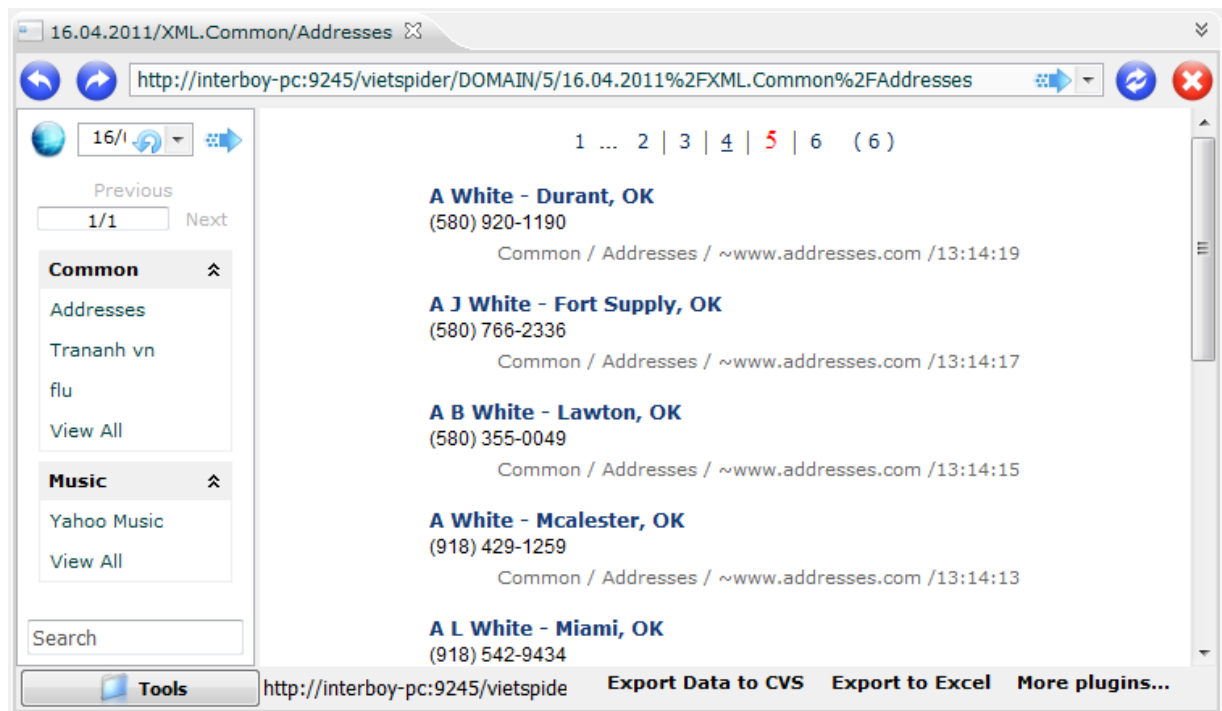


### 1.5.2 Left menu



### 1.5.3 Browsing content

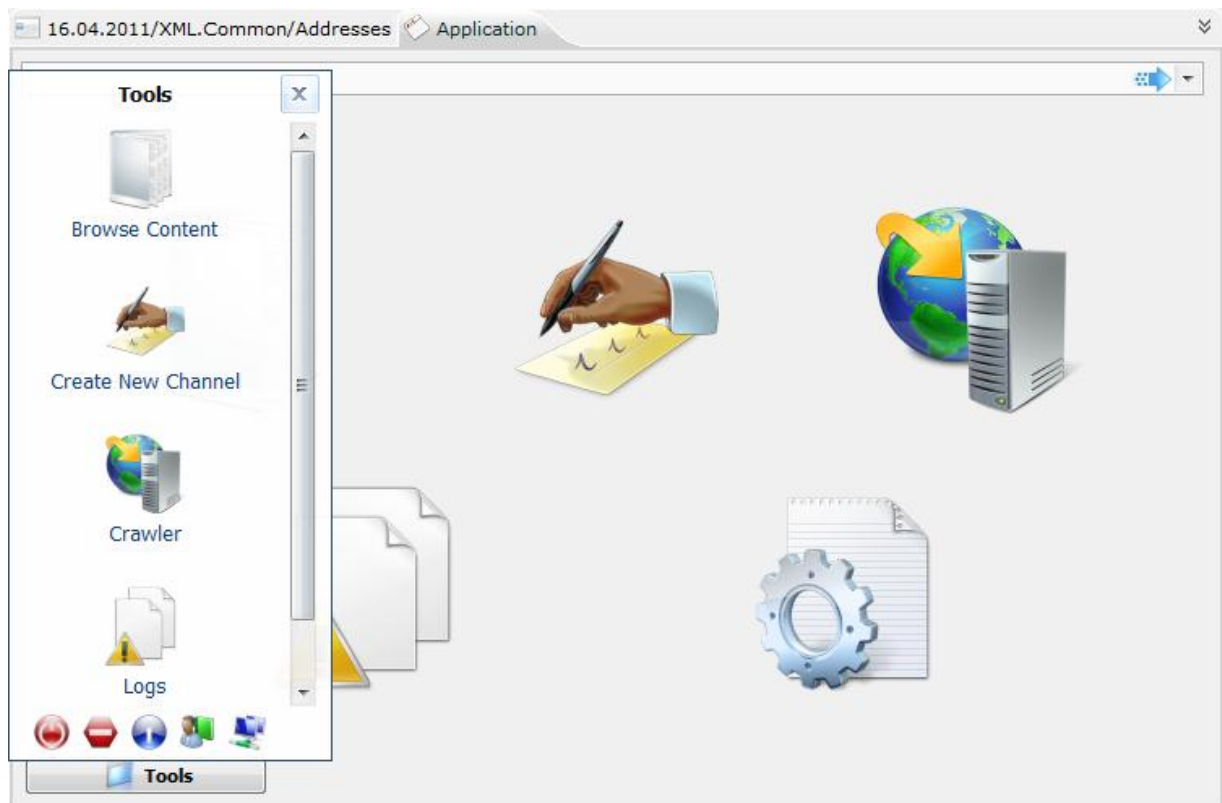
- **Content** part include list of channels that is containing data. This section provides ways of accessing data on different interfaces.
- When user selects a component in the **Content** category, the main window will show list of data as shown below.



### 1.5.4 Tools

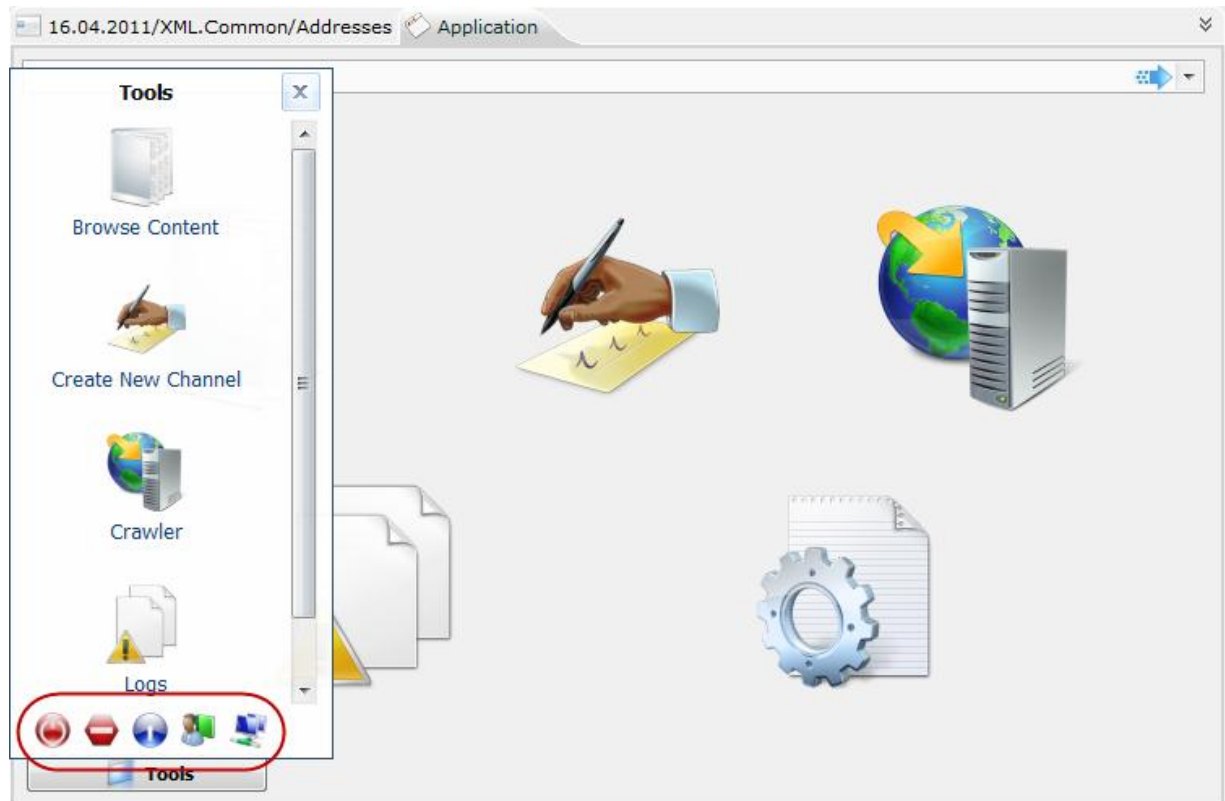
- They are components allowing management actions being implemented on the system
  - ✖ Creating and managing news channels.
  - ✖ Managing list of data flows and exploitation process.
  - ✖ Monitoring errors, warnings in data downloading process.
  - ✖ Monitoring and managing user.
  - ✖ Configuring parameters for crawling program and application system.

The picture below illustrates **Tool** section.

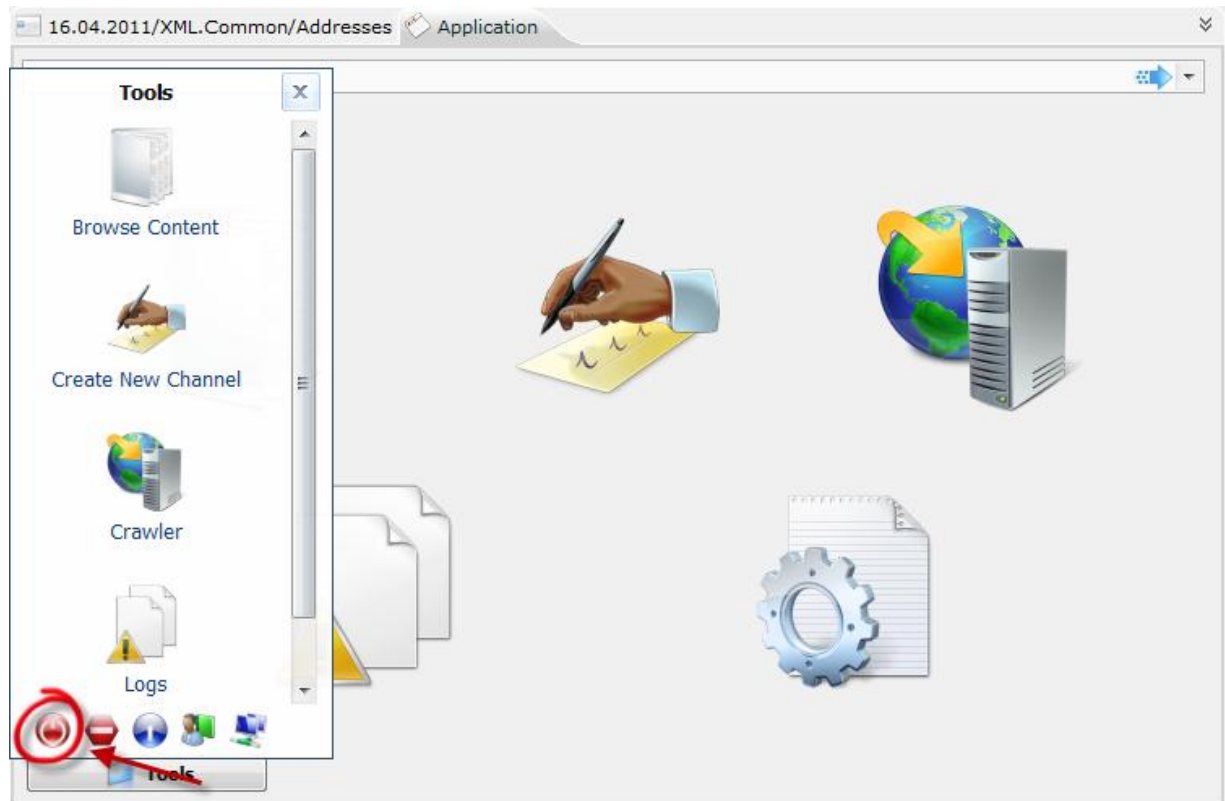


### 1.5.5 Expansion functions

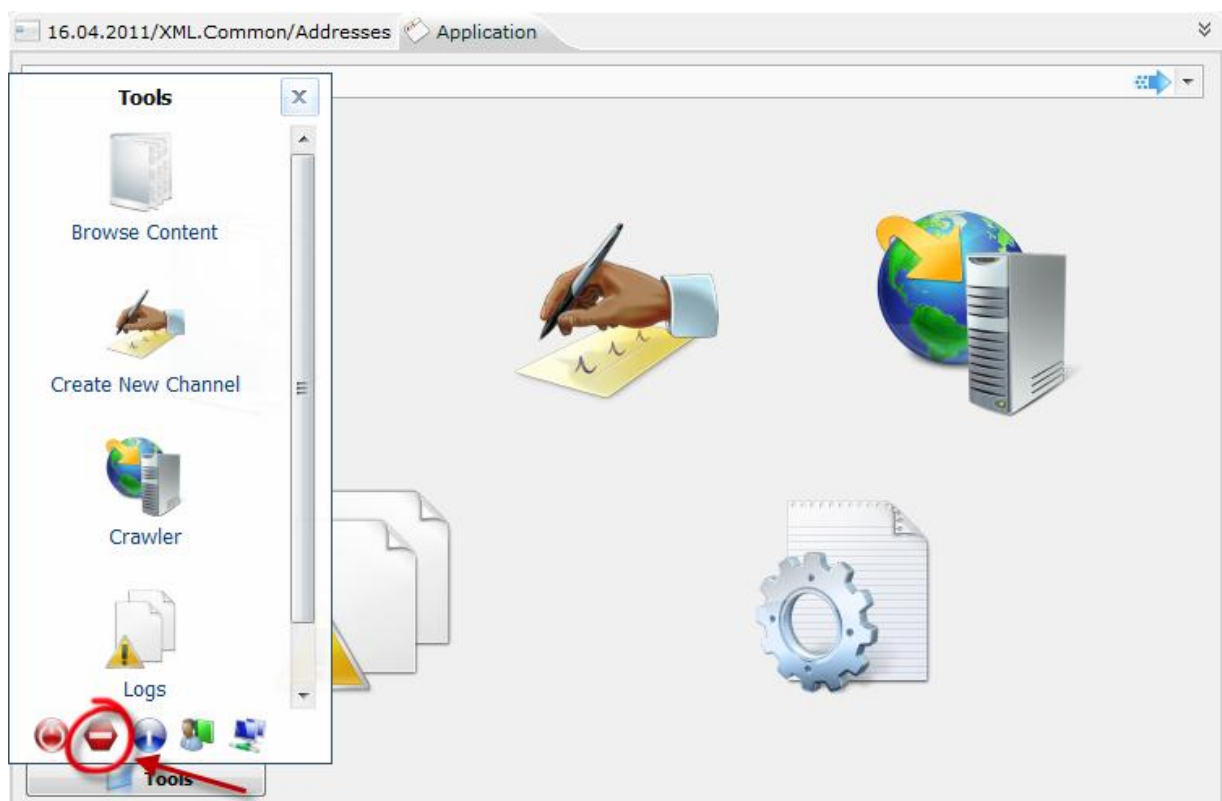
- Bottom of **the menu left** is expansion functions of program such as using guide document, logout, information about program, ...



- 1. Turn off VietSpider Server:** when user click on this function, the server of program immediately disable. Then if user wants to continue working on the program, he or she can enable VietSpider server again.

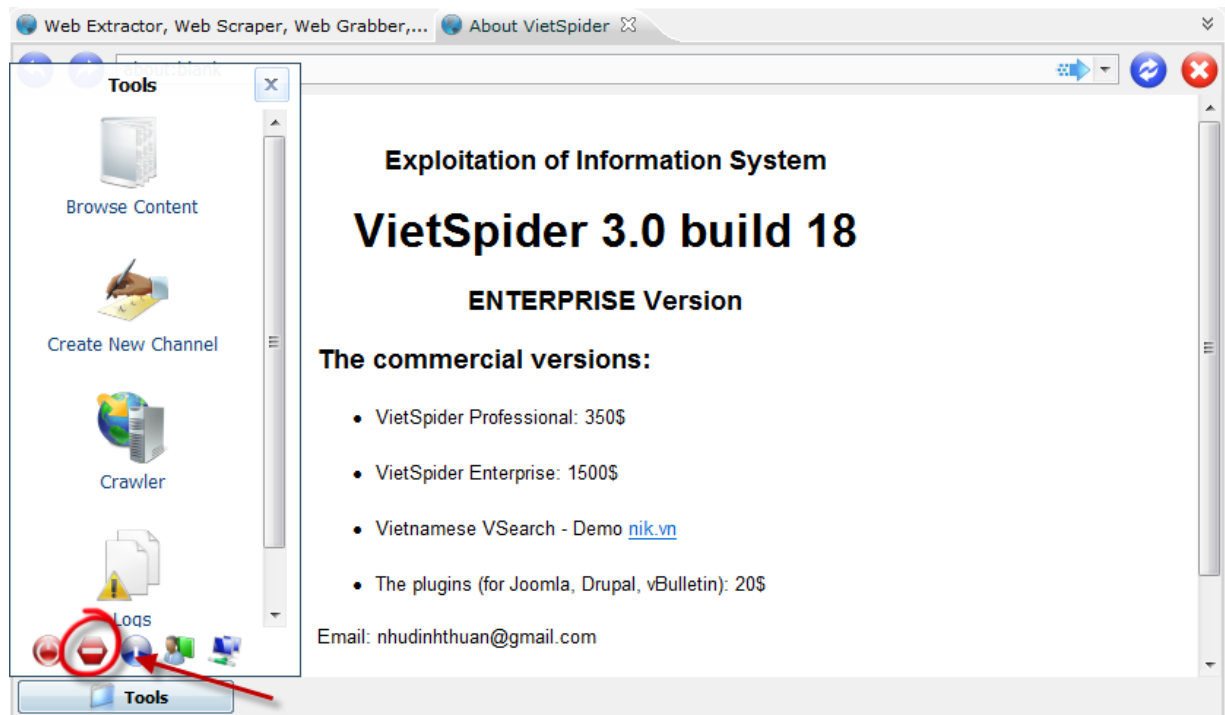


2. **Logout:** When user click on this function, the program will shutdown. Then, in order to work with the program, user must login the system again.

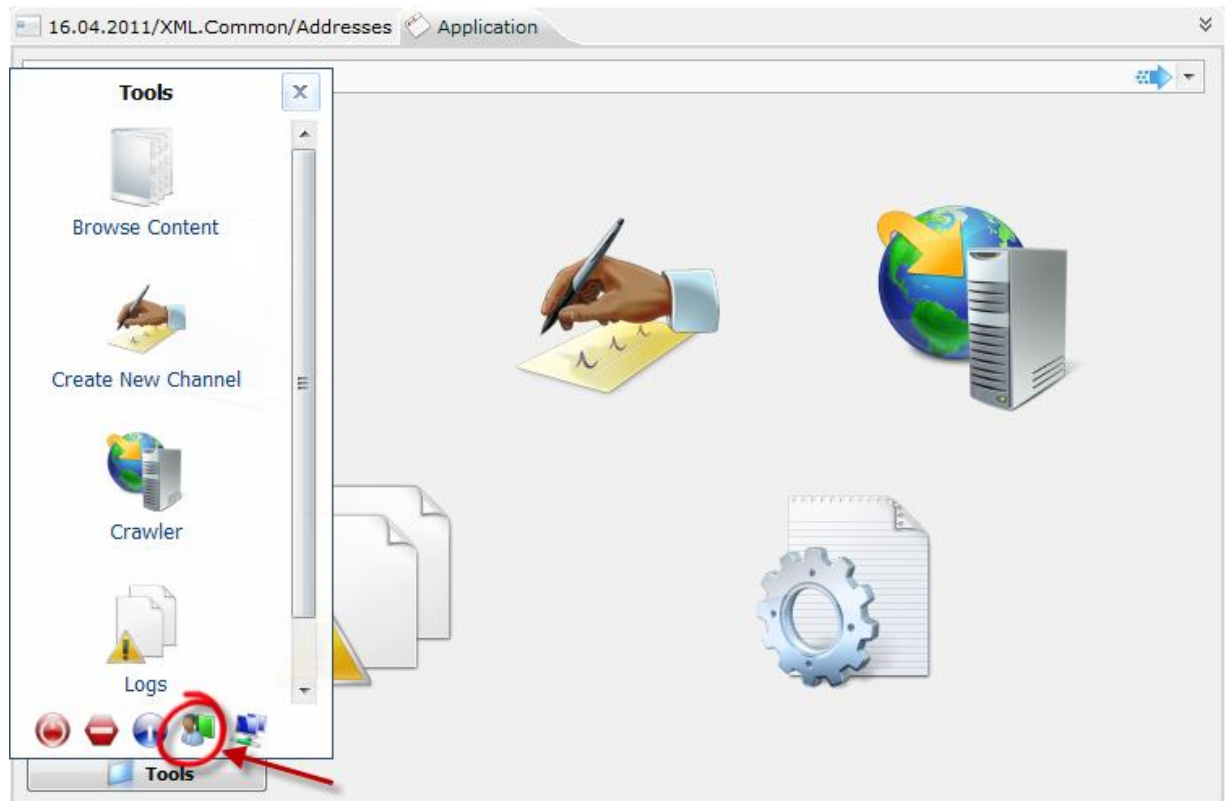




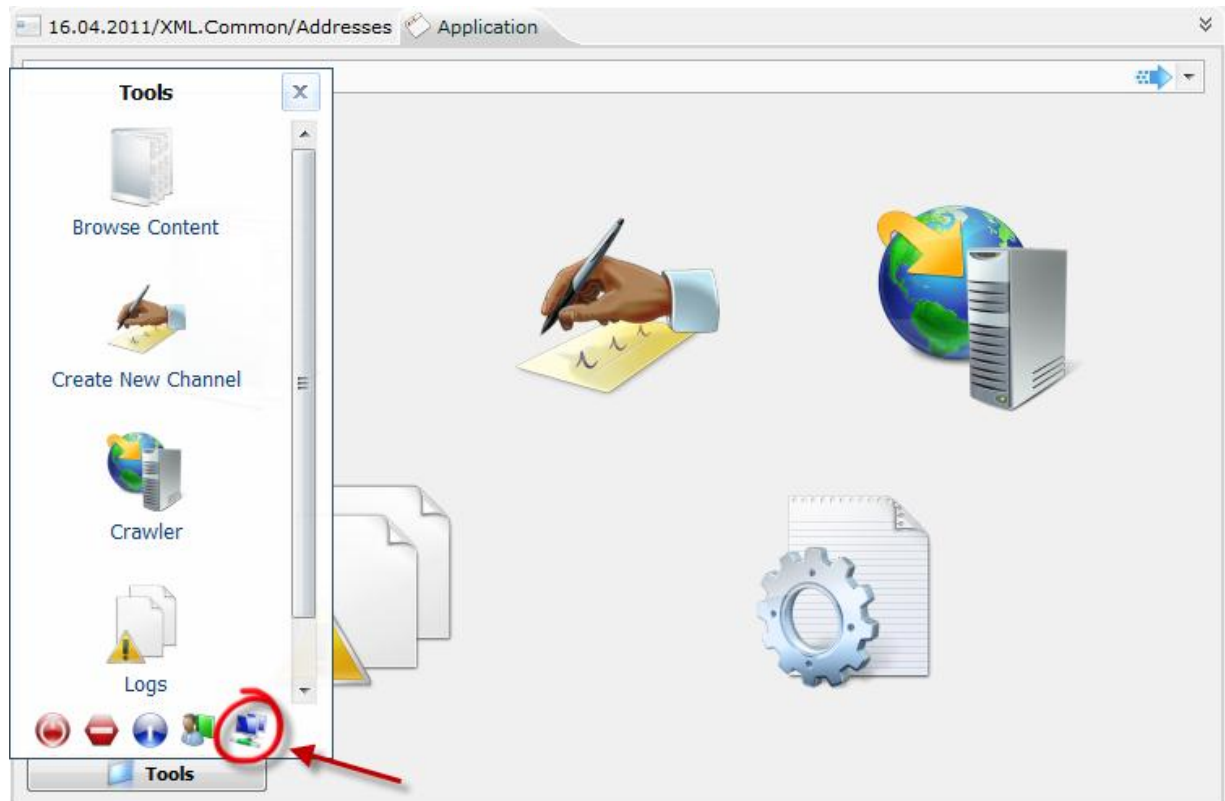
- 3. About program:** When user clicks on this function, the basic information about VietSpider program such as versions, contact details, will be displayed.



- 4. User guide:** This function provides user detailed guide documentation about the program. User should follow it then he or she can work with the program easily.

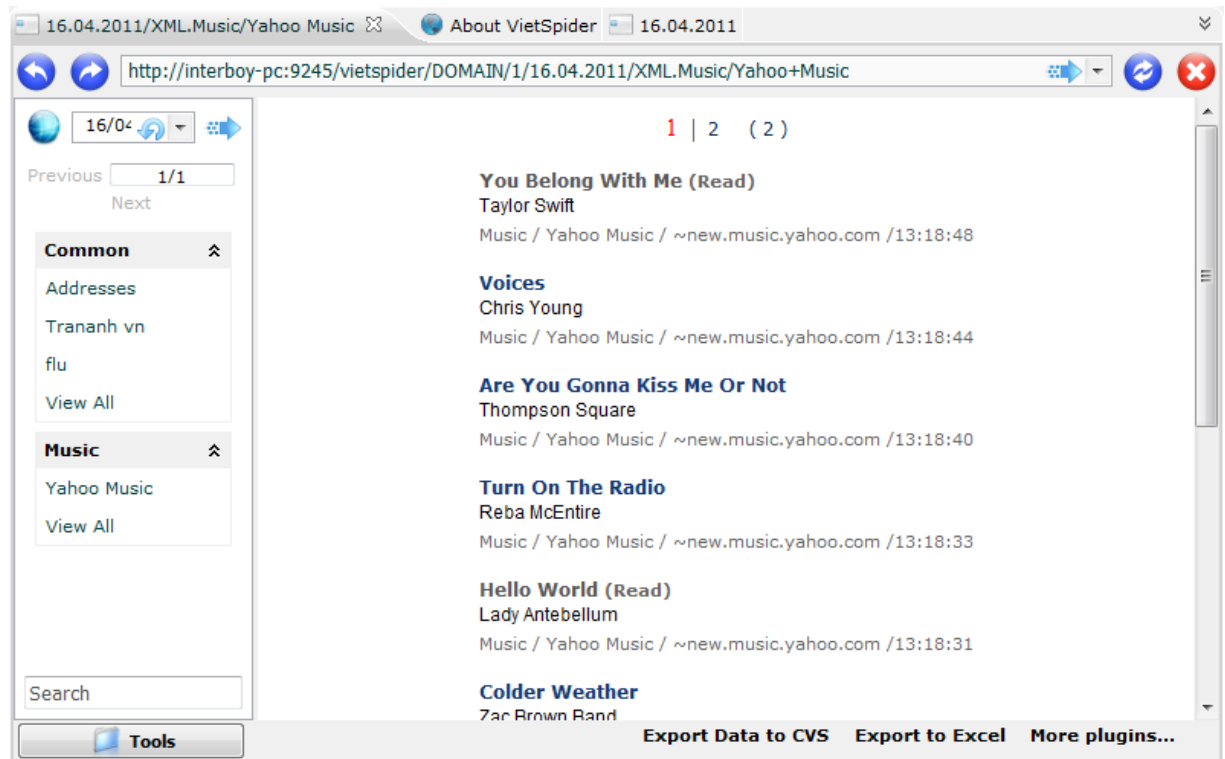


- 5. Connecting to other computers:** This function can help user connect to another computer that is also running VietSpider program. It's very useful and convenient because user can manage many machines remotely without directly sitting in front of them to control.

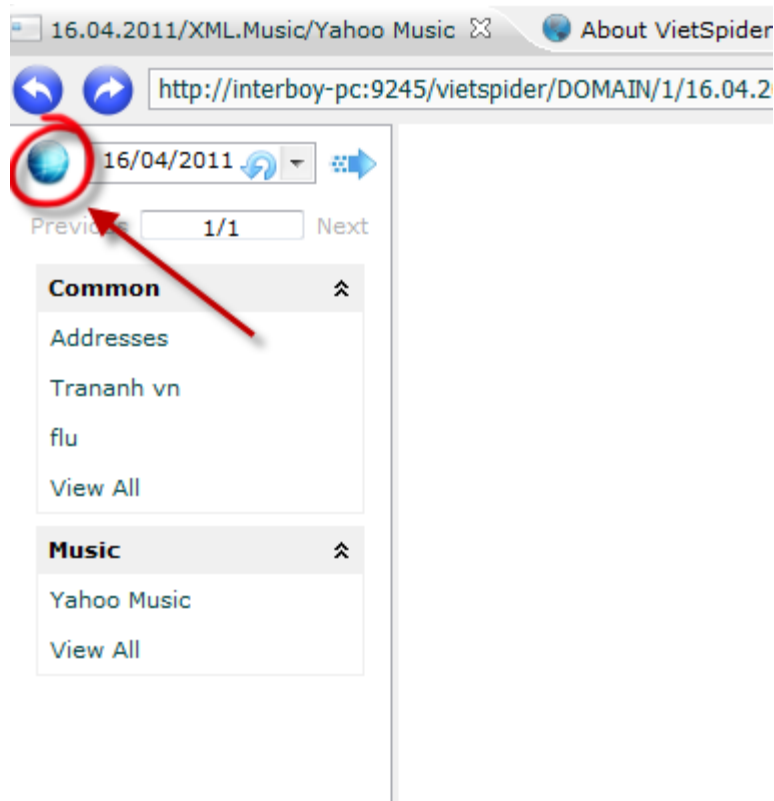


## 2. Data exploitation

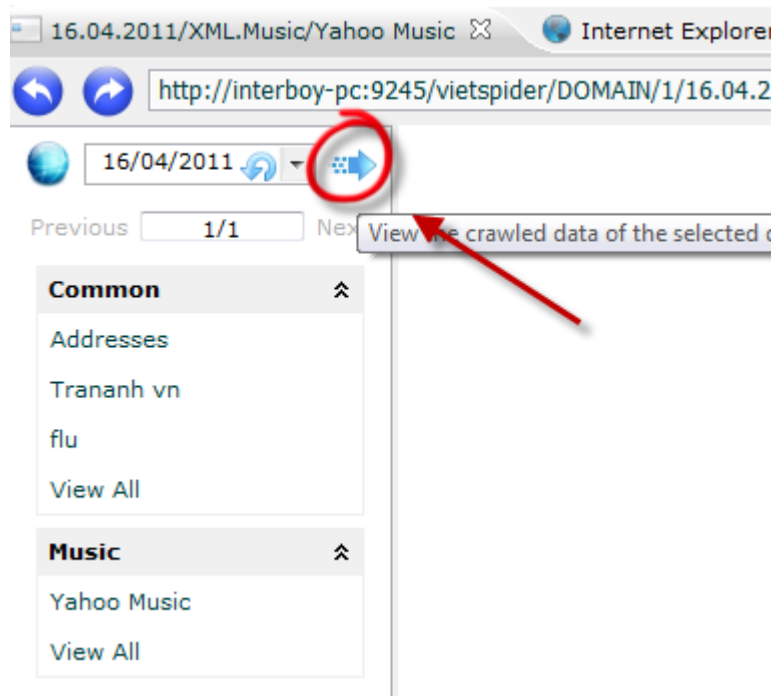
- This function exploits downloaded data from application. Data exploitation provides ways to access, search, browse... differently on areas of information that user is allowed to access.



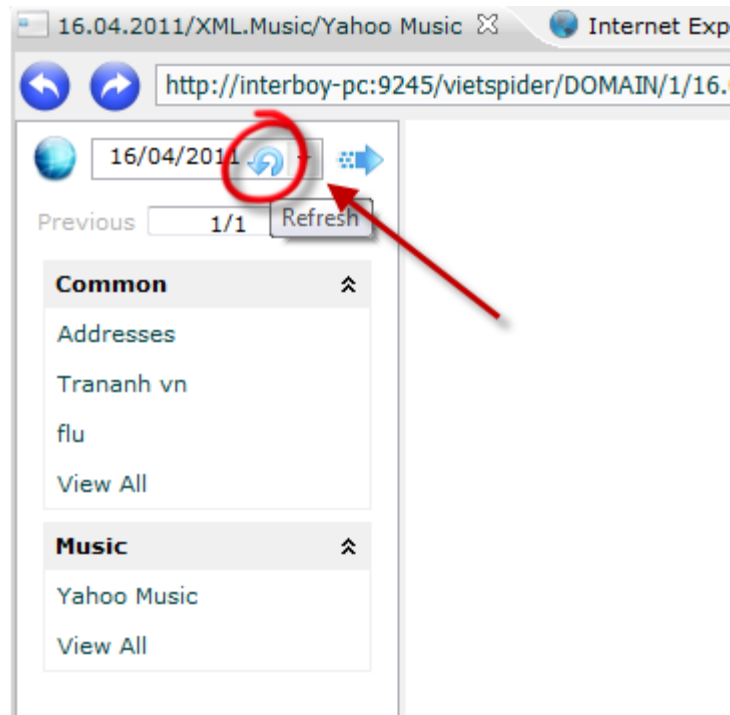
- Data exploitation includes the following features:
  - Crawling on the user interface
    - Functions of **Content** section on the user interface
    - Some main manipulations supporting data browsing and management
    - ✖ **Notable data:** Remarkable data downloaded in the day.



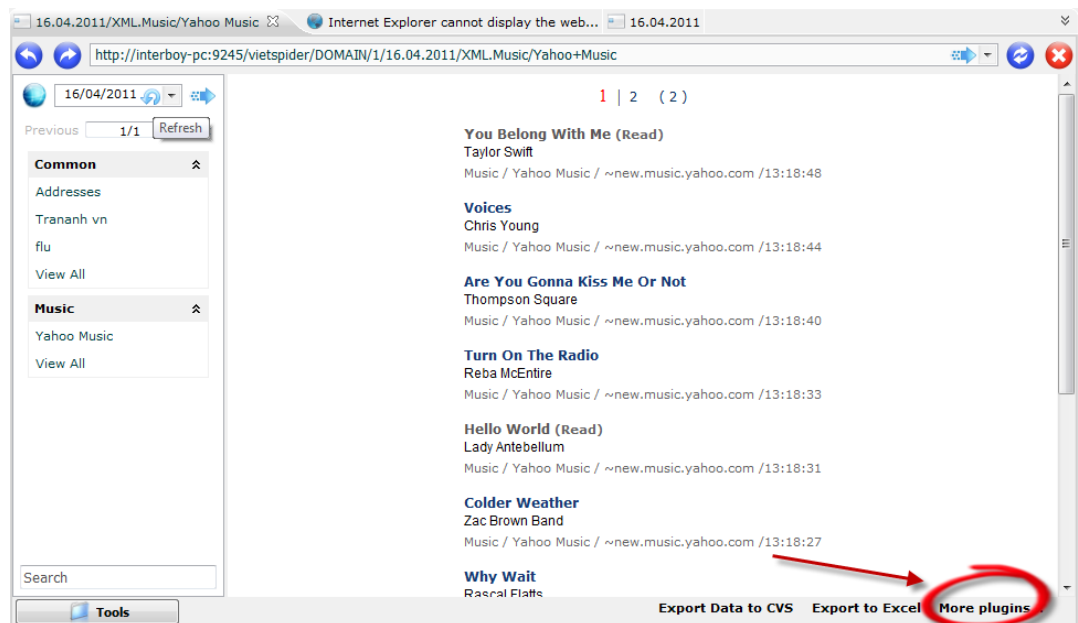
✱ **Browsing news of the day:** View all downloaded data in the day of the program.



✱ **Update:** Updating latest news that the program crawled.



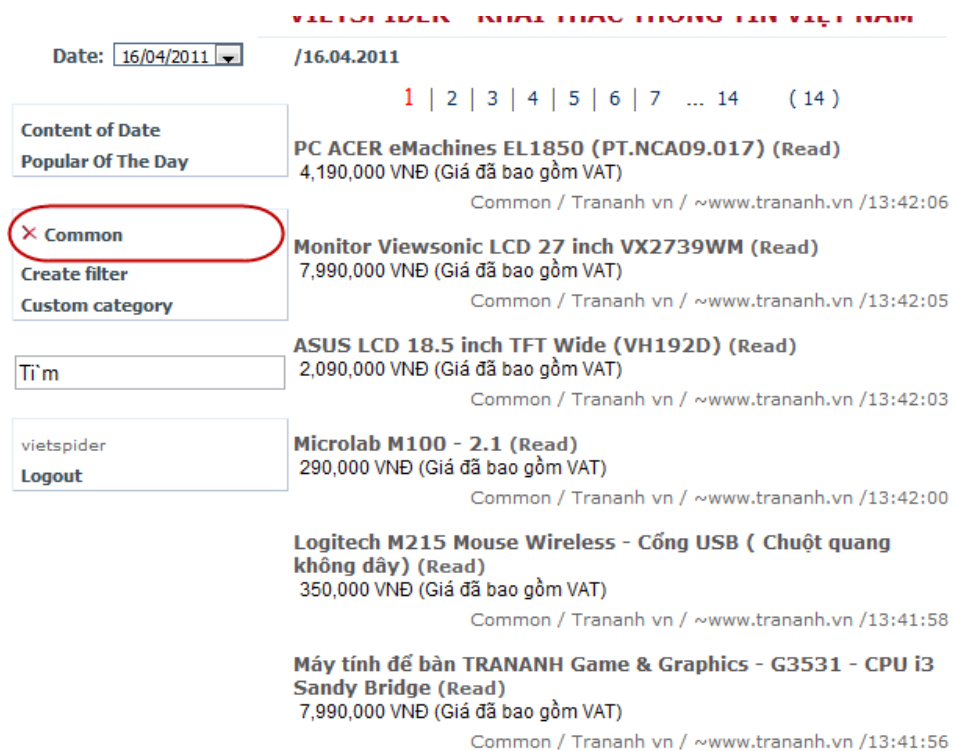
✱ **More plugins....** Some features of programs such as: deleting the content, exporting data, downloading configuration from center flow, adding more channels needing to get data, turning off server, index.



- Exploitation of content on the browser interface

- Including the following functions:

- ✱ **Browsing content by category or data source:** Allowing user to browse content based on category or corresponding data source.
    - ✱ **Personalized access:** Personalizing categories for browsing data and creating filters to purify the content.

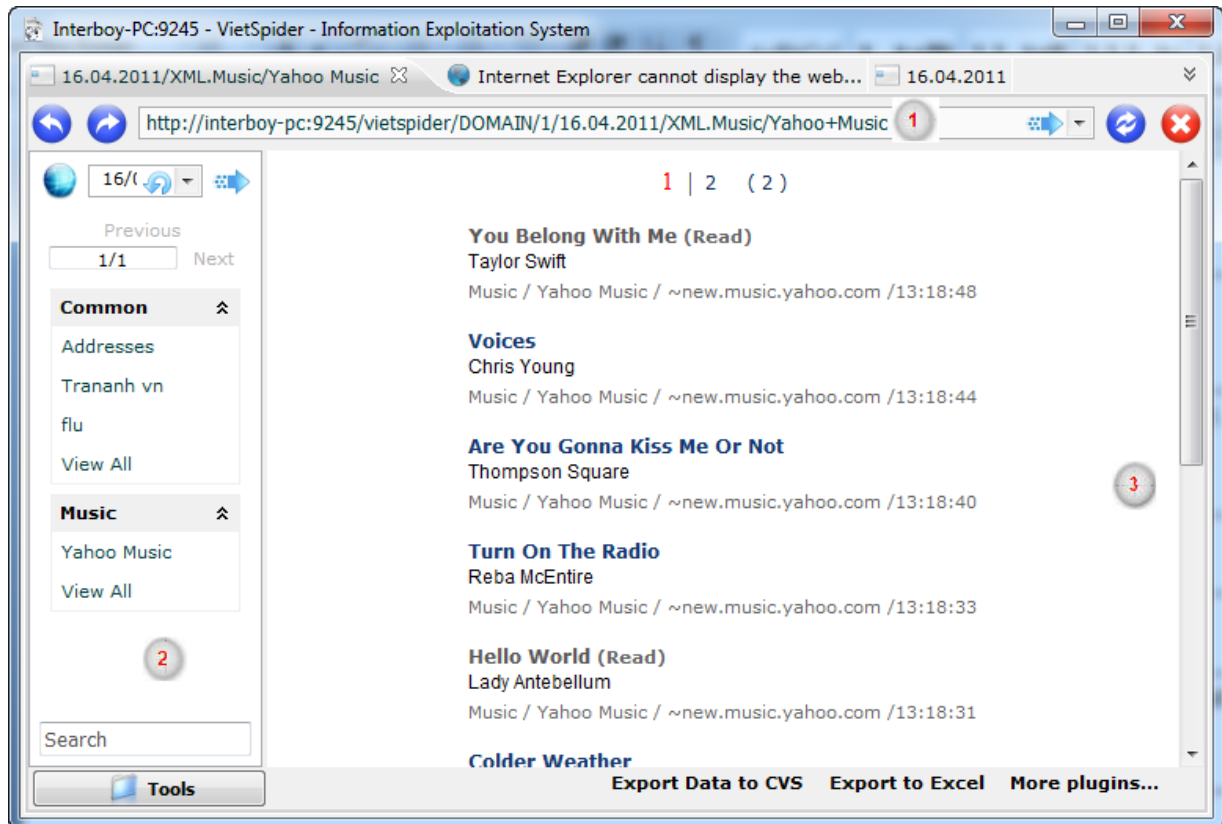









- ✱ **Browse data overall:** This function allow browsing all content of selected day , then collecting entire of outstanding events in that day.
    - ✱ **Search data:** Find data in the system based on given keywords.

## 2.1 Data exploitation on the user interface

### 2.1.1 Interface

- Used to view and manage basically **Content** scanned by the program.



Title	Content
	Crawling address
	List of categories and sources that have data.
	List of displaying contents that based on a browsing way.
	Symbol representing for return previous action.
	Symbol representing for going to next action.
	Symbol representing for updating categories containing data of selected day.
	Icon representing for stopping update data in the day.





Browsing news in the day.



## 2.2 Data exploitation on the browser interface

### 2.2.1 Interface

- The system allows users could exploit data on browser with the support of many functions and different ways.
- You can use internet explorer or Firefox browser for implementing the program.
  - At the address bar of browser, user will type address form like: <http://host address : port/> (ex: on my computer, the address I typed is: <http://localhost:9245/>). The following interface will appear.



Title	Content
-------	---------

	Providing functions about browsing, filtering, searching,... different information.
	List of categories or detailed contents will be displayed.

### 2.2.2 Login by account on browser

- When users logged into the program on browser, they can have more personalized functions for browsing, filtering data such as creating categories or filters,...
- On the interface of the browser, when user clicks on **Sign in** function.
- Users have to fill their own information account on the system like **account name**, **password** then click on **login** button to continue.
- After login successfully, the system will redirect to homepage and display name of accessing account.
- Some special functions of system like creating categories, filters are only available when user logged in to the program.

### 2.2.3 Logout system.

- Log out (exit) from the account signed. The program returns status of public user.
- In order to logout, user can click on **logout** button on left menu of program.

### 2.2.4 Search

- **Search** function help user find information based on keywords.
- On the browser interface, user will fill keywords needing to find in searching textbox at left menu.

**Note:** +If user want to search exactly one phrase like **A B**, he or she should put that phrase in quote mark like “**A B**”. Then, the result (if satisfied) will contain exactly the phrase **A B**.

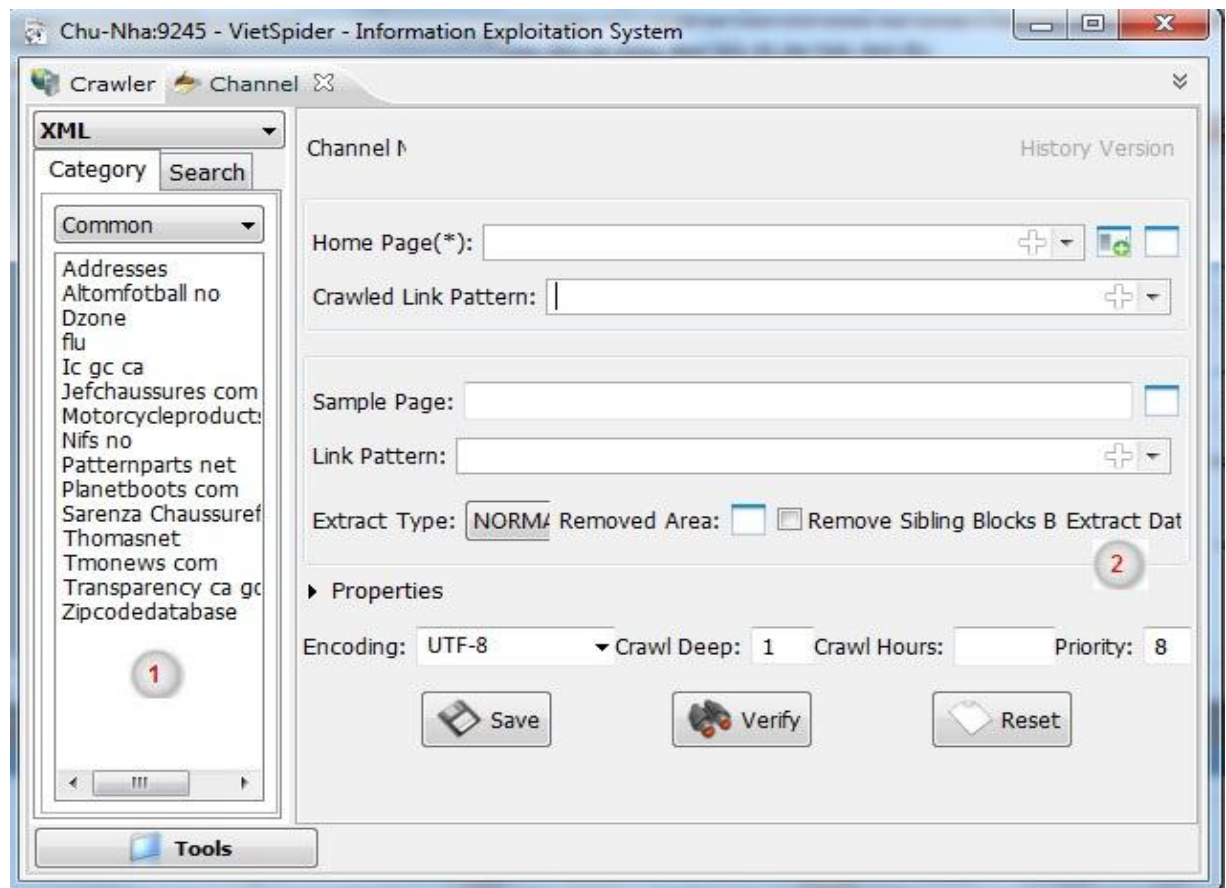
+If keywords separated by space, the system will implement **OR** algorithm. This means the program will find keyword **A** or **B** instead of both **AB**.

+ If keywords written together, the system will carry out **AND** algorithm. This means the program will find data containing keyword (**A**) + (**B**) or **+A+B**

### 3. Channel management

#### 3.1 Create new channel

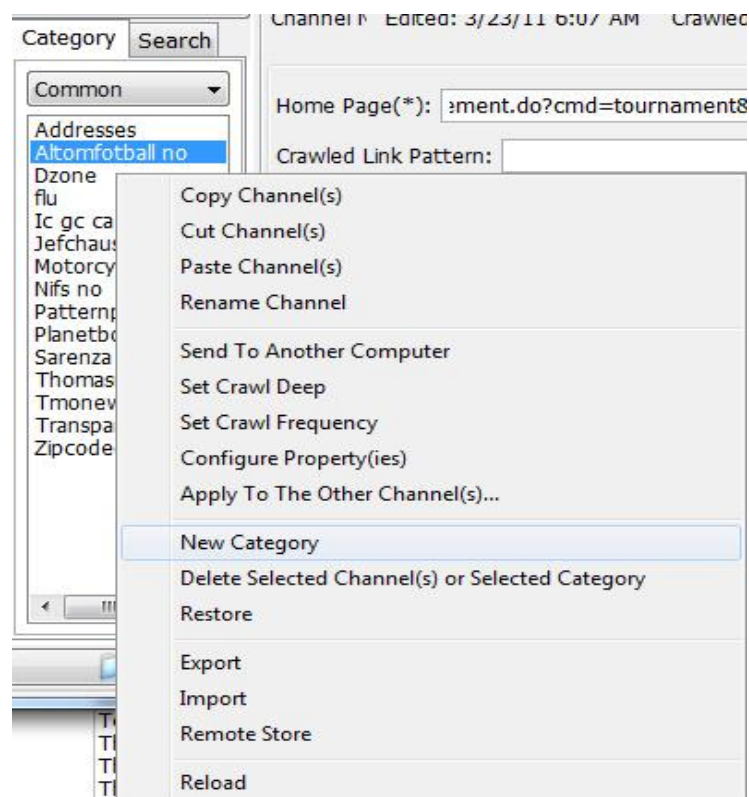
- **Create new channel** is use to manage configuration of channel.
- On **Tool** menu click on **Create new channel** icon. The interface will look like this:



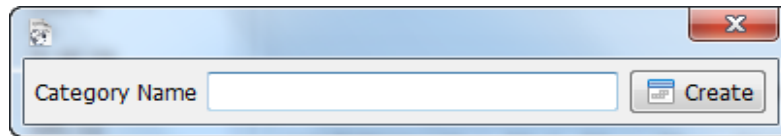
Title	Content
1	List of all chosen categories and channels
2	All information about channel, you can modify it in two mode: Basic and Advance.

### 3.2 Make new category

- This function to create a category. **Channels** will be put in **Category**. It's very useful and easy to manage, monitor and browse content.
- In **Category**, right click and select **New category**.



- A new dialog box will ask you enter a category name.



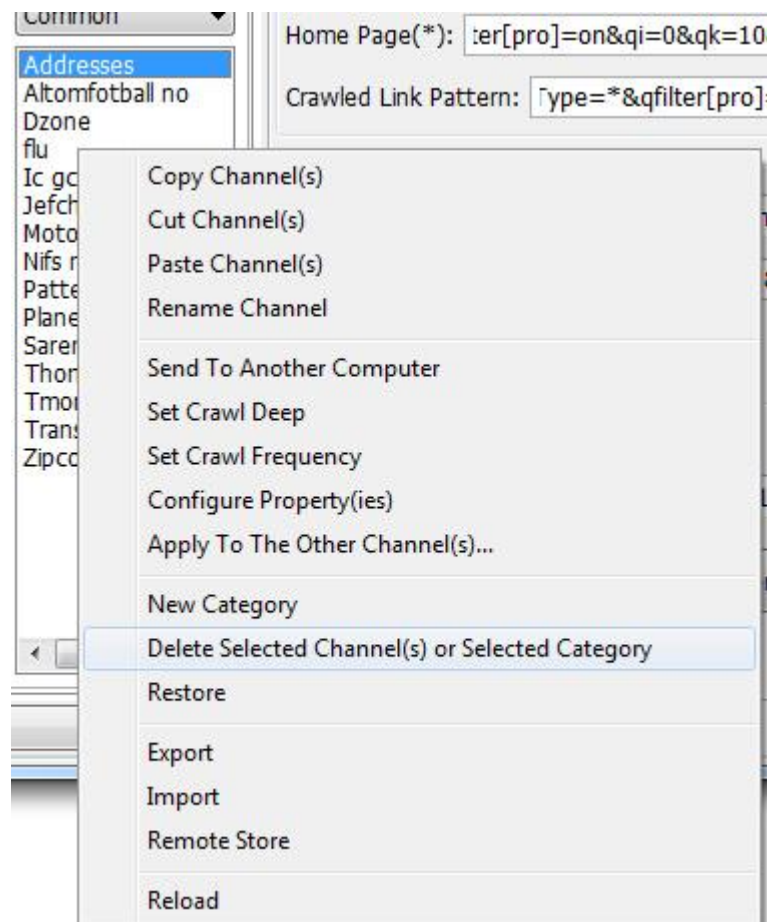
- After entering the name, then click **Create** to make a new category. Otherwise, close this dialog.

*Note:*

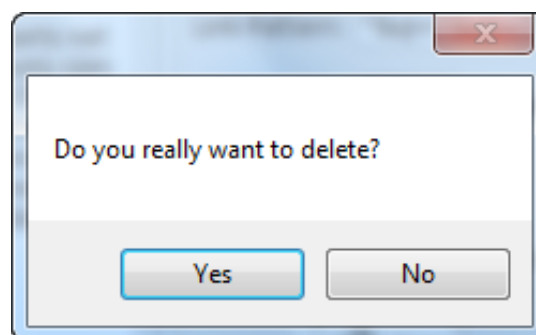
- Note: Name of category is not allowed including special characters: (-), (,), (.), (<), (>), (\*), (?), (/), (|)...

### 3.3 Remove a category or channel

- This function to delete categories or channel.
- In category list, right click on channel that you want to delete and select **Delete selected channel(s) or selected category**.



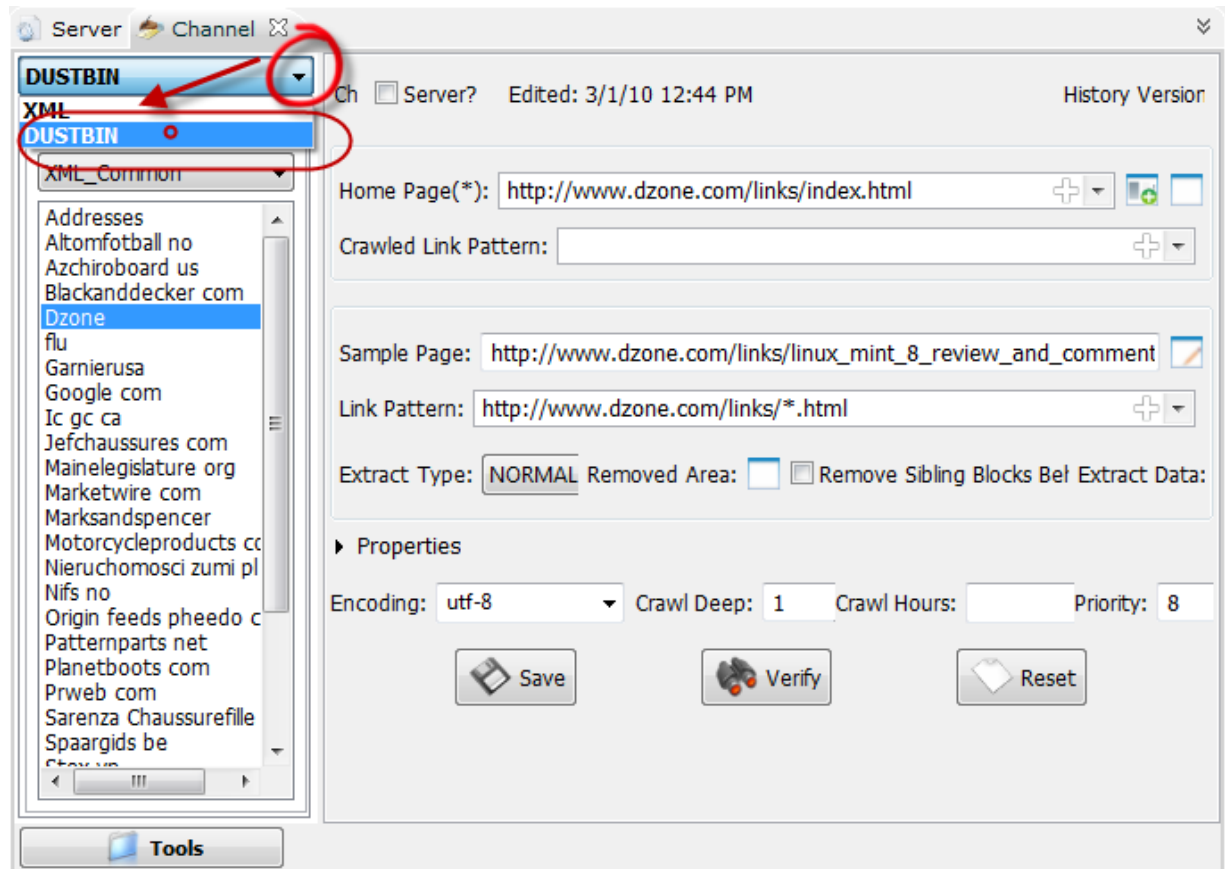
- A confirmation dialog box will appear:



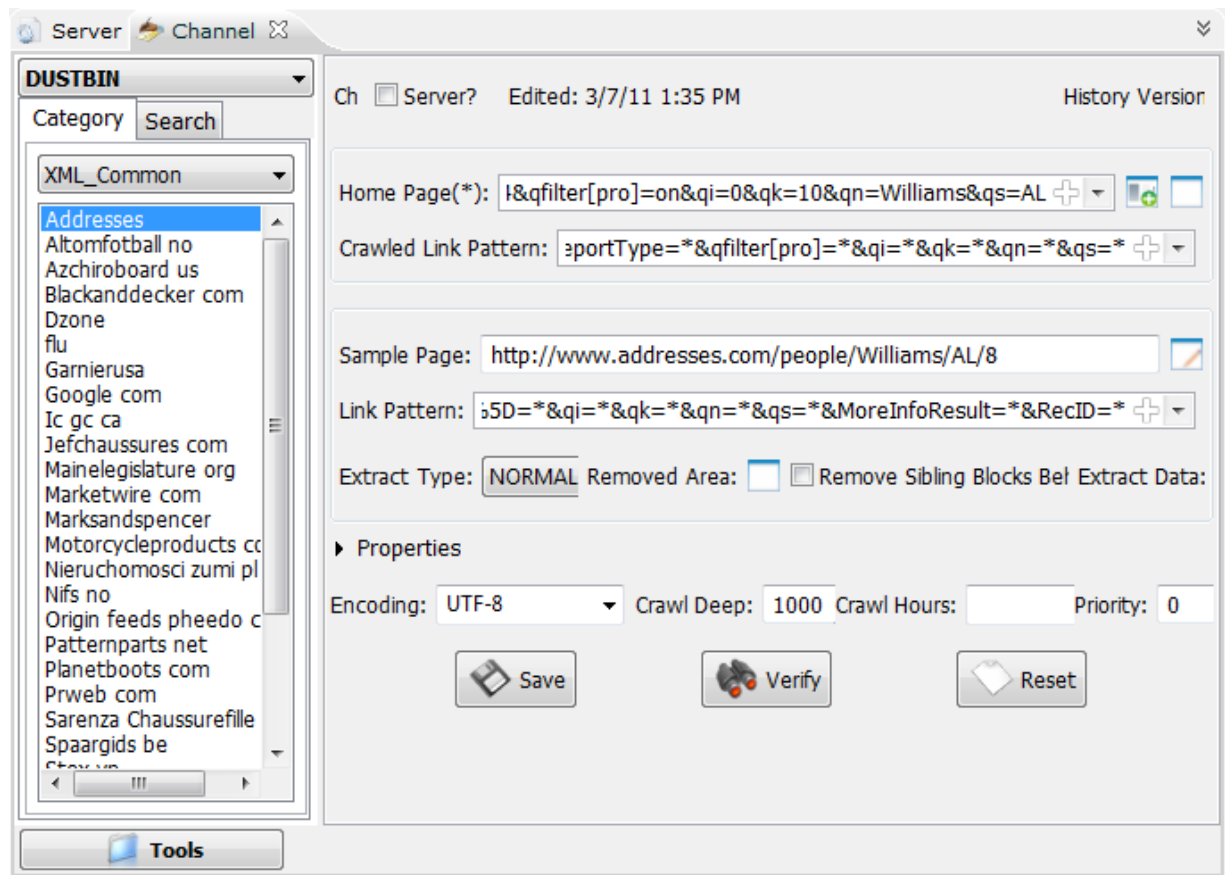
- Click **yes** to remove or **No** to cancel. After removing, channels or categories will move to **Dustbin**.

### 3.4 View current and removed categories

- This function help managers check deleted threads and restore configurations of flows or categories that users have deleted.
- In **Creating data flow section**, at the menu left of program interface, users click on the arrow (red areas are delineated as shown below), then select the **DUSTBIN**.



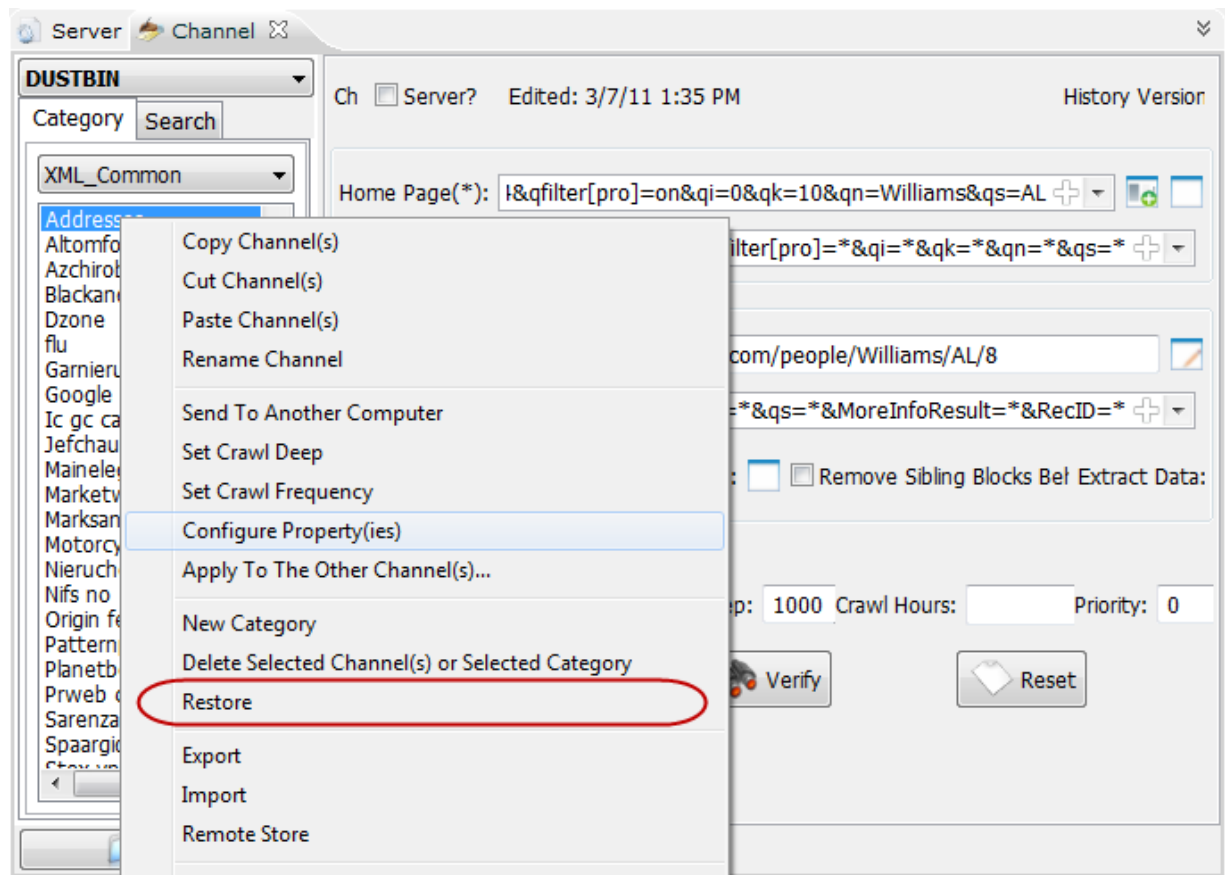
- User will see just below the **DUSTBIN** categories. **DUSTBIN** category contains categories and flows deleted from configuration repository.



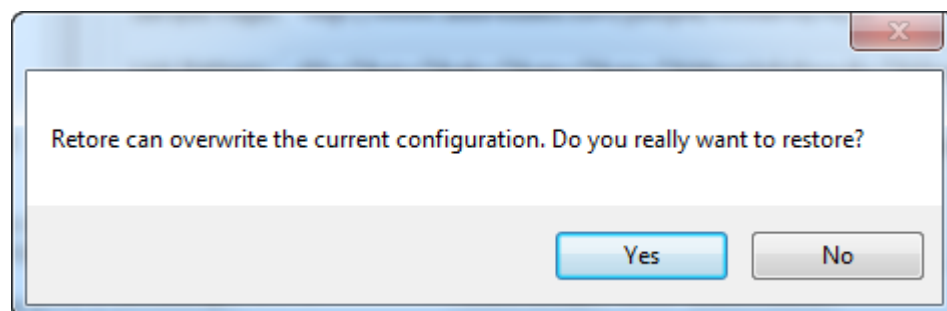
### 3.5 Restore deleted categories or data flows

- This feature helps the administrator can restore deleted categories or flows in the past.
- On **View thread or deleted categories** user selects flow or category to restore, then right-click on them. The application interface displayed as shown below:





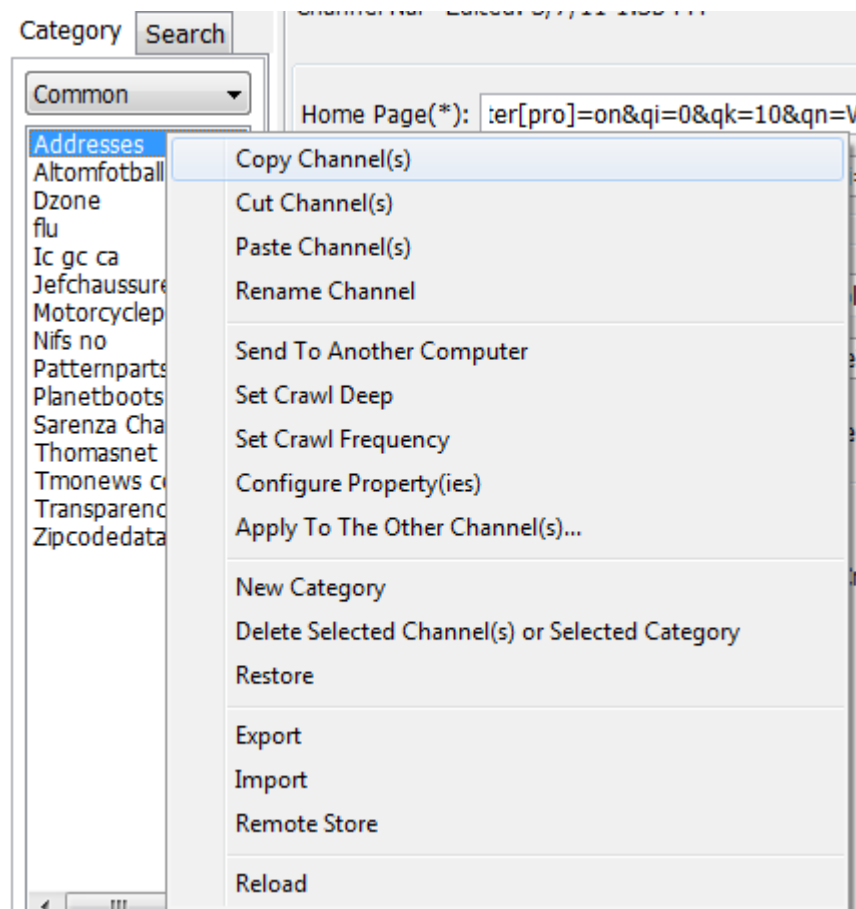
- Users select the **Restore** function, on the interface of program, a warning for getting user's confirm will appears to make sure that the action of restoring is decided.



- User have to press on **Yes** button to restore configuration or **No** to cancel the action..

### 3.6 Copy channel

- Copy channel function to allow user copy configuration of channel to other **Category** or **Data zone**. User does NOT need to fill the same configuration of channel or move a channel to other category.
- In **Category** list, then right click to select channel that you want to copy.



- Select **Copy channels**, after that, system will copy all configurations into memory.

### 3.7 Cut channel

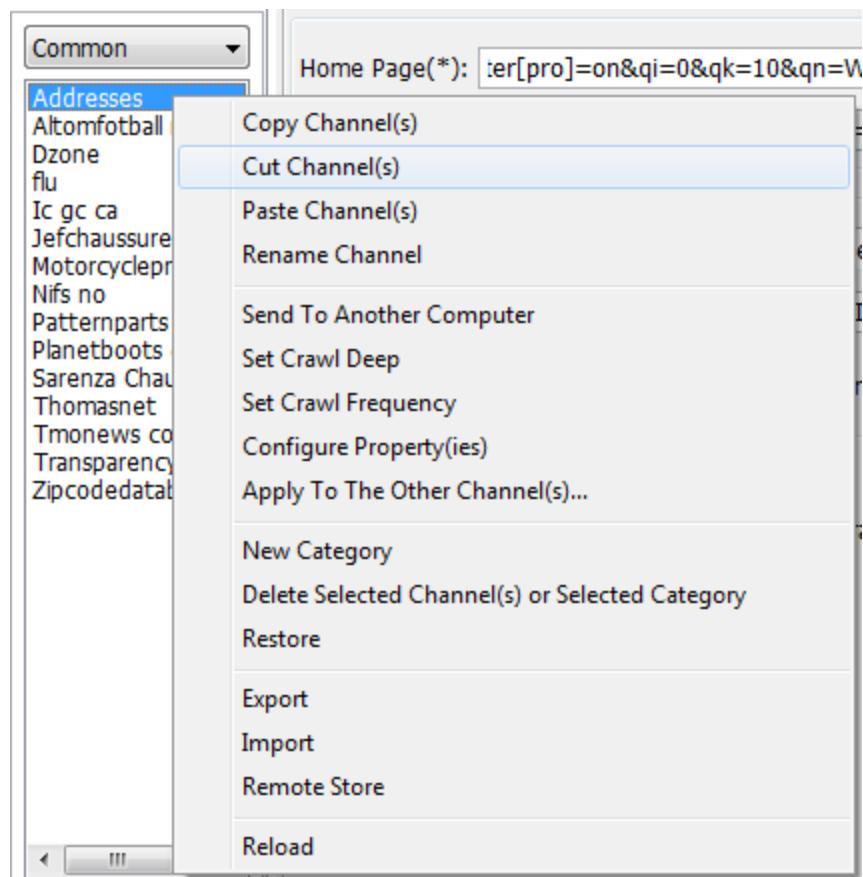
- **Cut channel** to allow user to copy and remove channels to other category.
- Select a channel and right click, then select **Cut channel(s)**.

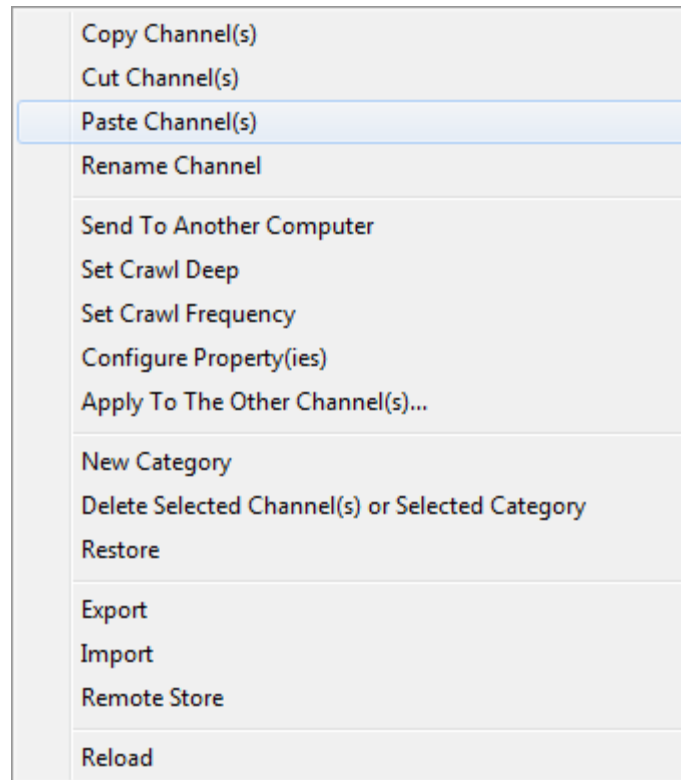
- ✖ User goes to destination category, and right clicks, then chooses **Paste channel(s)**.

**Note:** this function can move multiple channels to other category.

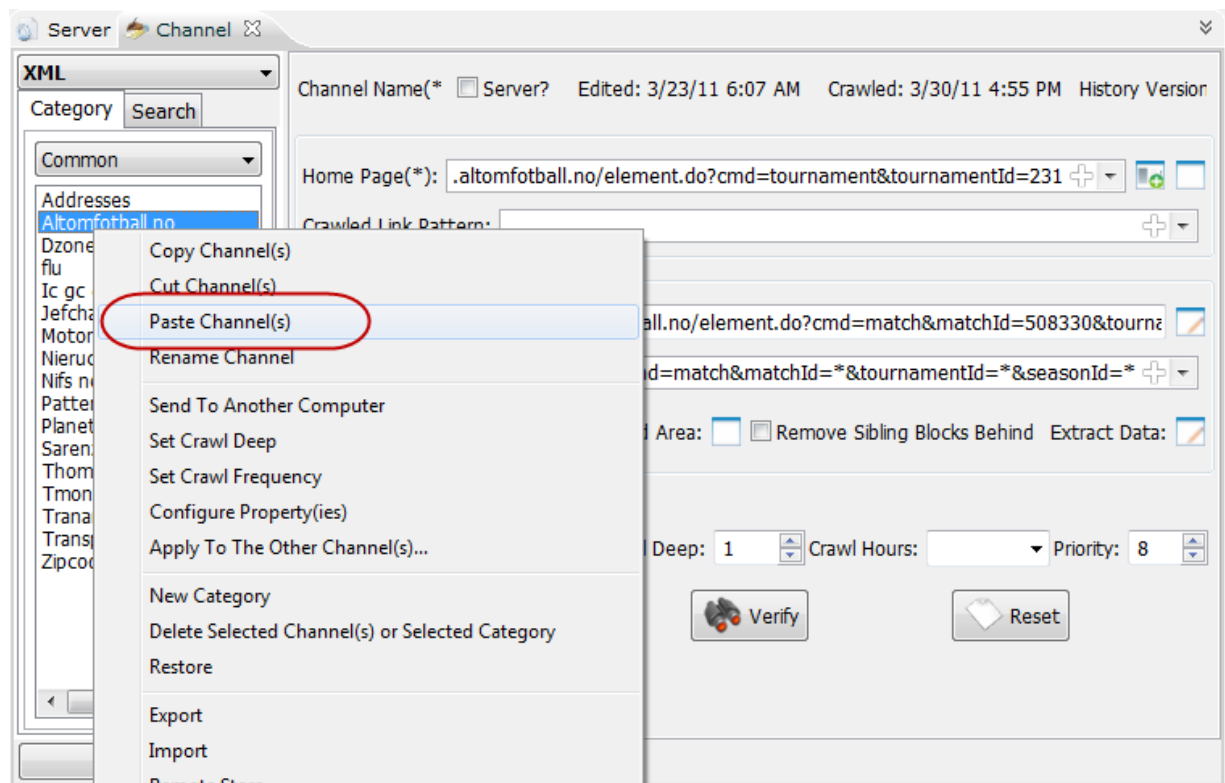
### 3.8 Paste channel

- ✖ **Paste channel** to allow user paste all configuration of channel to destination category.





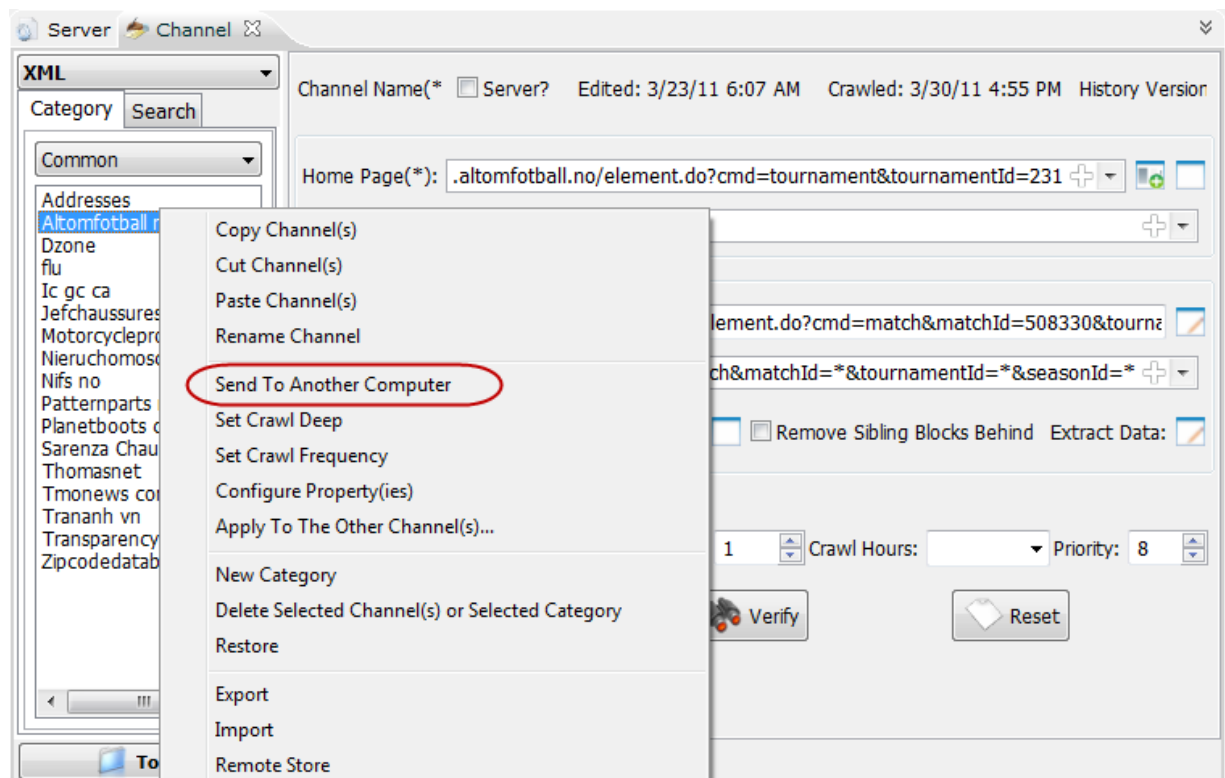
- User goes to destination category, and right click, then choose **Paste channel(s)**.



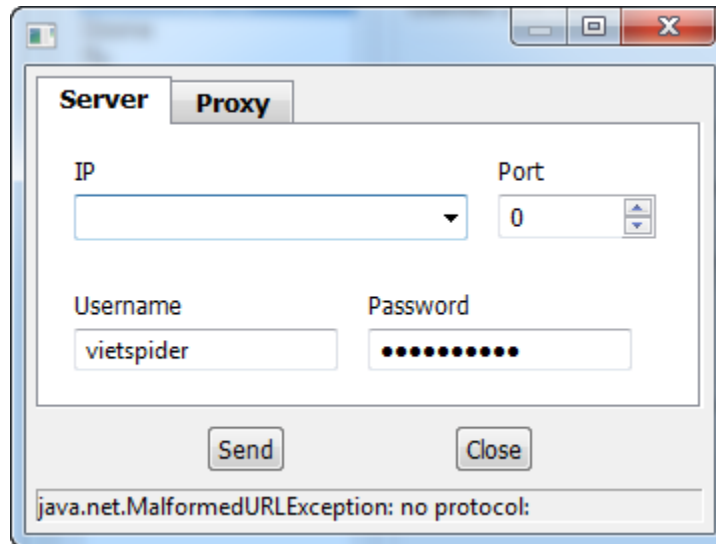
- User select **Paste channel(s)**, new channel will create or overwrite a exist channel.

### 3.9 Send channel or categories to other computer.

- **Send to another computer** to allow user to exchange configuration of channels or categories to other system.
- In category, let right click on channels and then choose **Send To Another Computer**



- After selecting **Send To Another Computer** , a dialog box will appear:



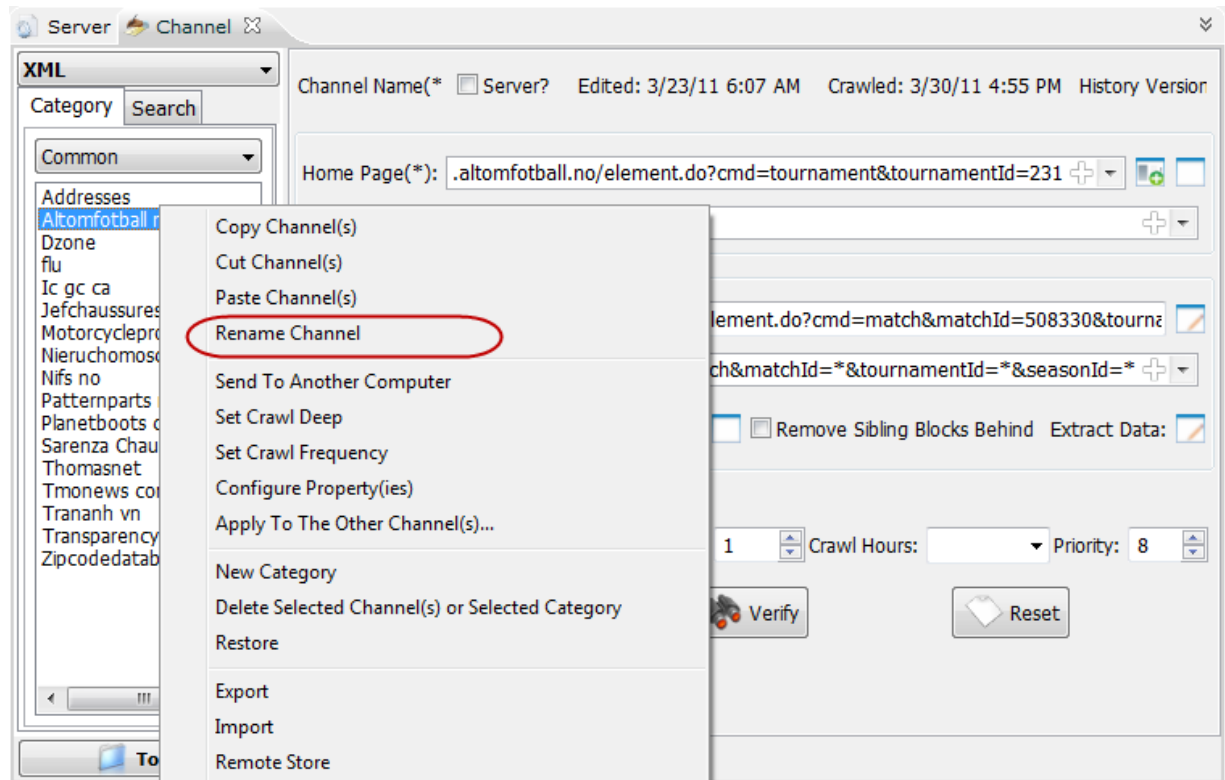
- ✖ The dialog box will ask you to enter configuration, including: Destination IP address, Port, User name and Password

*Note:* + User can select multiple categories

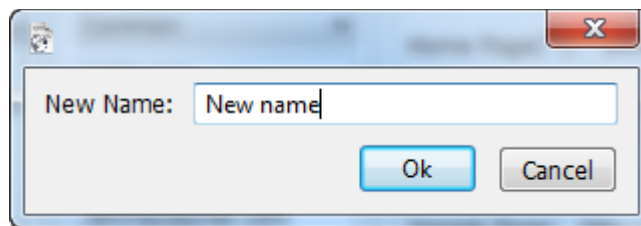
- ✖ After finish fill the form, click **Send** to send configuration to chosen computer, otherwise, close the dialog box to cancel.
- ✖ Look at category list of destination computer, you will see the sent channel as same as source computer.

### 3.10 Rename

- Rename function allow user can change name of channel or category. Note: no change in configuration of channel.
- In category, let right click on channel, and then select **Rename channel**.



- When click function **Rename channel**, there is a dialog box allowing to change name of channel or category.



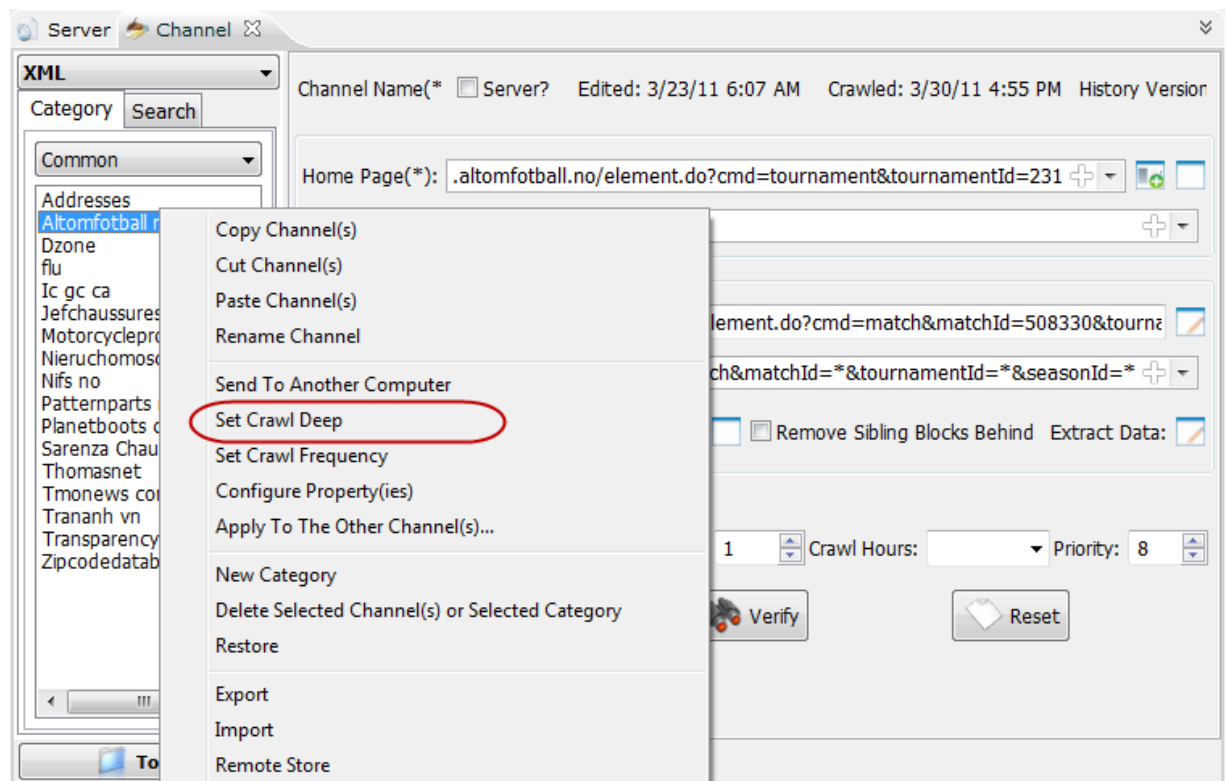
- If you really want to change name, click **OK**, otherwise, close the dialog box, the VietSpider will NOT do anything.

### 3.11 Crawl deep

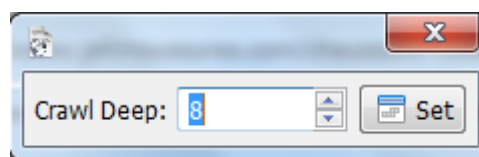
**Set Crawl deep** function allows system crawling content in deep.

For example: If homepage of website is level 0, subpage is level 1 count from homepage. The next is level 2, 3... Therefore, crawl can get more data when it go in deep, and it takes more time.

- Let right click in category list on the left menu:



- User change value of **Set Crawl deep**, this value is low or high depends on user purpose.



*Note:* + if the configuration information of the stream only **Homepage** approved without any **type of site**, users should fix **Crawl Deep** is 1, so program scans only the surface of the site rather than take more time to scan deeper.

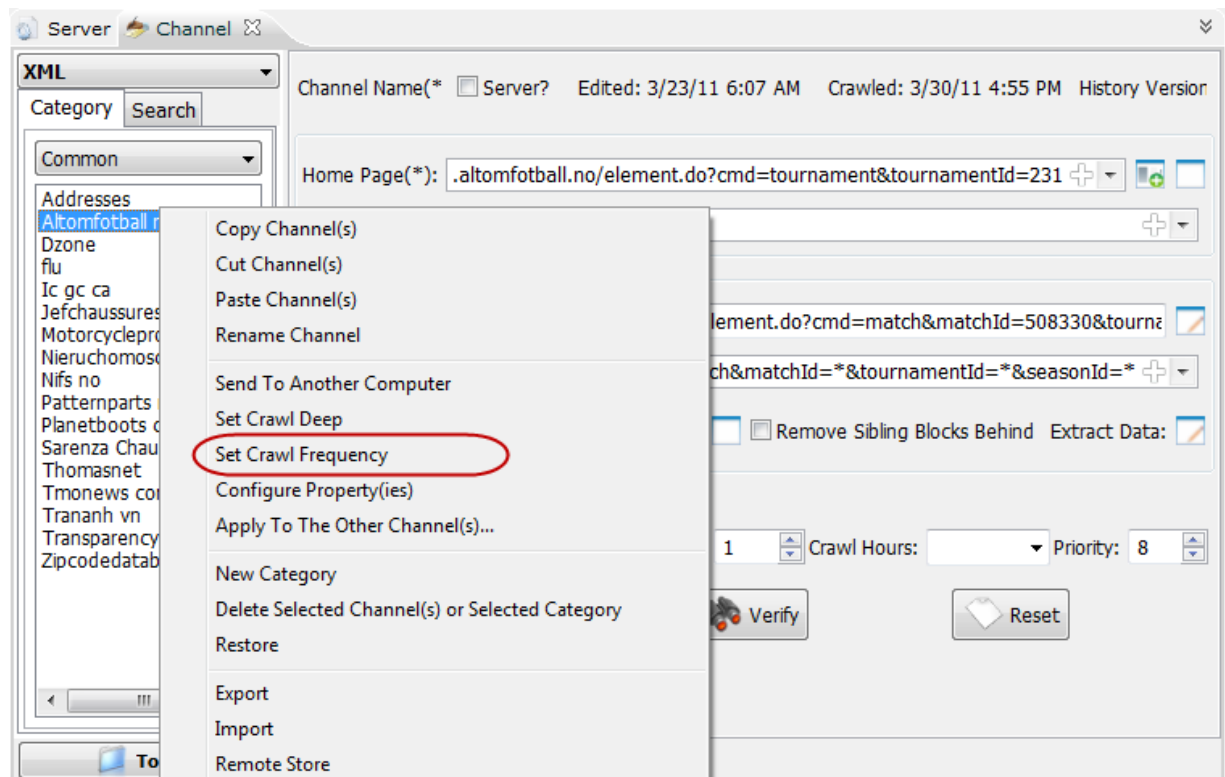


+ If the configuration of information is **form pages**, in default, the browser will scan from two. Higher levels set by the user.

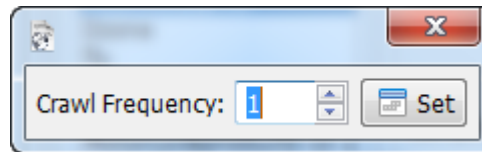
+ If the website has more information is required to download. Then estimate **Crawl deep** to scan.

### 3.12 Set crawl frequency

- ✱ This feature helps determine system load frequency scan of the site for a certain period. In **Priority scan**, the default, provisions is corresponding to 30 minutes. Value for the priority is set in the **System** Configuration instructions in the next section of this document. If users set, the priority crawl is 48 = 1 day, after a day on the program will automatically scan the website to get the web data.
- In the list of categories, user right click and then set **Priority scan**



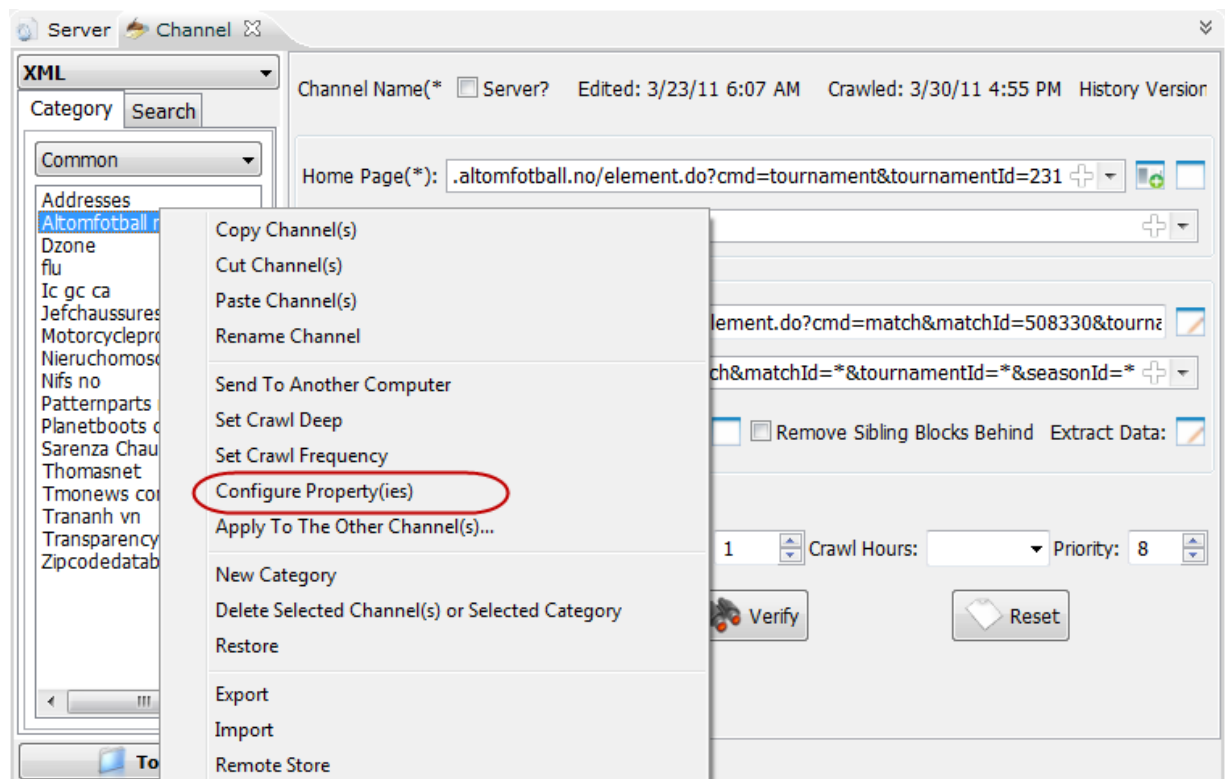
- After the user selects set **Crawl Frequency**, the program will display a dialog that allows user to select a scan priority as shown below:



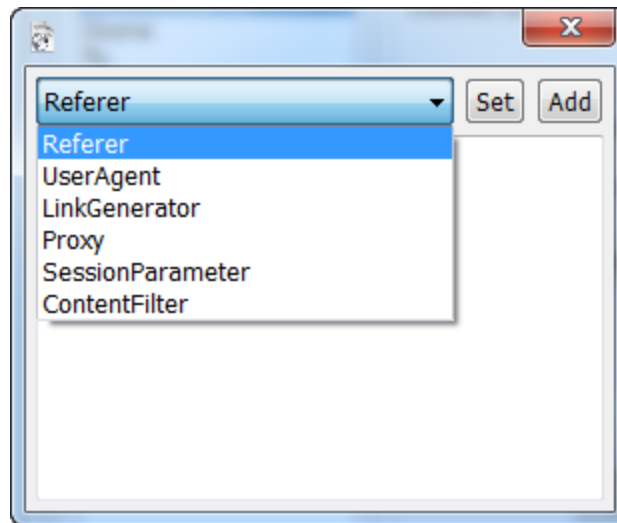
- User can change the value that user want to set for Channel. After you have identified the **Crawl frequency** users click the **Set** button to save.
- User can set the **frequency level** for a channel scan or a few channels in the same category by pressing the Ctrl key or Shift key to select multiple channels.

### 3.13 Configure properties.

- Properties are to provide additional data in the configuration of channel. For websites with special parameters, then we must add some attributes. These attributes support download programs work channel more efficiently.
- In the list of categories, the user right click, after the menu appears, select **Set property**.



- After the user clicks on **Set property**, a window with the attributes are displayed as shown below



- ✖ Depending on the system, users have different lists of attributes. Current system has the basic attributes are: ContentFilter, Login, Referer, JSonclickPatten, LinkGenerator, HomepageTemplate, Proxy, SesionParameter, Address.
- After completing all the attributes, click **Set** to save the set value of the attribute, click **Add** to add to the collective value of the property.

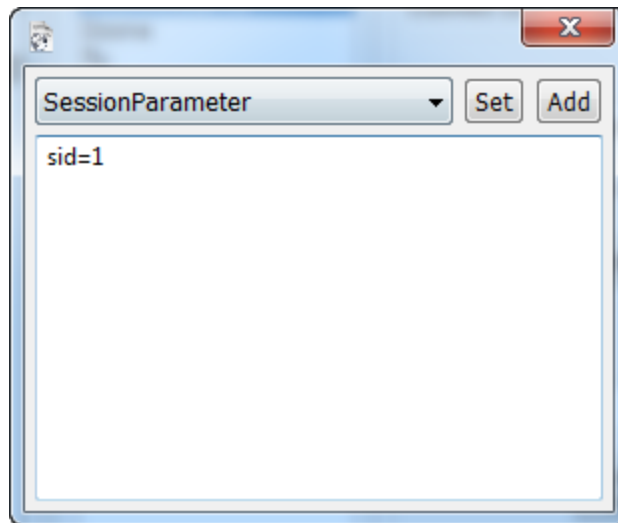
*Note: Some specific examples of the channel and adding special attributes*

+ ContentFilter allowed downloading content filtering by word or phrase. Download program checks to see content that contains keywords and then filtered or logged. Support operations OR (or) with a comma (,) and bitwise (and) with the plus sign (+).

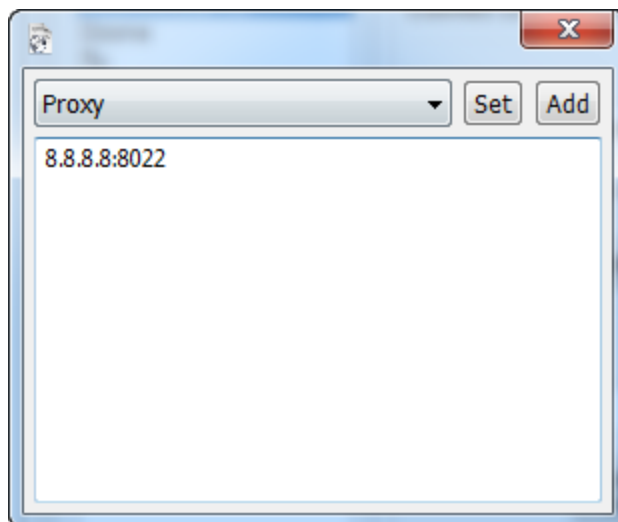
+ Some special website to access links through JavaScript functions. JSonclickPattern belonging to convert these values into the link function.

+ Some Forums add **Session parameter** in the link browser as: s =, = sid, PHPSESSID =, SID =... as shown below. Adding programs **SesionParameter**

help avoid reloading all the data mined by the value of the session id changes to the program download page mistook the new data.



+ If you must access the site via proxy, the proxy adds value to this attribute as shown below:



## 1. Login

Some websites require login to access data or content. In order to help VietSpider could crawl these pages, you must have an account. Then, you have to add these accounts to login part of VietSpider with the following structure:

*Address of login page*

*Username: password*

Ex: If I want to access **java.net** website, I have to add information to login part of VietSpider like:

```
https://www.dev.java.net/servlets/TLogin
thuannnd:zhmllj
```

## 2. Referrer

Some websites do not allow a direct access, they require user go to home page before returning to content page. For example, when you access a link, it gives a message “Click here to view your required page”. You should fill any link of that website into **Referrer** part to help VietSpider can crawl data.

## 3. User Agent

(Faking browser, Google or Yahoo bot)

If you want to fake Google bot, you have to fill Google keyword into **User Agent** part and similarly for yahoo also.

## 4. LinkGenerator

A mechanism that generates links. This property supports business customer.

Page indexing:

Index

```
http://ddth.vn/showthread.php?t=[index]
```

1->1000

Browsing by word:

Word

```
http://ddth.vn/showthread.php?t=[word]
```

Create link by JavaScript function:

Onclick

```
fn_LoadEntry_Lst_ByEntryId('*', '*')
```

```
http://chuyendong.gdc.vn/Controls/CDG/NewsCat/XML_Entry_List.aspx?ic={1}&p={2}
```

Homepage template configuration:

Homepage

`http://*.blog.360.yahoo.com/blog-*.cq=1`

`http://*.blog.360.yahoo.com/blog-*`

Search homepage and create new channel

Homepage

BLOG.BLOG VIET NAM.360 YAHOO

*Some example of LinkGenerator:*

Word

`http://ddth.vn/showthread.php?t=[word]`

Homepage

`http://*.blog.360.yahoo.com/blog-*.cq=1`

`http://*.blog.360.yahoo.com/blog-*`

Update document from JavaScript

`type org.vietspider.link.generator.UpdateDocumentGenerator`

`http://www.google.com/maps?f=q&source=s_q&output=*&hl=*&geocode=&q=*&btnG=*&vps=*&jsv=*&sll=*&sspn=*&abauth=*&absince=*`

Onload

`type org.vietspider.crawl.link.generator.FunctionFormGenerator`

`javascript: __doPostBack('*', '')`

`__EVENTTARGET={ 1 }`

## 5. Proxy

Set proxy for accessing the site from being blocked. Structure: *proxy-host: proxy-port*

Ex: Set proxy to access the page **talawas.org**, I should fill like below:

*203.178.133.10:3124*

#### 6. Session parameter

Some forums written by PHP language, they use parameter to save session id. For example, **ddth.com** has the link:

<http://ddth.com/showthread.php?s=fdskfjiu432432&t=123123>

Value: s= fdskfjiu432432 is session id. This value (fdskfjiu432432) is changed on each accessing times. In order to avoid reloading page of VietSpider, we set value for Session Parameter part. For above example, we set:

s=

So, value of parameter s = will be passed when VietSpider monitoring link.

#### 7. Content Filter

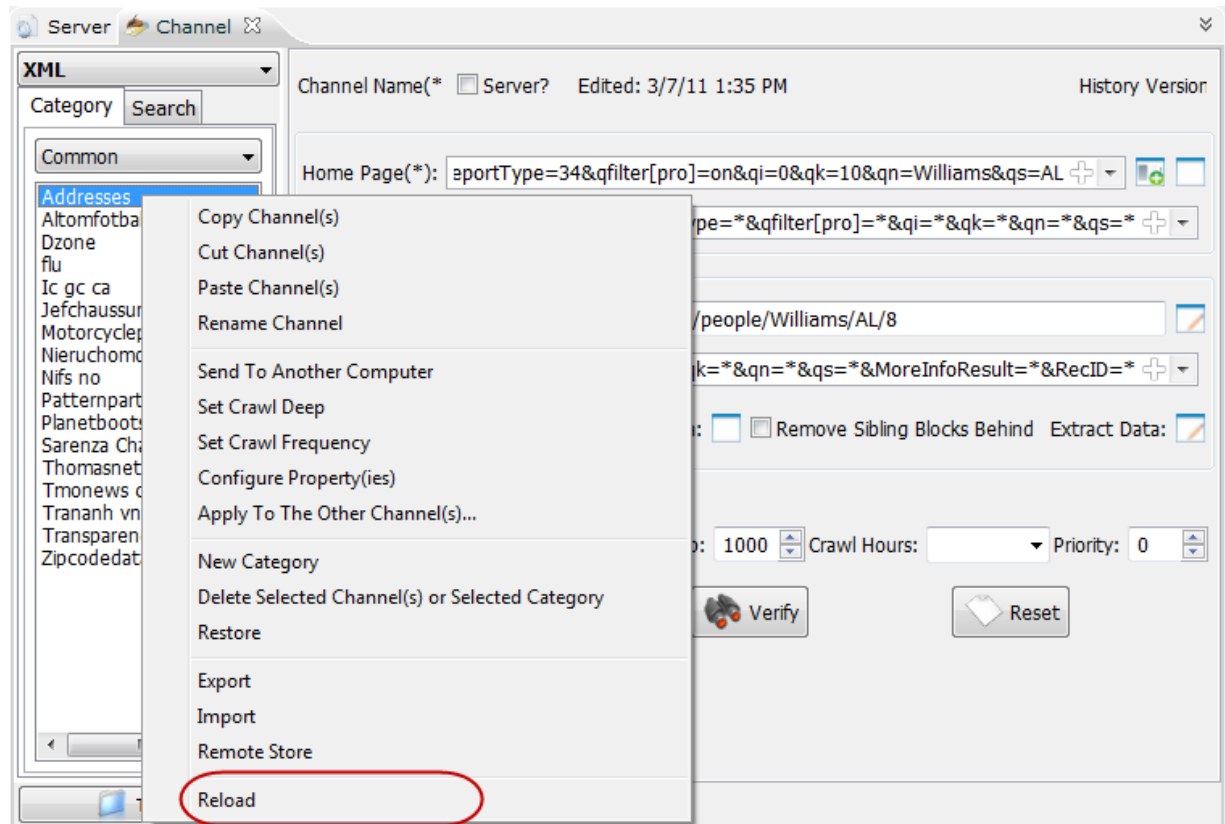
Content filter, this means VietSpider only save content containing corresponded keyword. OR operator replaced by the mark “,” and **AND** replaced by the mark “+”.

Ex: If I type **flood, rain, storm**

VietSpider will only save the content from the "flood" or "rain" or "storm".

### 3.14 Reload.

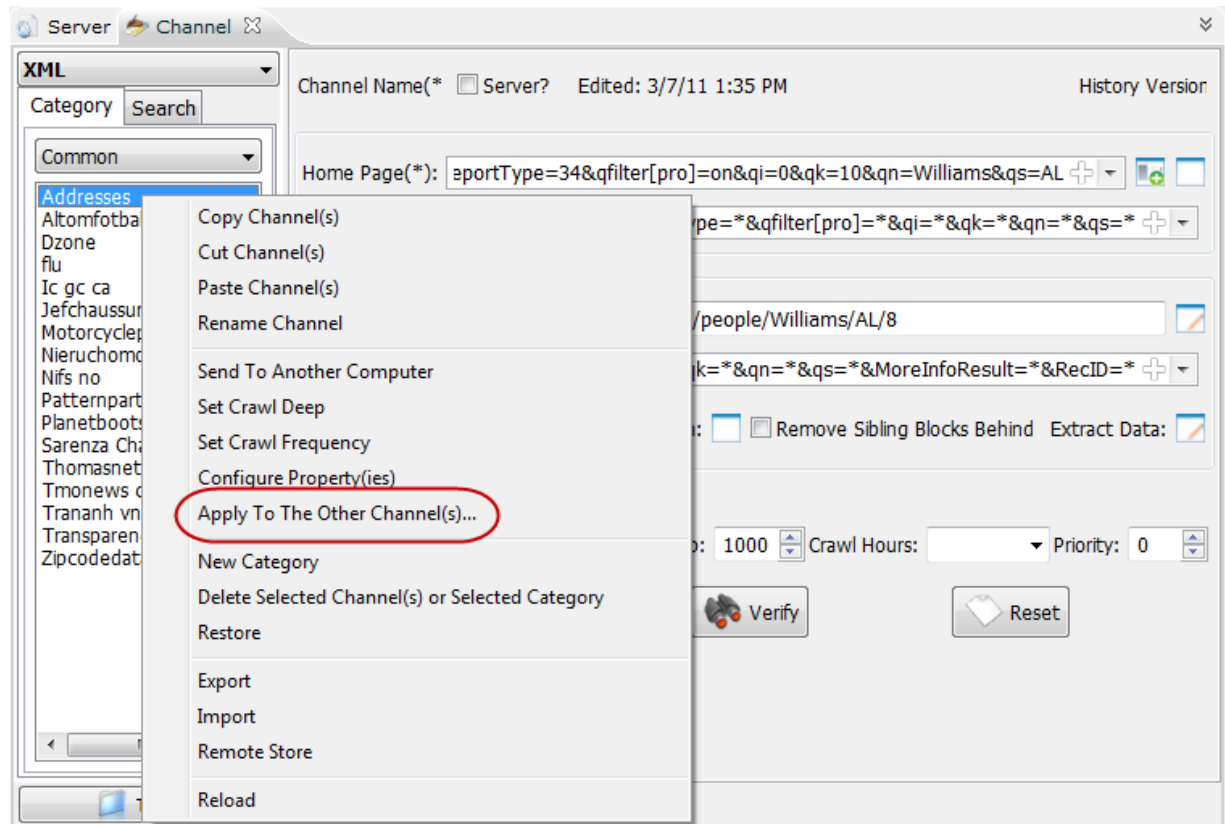
- This function is to update those systems portfolio flows have been configured in categories or new categories were created within the data.
- Select **Reload** in the right of the menu configuration list archive



### 3.15 Apply to other channels.

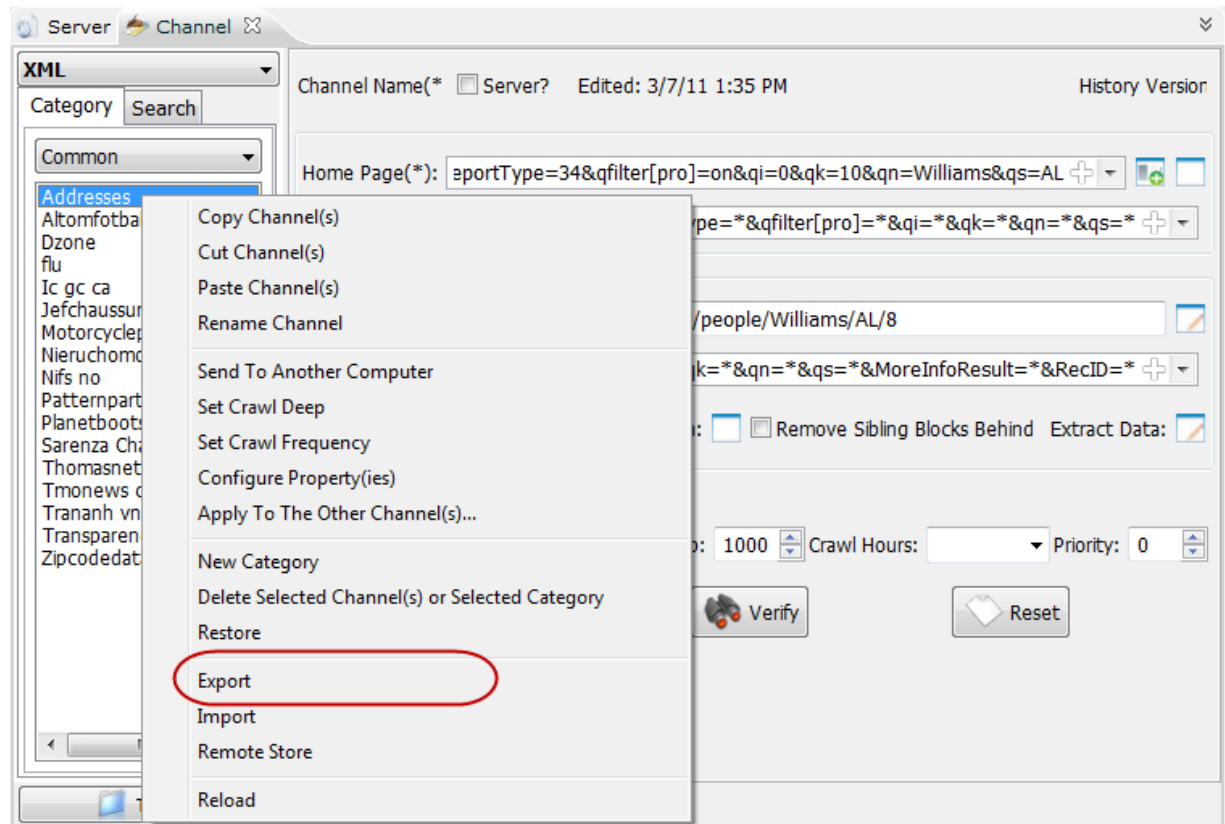
- ✱ Save configuration similar to the channel of the same name in stock, such person uses both the Culture and Sports Section of the newspaper BBC named in two different configurations. When the dissection in the modified configuration of the newspaper BBC Culture and users also want to update this peel for the Sports category in a different configuration similar name (such as channel configuration also named BBC is sufficient and Sports categories), then the user can use this function.



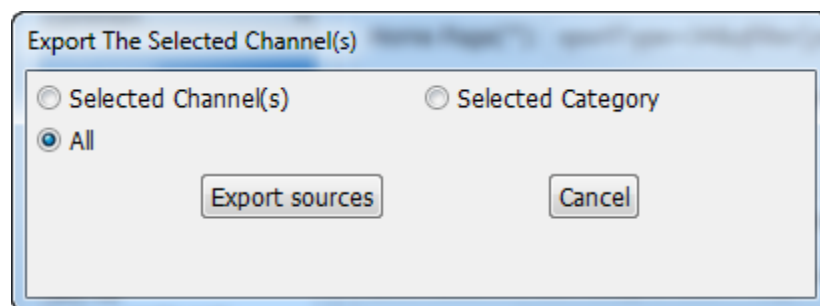


### 3.16 Export

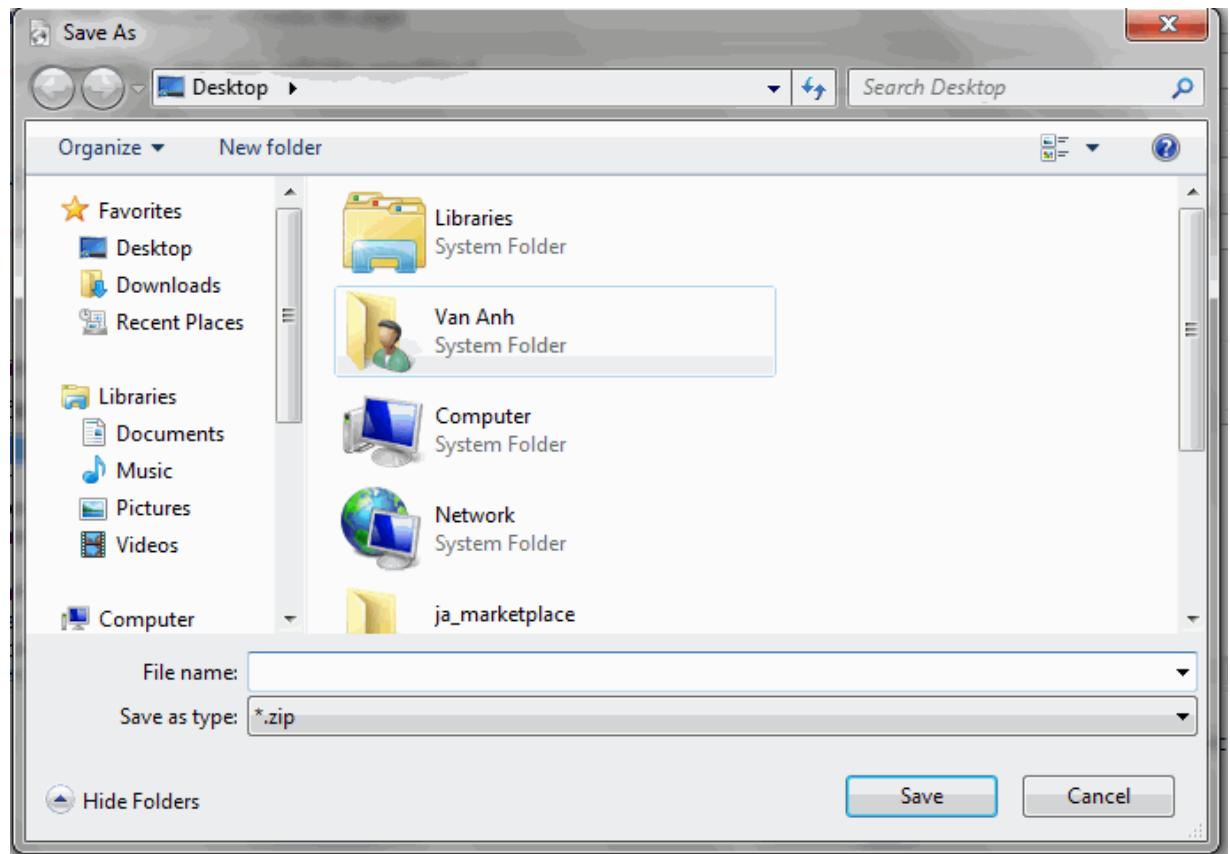
- The program allows users to produce a channel or multiple channels, one or more categories, or all categories of channels a compressed file. zip. This feature enables users to exchange configuration between multiple machines or can store the configuration of the threads in your machine.
- In the list of categories, the user right clicks. After displaying the menu, I select **Export**.



- When the program appears in a new window with more options for user selection as shown below: (The stream is chosen: the jets that are user selected; Section was selected as the categories that user is selected, the selected team is heading that the user data is selected, the entire amount in stock is all the channels in various categories included in the group are in the computer data of users in current time).



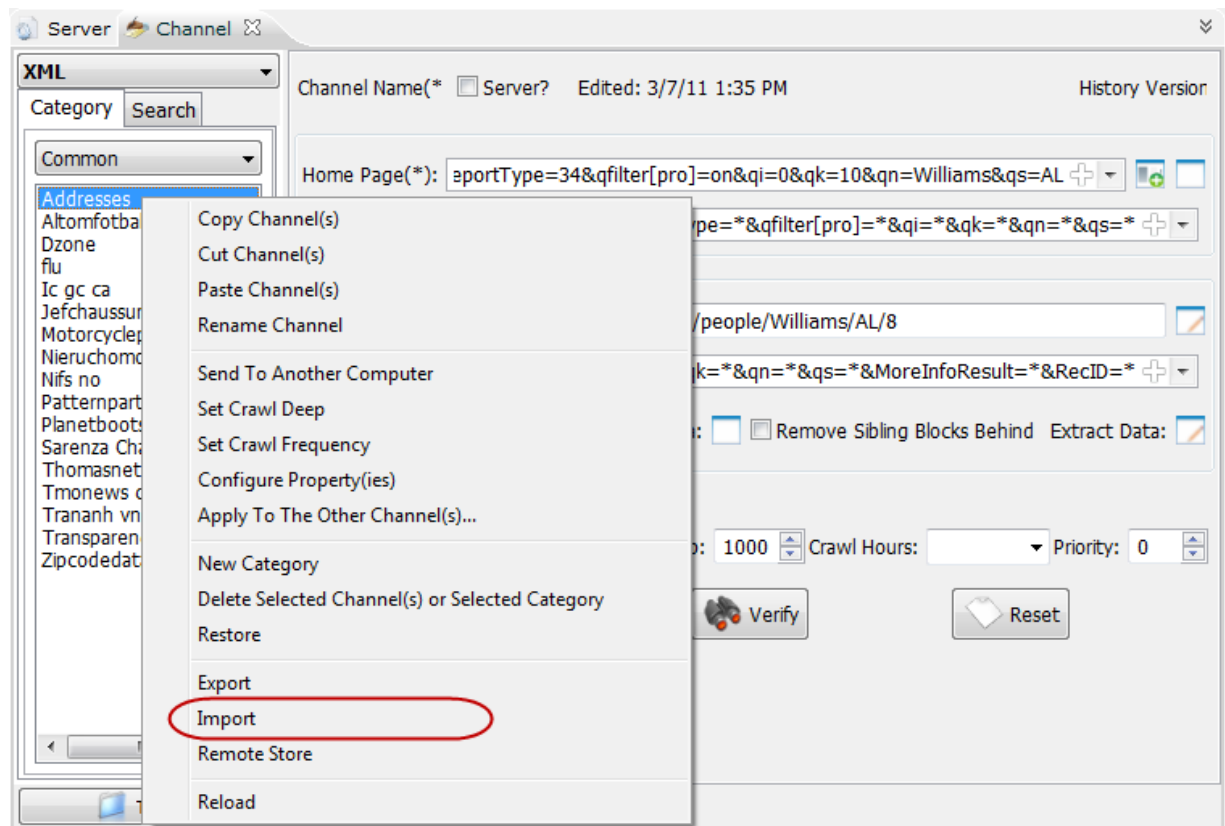
- After choosing, the options in accordance with the **requirements of their users** click **Select file** then program interface as shown if no click **Cancel**.



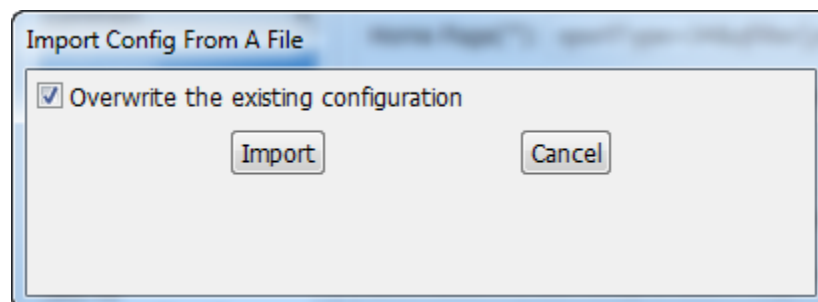
- The user name and choose where the files and then press the **Save** button, if not pressing the **Cancel** button.

### 3.17 Import

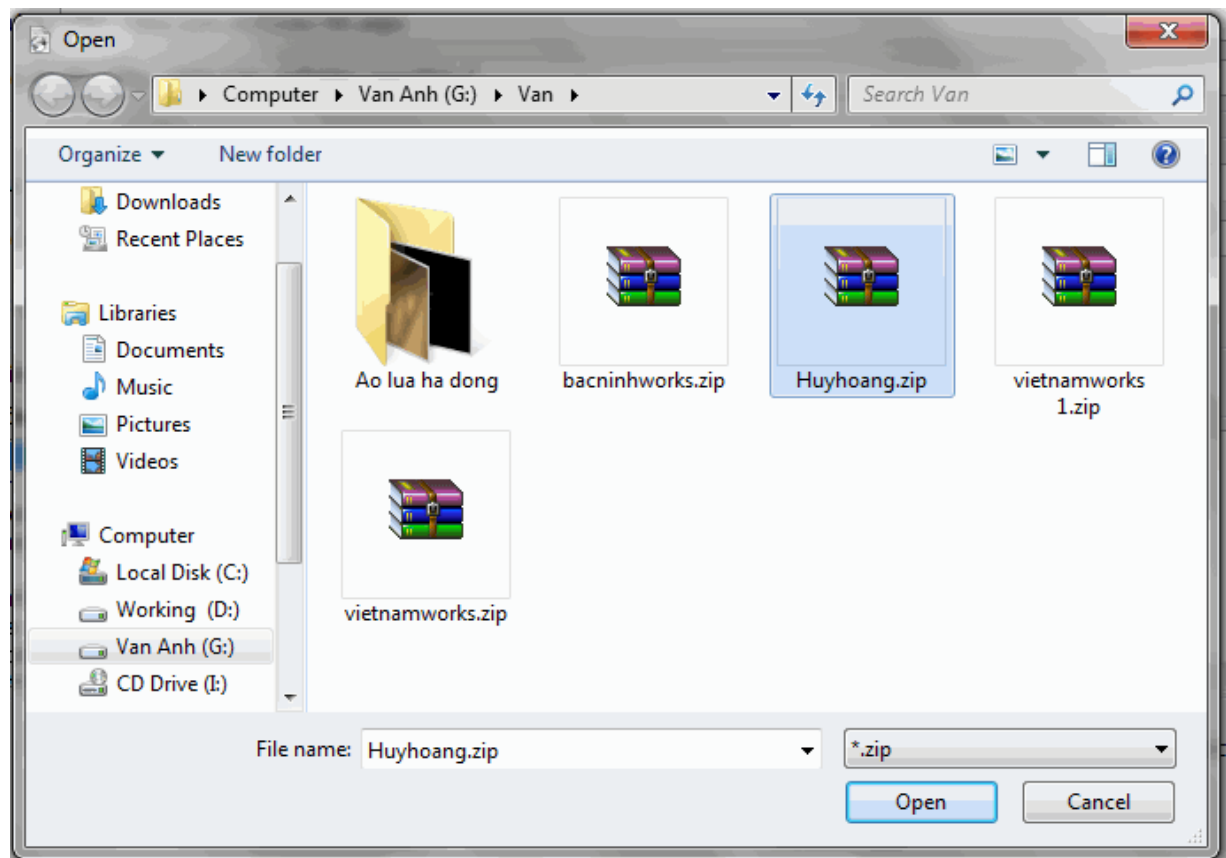
- The program allows users can import from a file \*.zip from one machine to another. This feature lets users help update the new configuration in the machine quickly.
- In the list of categories, the user right clicks. After displaying the menu, we choose **Import** from file \*.zip.



- When the program a window appears to notify users that new files will overwrite the configuration with the same name with the same group of data categories are included in the system, if user agree, let clicks the **Import** button, if not pressing the Cancel button.



- When the program displays a window that allows users to select the categories you need to import the file as shown below:



### 3.18 Create a new channel

#### The steps:

1. On the left list of the Channel tab, the user selects a section will contain ready to create channel.
2. In channel form, the user needs to enter the website address configured on the **Homepage**.
  - By default, when users input the **Homepage**, the program will automatically generate the domain name of the site to channel. Example: If the domain address or URL of the website is `http://siteit.vn/` the corresponding channel name of this website will be SiteIT.
  - Users can also manually set the channel name by entering the name that the user wants to put this channel.

Note: + Channel name, the user can manage with Latin.

+ Channel name cannot contain special characters (-), (.), (<), (>), (\*), (?), (\).....

3. Browse pages form a general pattern to download based program that goes deeper into the links in the website.

*Example:* + if users want to download the program go into each column, then the user should find a common denominator in order to provide value in the form of browser pages.

+ Channel GiaoDuc has URL:

<http://dantri.com.vn/c25/giaoduc-khuyenhoc.htm>

+ Channel *TheGioi* has URL:

<http://dantri.com.vn/c36/thegioi.htm>

+ So, general URL form is:

[http://dantri.com.vn/c\\*/\\*.htm](http://dantri.com.vn/c*/*.htm)

+ These *values are not fixed by replacing the asterisk (\*)*.

- ✖ Online sites have different layouts so can offer a wide assortment of sites approved for the program. Then download the program will compare the links have been then decided to go deep into it or not. Browser page is a form of value is not required, it is intended to optimize the program downloads program downloads and help focus on areas where data needs to take.

*Example:* + in <http://tintuonline.com.vn/vn/index.html> website has many categories such as: daily event, The World, Entertainment.... with the domain name or address different URL as follows:

- ✖ *The World:* <http://tintuonline.com.vn/vn/thegioi/index.html>

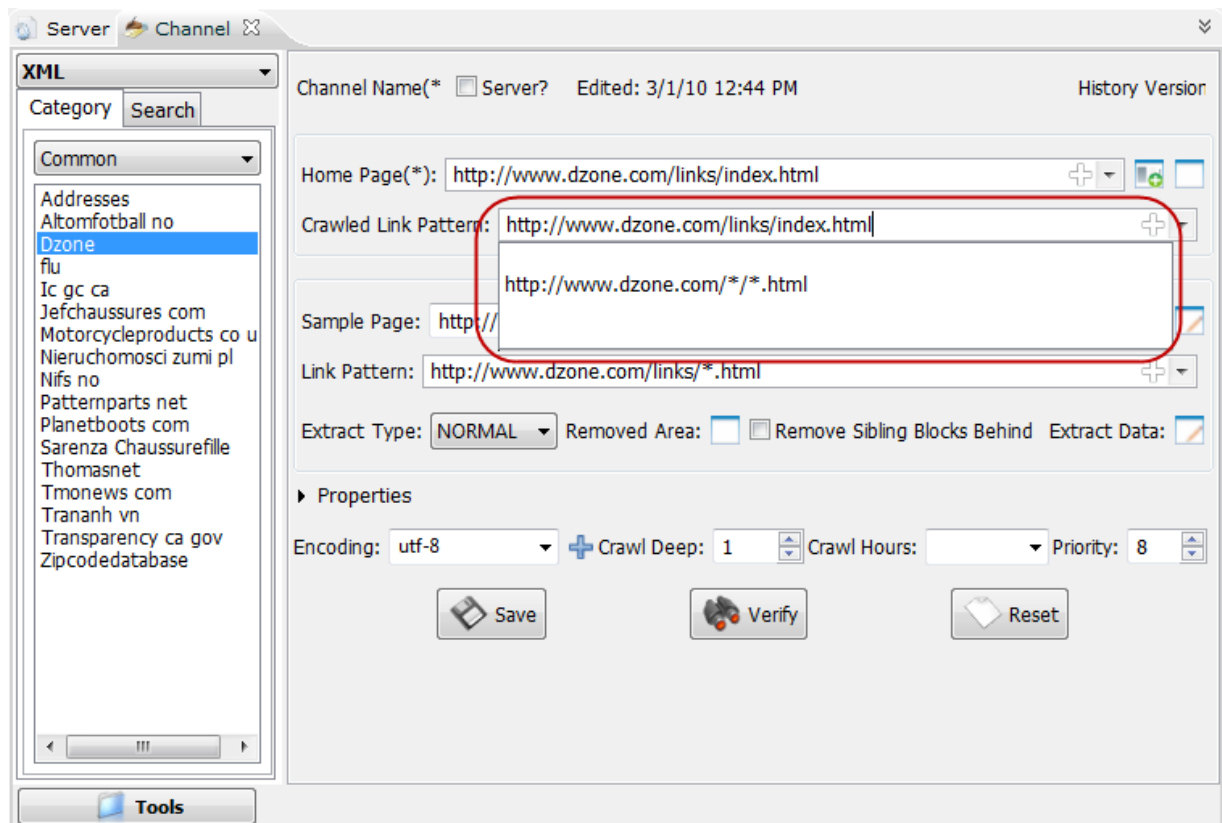
- *Economy:* <http://tintuonline.com.vn/vn/kinhte/index.html>

- Science: <http://tintuonline.com.vn/vn/khoahoc/index.html>

+ Three categories of Website have sections

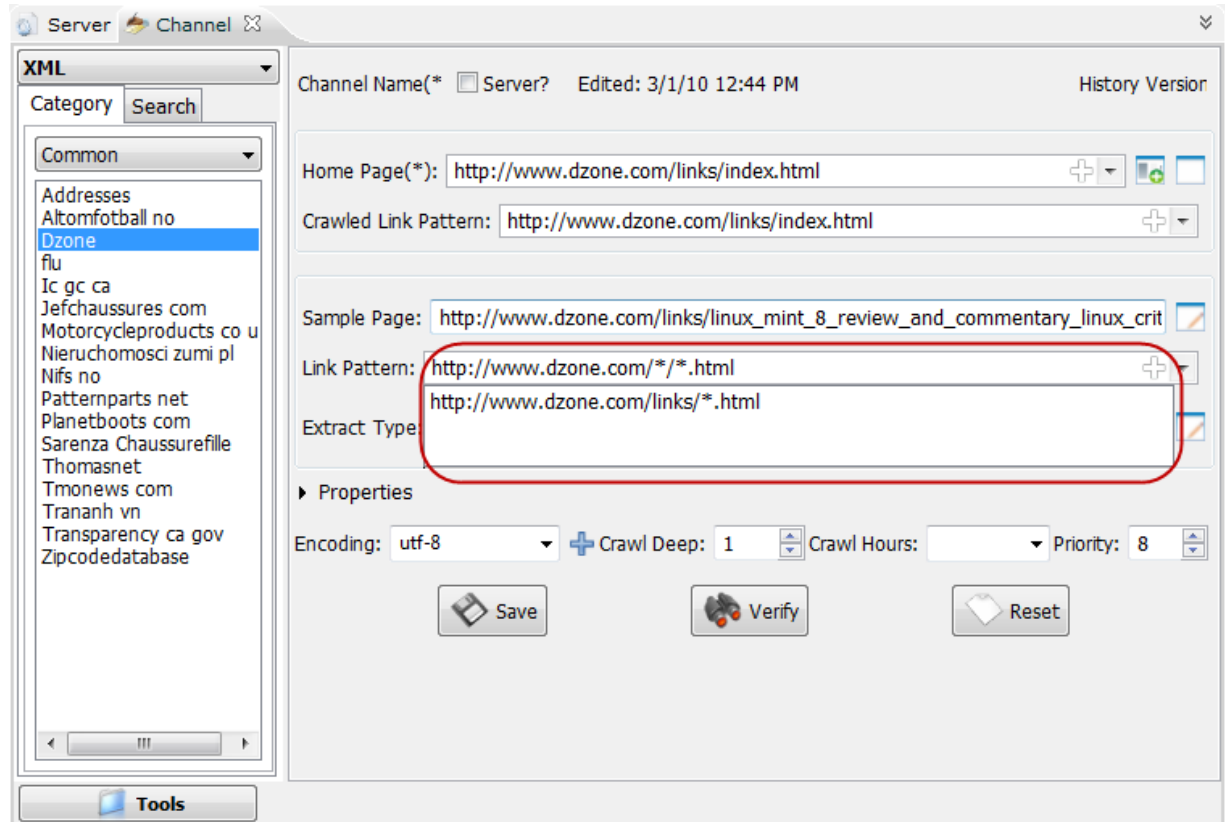
<http://tintuonline.com.vn/vn/> <http://tintuonline.com.vn/vn/index.html> is the same also different parts there are red letters thegioi, kinhte, khoaoc. When the user then replaces them with an asterisk (\*), such as a page is approved [http://tintuonline.com.vn/vn/\\* / index.html](http://tintuonline.com.vn/vn/* / index.html)

✱ The user clicks on **Type of page** to add to the list of approved sites as shown below:




#### 4. Users perform crawls a page template.

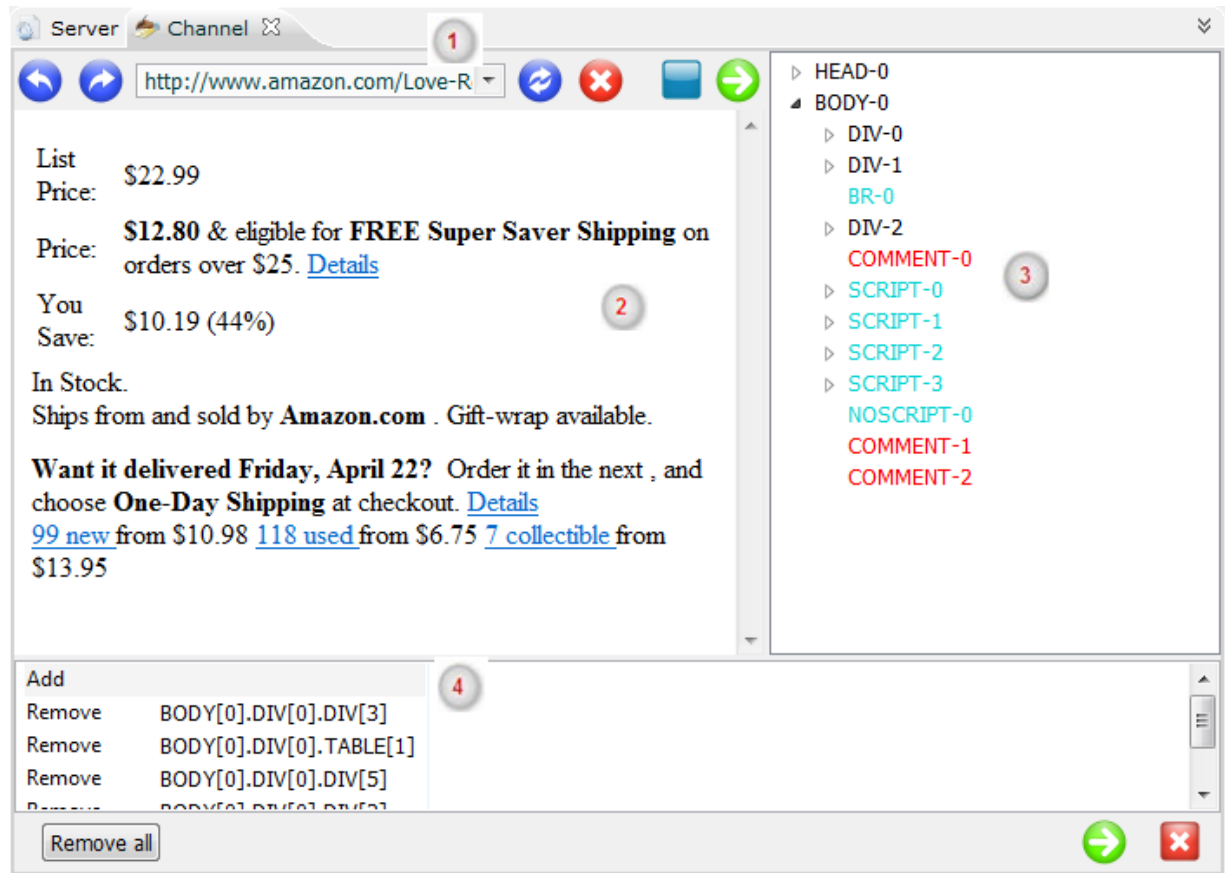
- ✖ **Template page** with similar purposes **browse pages**, but they are used to identify the exact program that is linked by link data. **Template page** is not required, optimize work load.
- ✖ The figure below illustrates the value and type of **Template page**.












## 5. Users select the area to get the web data.

- ✖ The user click on icon  in the bottom of **Crawl Area** to get data from web. The interface of **crawl area** is displayed as shown below:





Title	Content
	Template page URL.
	Displayed content.
	Tree structure of HTML page templates with the corresponding tags.
	List of selected blocks, shows the value of the block is approved.
	Display full page or not?

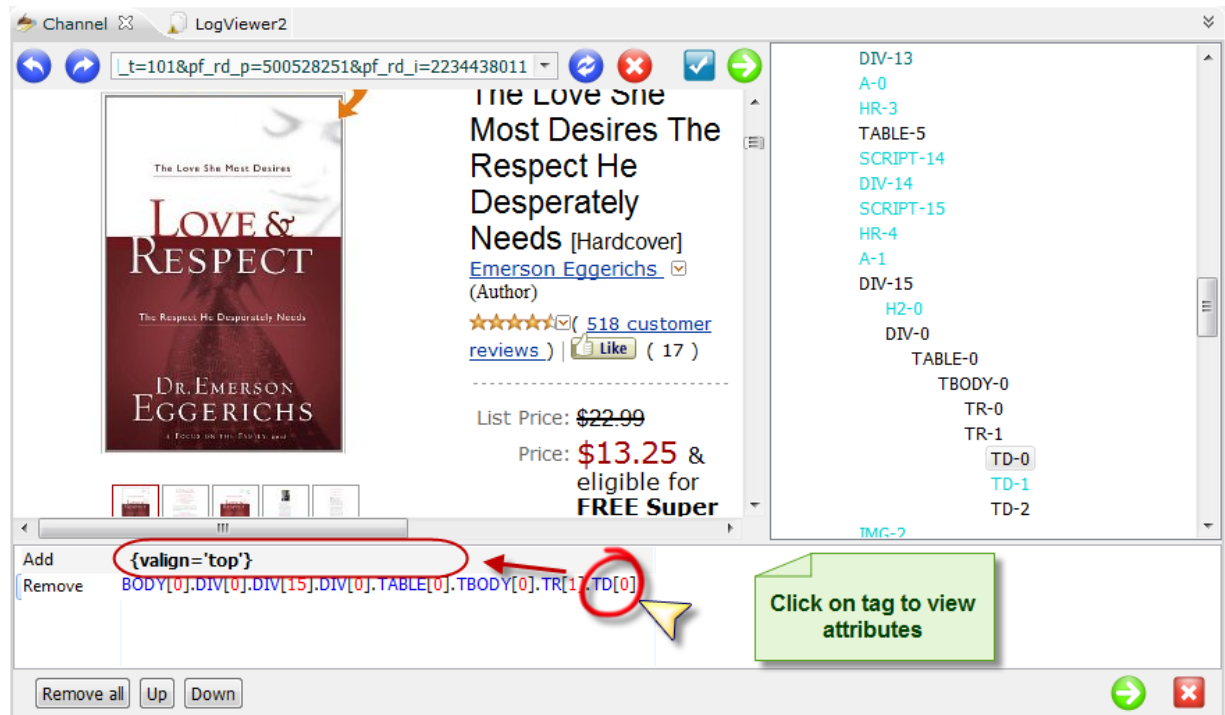
	Update content from Internet.
	Stop.
	Browse the site to download the full attributes of the site or the user selects Done when finished.
	Clear all block.

- ✖ On a web page, the user need only peel a couple of data regions. HTML trees right help users navigate through the site's data. Each node (branch) corresponds to a range of data. The data region may contain many areas the data. The user selects one or more buttons just tell the program the user simply gets the data associated with it.
- ✖ After selecting the button corresponding to the data user should take to get the address. Specific address of the button will display the bottom (like the red frame). I press the button (+ additional blocks) to add to the list of blocks are blocks

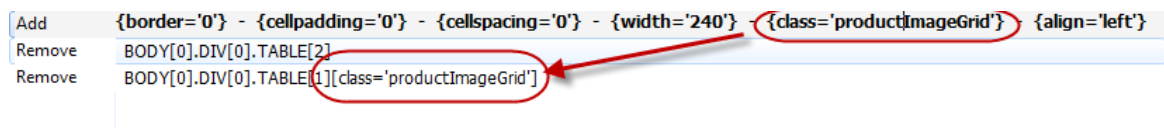
✖ After that the interface program will be as follows:



- ✖ To find the node containing the data more quickly, highlight the text after (the text displayed on the page), then the program automatically captured and highlighted areas indicate data corresponding button on the side as shown.
- ✖ Viewing and editing attributes of selected tag.



**For example:** User wants to get elements of class *productImageGrid* only. So block tag will be: `BODY[0].DIV[0].TABLE[1][class='productImageGrid']`





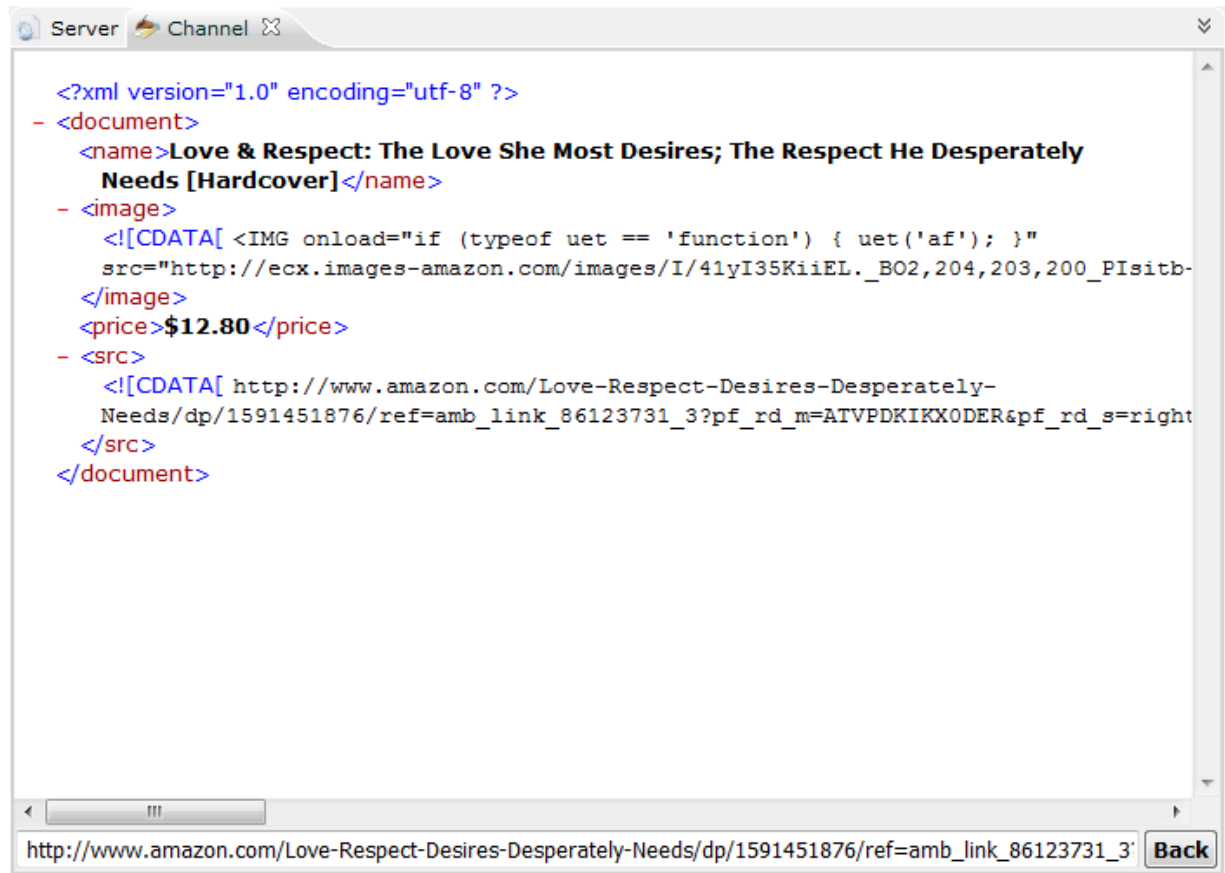
✳ **Math expression:** This feature allows user can config channel flexible with math expression. Some common math expression:

- \* : for all
- $i\%2=0$  : only even ;  $i\%2 = 1$ : only odd
- $i>3$  : for all greater than 3.
- And other math operator: =, <=, >=, !=, >, <...

✳ **Auto detect:** This feature allows user finding the tag more easily. He or she just selects a text area in the left side of browser. Vietspider will automatically detect the tag on the tree on the right as shown below:



- ✖ After selecting, click the button  otherwise, user click the button  to clear the entire block.
- ✖ If user want to check the crawl area, click **Verify** button , a window will be show as bellow:



- ✱ After verifying, the entire configuration is correct, please click **Back** button to get back and click **Save** to save configuration of channel with Name, crawl area... as defined.
6. **Remove area:** To allow user remove block, tag that not necessary in crawl area.
    - ✱ In area where web data need to crawl including unnecessary block, user clicks on **Remove area** icon at the bottom.
  7. Users can eliminate unnecessary block when the user has selected.
  8. Users select the **type of crawl** to suit the information content of the website shows.
  9. Depending on the information content is shown on the website that we choose appropriate type of separation like Normal (peeled normal) or type Row (split into the web page format).
    - Select the type of data format, scan rate, consistent with the priority of the website contents in section format, scan rate, level of priority.

- ✖ Selecting **Scan level** to crawl website in deep, value of **Scan level** depend on content of website.
- ✖ This feature helps determine system load frequency crawl of the site for a certain period. In frequency crawl, the default, provisions is corresponding to 30 minutes. Value for the priority is set in the **System** Configuration instructions in the next section of this document. If users set, the priority crawl is 48 = 1 day, after a day on the program will automatically scan the website to get the web data.

**10. Priority:** When two or more channels have same schedule to crawl data, which one has smaller priority will be crawled first.

Channel Name( ☐ Server? ) History Version

Home Page(\*):  + -

Crawled Link Pattern:  + -

Sample Page:  + -

Link Pattern:  + -

Extract Type:  Removed Area: ☐ ☐ Remove Sibling Blocks Behind Extract Data: ☐

► Properties

Encoding:  + Crawl Deep:  Crawl Hours:  Priority:

For example: Amazon and eBay channel crawl at the same time, Amazon's priority value is 8, and eBay's priority value is 1. Therefore, eBay channel will be crawled first, and then Amazon.

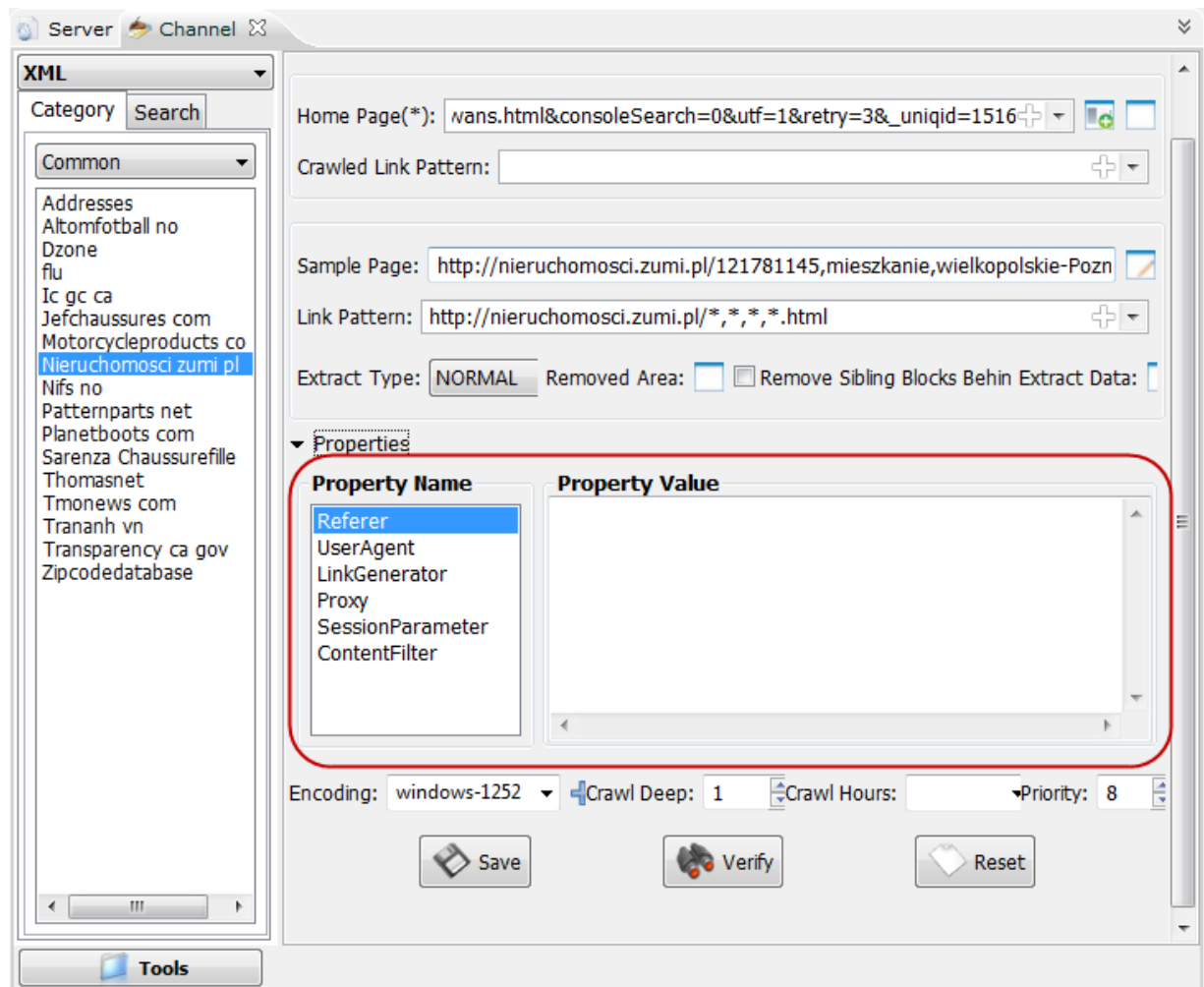
- 11. Crawl hours:** By default, VietSpider will crawl 24/24hours per day. If user want to schedule VietSpider exact time to crawl, user set value as form: “**From time – End time**”

The screenshot shows the VietSpider configuration window. At the top, there's a 'Channel Name' field with a 'Server?' checkbox and a 'History Version' label. Below this is a 'Home Page(\*)' field containing 'http://www.amazon.com' and a 'Crawled Link Pattern' field. The 'Sample Page' field contains 'http://www.amazon.com/Kindle-Wireless-Reader-Wifi-Graphite/dp/B003DZ1'. The 'Link Pattern' field contains 'http://www.amazon.com/Kindle-Wireless-Reader-Wifi-Graphite/dp/B003DZ'. The 'Extract Type' is set to 'NORMAL', and there are checkboxes for 'Removed Area', 'Remove Sibling Blocks Behind', and 'Extract Data'. Under the 'Properties' section, the 'Encoding' is 'UTF-8', 'Crawl Deep' is '1', 'Crawl Hours' is '1-12' (highlighted with a red circle), and 'Priority' is '8'. At the bottom are 'Save', 'Verify', and 'Reset' buttons.

For example: 1-12; the program will crawl from 1 to 12 hour.

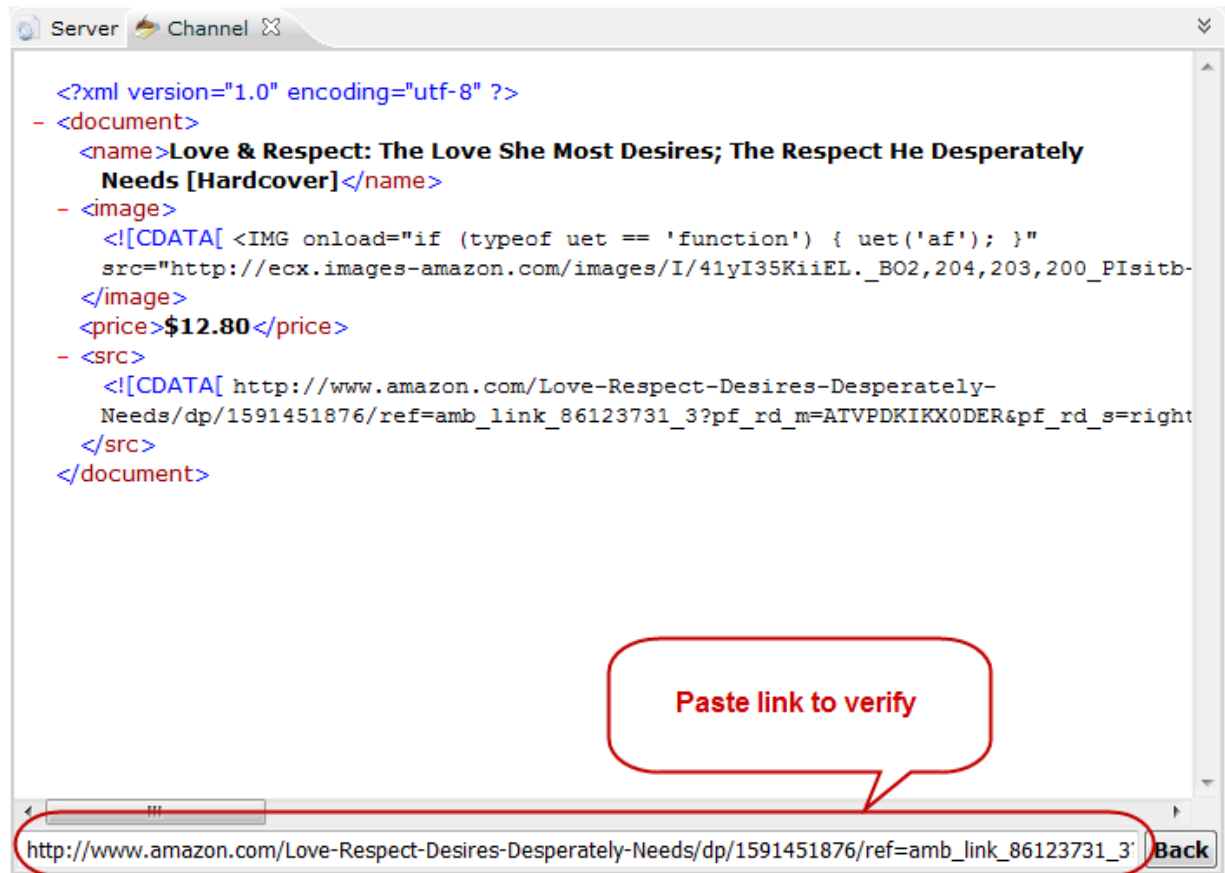
- 12. Channel property:** adding a number of extended attributes for data mining. Some sites will require additional parameters can be exploited. The figure below illustrates the interface of the source attribute:





### 13. Verify and Save the configuration:

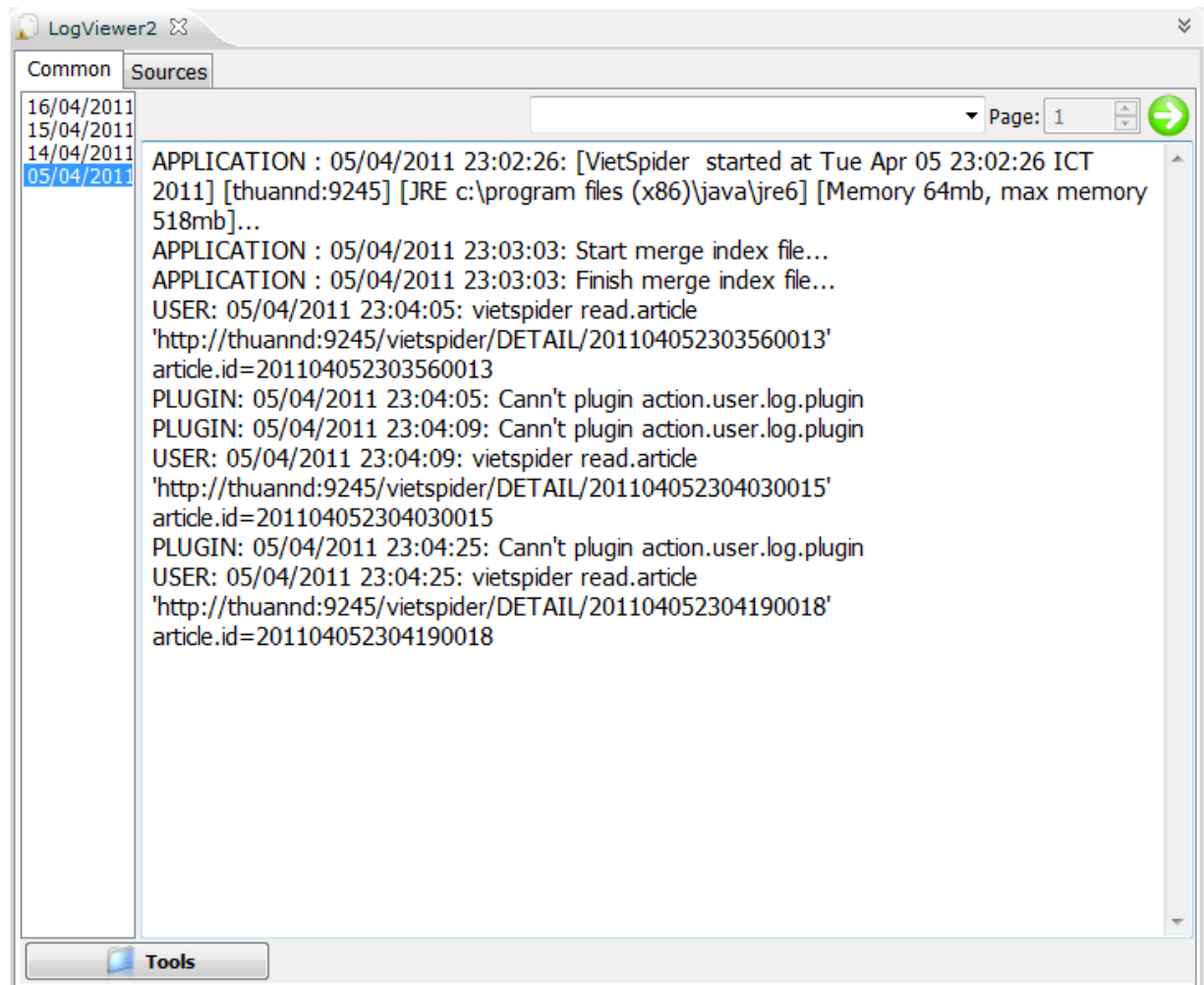
- Click on **Verify** to check the crawl area is correct or not. Put some more URLs to verify **Template channel**.



- Click on **Save** to save configuration. Or click **Reset** to reset configuration

### 3.19 View log

- On the toolbar select View log function will be the interface program as shown below:



#### 1. Graphic interface support:

- APPLICATION: Application error
- SOURCE: Errors during the scan and the channel configuration.
- SOURCE&ERROR: Alert wrong configuration channel.
- USER: Monitoring users.

## 2. Sources

This function helps to collect statistics about downloaded data of VietSpider and status of each channel such as: Score, Message, saved, home, link, data, invalid, script, and bytes. These statistics are arranged in day on the right side of the page like image below. Besides, a list of all downloaded channels in that day is also displayed.

Name	Score	Message	Saved	Home	Link	Data	Invalid	Script	Bytes
2011_04_16.Common...	0.0	{source.not.found}	0	0	0	0	31	0	0 Bytes
Addresses.Common.XML	-1.0	{crawling.stop.by.user}	0	15	0	2	0	0	0 Bytes
Amazon Books.Product...	-7.0	{crawling.end}	0	18	0	0	0	0	0 Bytes
Download CNET.Produc...	-8.5	{crawling.stop.by.user}	0	19	0	2	0	0	22.66 KB
Ebay.Product.XML	-21...	{not.found.response}	0	403	0	0	0	0	0 Bytes
Mp3 Artist.Music.XML	-7.0	{crawling.end}	0	17	0	0	0	0	0 Bytes
Tmonews com.Commo...	-2.0	{crawling.end}	0	28	0	0	0	0	0 Bytes
Trananh vn may tinh.C...	0.0	{validate.fail.no.crawl}	0	15	0	0	20	0	0 Bytes
Trananh vn.Common.X...	-96...	{out.of.depth.downlo...	0	194	102	1606	0	0	18.06 MB
Yahoo Music.Music.XML	-28.0	{document.is.empty}	0	15	0	22	0	0	1.27 MB
Zipcodedatabase.Com...	-2.0	{crawling.end}	0	8	0	0	0	0	38.37 KB

### Summary tab

**Name:** Name of channel

**Score:** Accurately assessing data got from the channel. If score number is smaller than zero (score<0), it means that you have configured wrongly the channel (you need to reconfigure).This is illustrated by red line in the picture below.

**Message:** Notice of VietSpider for each channel.

**Saved:** Number of data pages in database.

**Home:** Total number of links is taken by VietSpider from all pages in the Home page.

**Link:** Number of links that VietSpider got in **Crawled Link Pattern**.

**Data:** Total of links today.

**Invalid:** Number of error links (because they are not compatible with Sample page).

**Bytes:** Total off traffic for each channel.

Name	Score	Message	Saved	Home	Link	Data	Invalid	Script	Bytes
2011_04_16.Common....	0.0	{source.not.found}	0	0	0	0	31	0	0 Bytes
Addresses.Common.XML	-1.0	{crawling.stop.by.user}	0	15	0	2	0	0	0 Bytes
Amazon Books.Product...	-7.0	{crawling.end}	0	18	0	0	0	0	0 Bytes
Download CNET.Produc...	-8.5	{crawling.stop.by.user}	0	19	0	2	0	0	22.66 KB
Ebay.Product.XML	-21...	{not.found.response}	0	403	0	0	0	0	0 Bytes
Mp3 Artist.Music.XML	-7.0	{crawling.end}	0	17	0	0	0	0	0 Bytes
Tmonews com.Commo...	-2.0	{crawling.end}	0	28	0	0	0	0	0 Bytes
Trananh vn may tinh.C...	0.0	{validate.fail.no.crawl}	0	15	0	0	20	0	0 Bytes
Trananh vn.Common.X...	-96...	{out.of.depth.downlo...	0	194	102	1606	0	0	18.06 MB
Yahoo Music.Music.XML	-28.0	{document.is.empty}	0	15	0	22	0	0	1.27 MB
Zipcodedatabase.Com...	-2.0	{crawling.end}	0	8	0	0	0	0	38.37 KB

## Log tab

When admin double click on a row, you will see all detail report for each channel. Or you can select a channel and click on **Log tab**, a window will look like this:

Summary

Log

Quick View

Name	Score	Message	Saved	Home	Link	Data	Invalid	Script	Bytes
2011_04_16.Common...	0.0	{source.not.found}	0	0	0	0	31	0	0 Bytes
Addresses.Common.XML	-1.0	{crawling.stop.by.user}	0	15	0	2	0	0	0 Bytes
Amazon Books.Product...	-7.0	{crawling.end}	0	18	0	0	0	0	0 Bytes
Download CNET.Product.XML	-5	{crawling.stop.by.user}	0	19	0	2	0	0	22.66 KB
Ebay.Product.XML	-217.0	{not.found.response}	0	403	0	0	0	0	0 Bytes
Mp3 Artist.Music.XML	-7.0	{crawling.end}	0	17	0	0	0	0	0 Bytes
Tmonews com.Commo...	-2.0		28	0	0	0	0	0	0 Bytes
Trananh vn may tinh.C...	0.		15	0	0	20	0	0	0 Bytes
Trananh vn.Common.X...	-9		194	102	1606	0	0	0	18.06 MB
Yahoo Music.Music.XML	-2		15	0	22	0	0	0	1.27 MB
Zipcodedatabase.Com...	-2		8	0	0	0	0	0	38.37 KB

Double click

Or

Summary

Log

Quick View

Name	Score	Message	Saved	Home	Link	Data	Invalid	Script	Bytes
2011_04_16.Common...	0.0	{source.not.found}	0	0	0	0	31	0	0 Bytes
Addresses.Common.XML	-1.0	{crawling.stop.by.user}	0	15	0	2	0	0	0 Bytes
Amazon Books.Product...	-7.0	{crawling.end}	0	18	0	0	0	0	0 Bytes
Download CNET.Product.XML	5	{crawling.stop.by.user}	0	19	0	2	0	0	22.66 KB
Ebay.Product.XML	-7.0	{not.found.response}	0	403	0	0	0	0	0 Bytes
Mp3 Artist.Music.XML	-7.0	{crawling.end}	0	17	0	0	0	0	0 Bytes
Tmonews com.Commo...	-2.0	{crawling.end}	0	28	0	0	0	0	0 Bytes
Trananh vn may tinh.C...	0.0	{validate.fail.no.crawl}	0	15	0	0	20	0	0 Bytes
Trananh vn.Common.X...	-964.5	{out.of.depth.downlo...	0	194	102	1606	0	0	18.06 MB
Yahoo Music.Music.XML	-28.0	{document.is.empty}	0	15	0	22	0	0	1.27 MB
Zipcodedatabase.Com...	-2.0	{crawling.end}	0	8	0	0	0	0	38.37 KB

The log tab will be shown as bellow:

The screenshot shows the LogViewer2 application with the Log tab selected. The interface includes a left sidebar with a list of sources and a main area displaying log entries. Annotations include:

- Click to modify the channel configuration:** Points to the "Edit XML.Common.Addresses" link.
- Click to view webpage in browser:** Points to the "Edit Test Website" link.
- Suppend to crawl the channel:** Points to the "Disable Channel" link.
- Statistic:** Points to the log entry details.

The log entry displayed is:

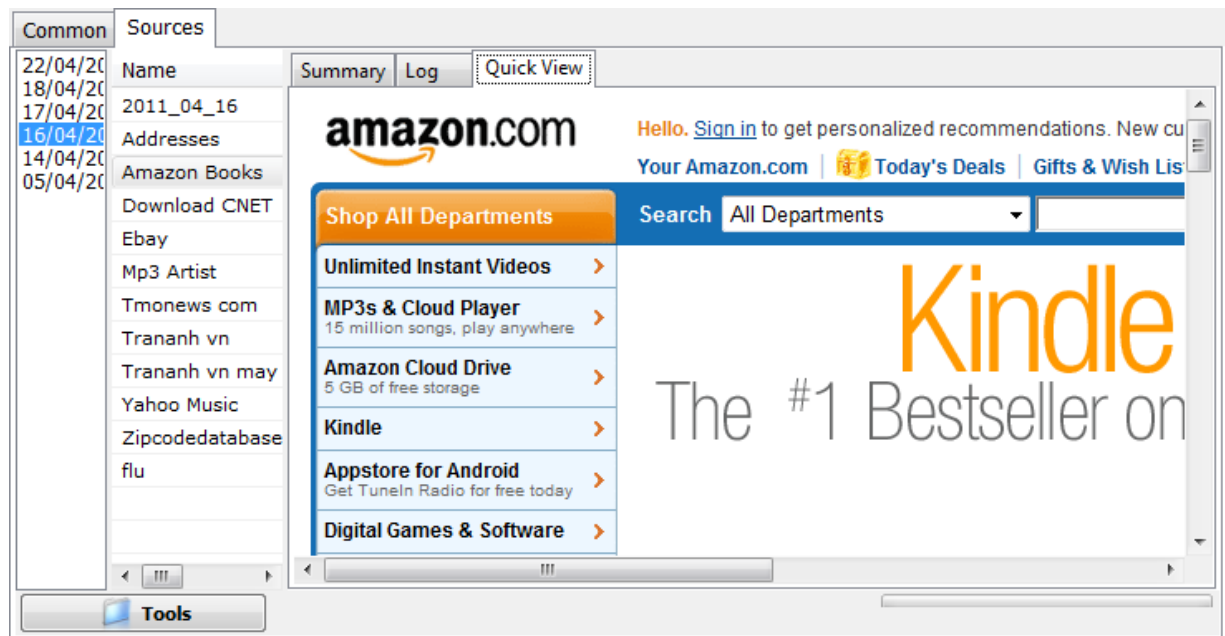
```

http://www.addresses.com/results.php?ReportType=34&qfilter[pro]=on&qj=0&qk=10&qn=Williams&qs=AL
{crawling.stop.by.user}
16/04/2011 12:57:05 | ID: 26 | URL Code: 2076534581 | Level: 0 | Phase: Source | Type:
Homepage | Size: 0 | N/A

```

## Quick View Tab

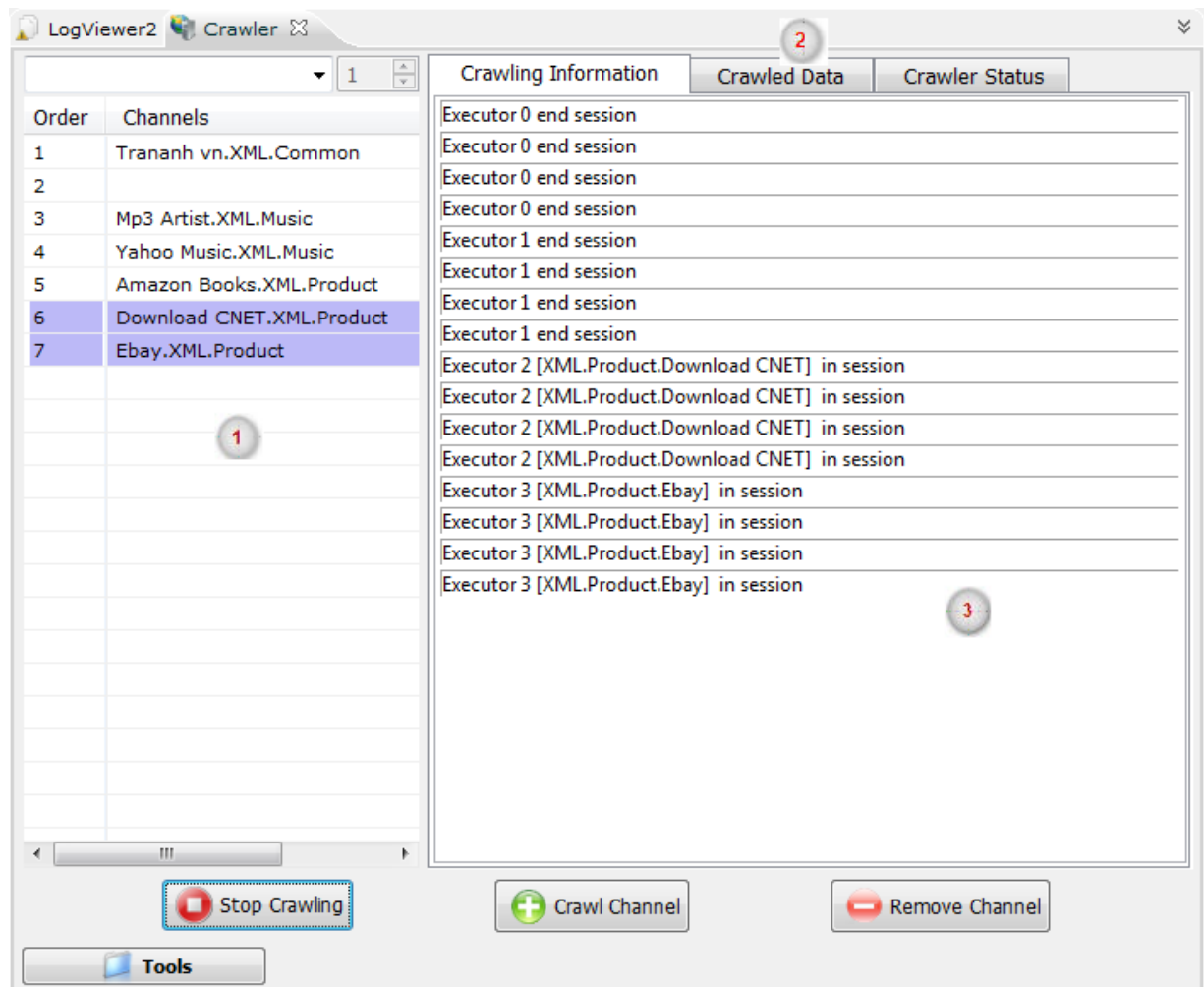
The program will allow viewing quickly the address at the **Home** page in channel configuration. Take an example of Amazon.com website.



## 3.20 Crawl Data

### 3.20.1 Interface

- **Crawl data interface** allows monitoring and management... portfolio flows to load and process the information load.
- On the toolbar, click on the **Crawl Data** Item in the menu on the left. A window display as shown below:



Title	Content
1	List of channels.
2	List of downloaded content.
3	Other functions.



### 3.20.2 Crawl content

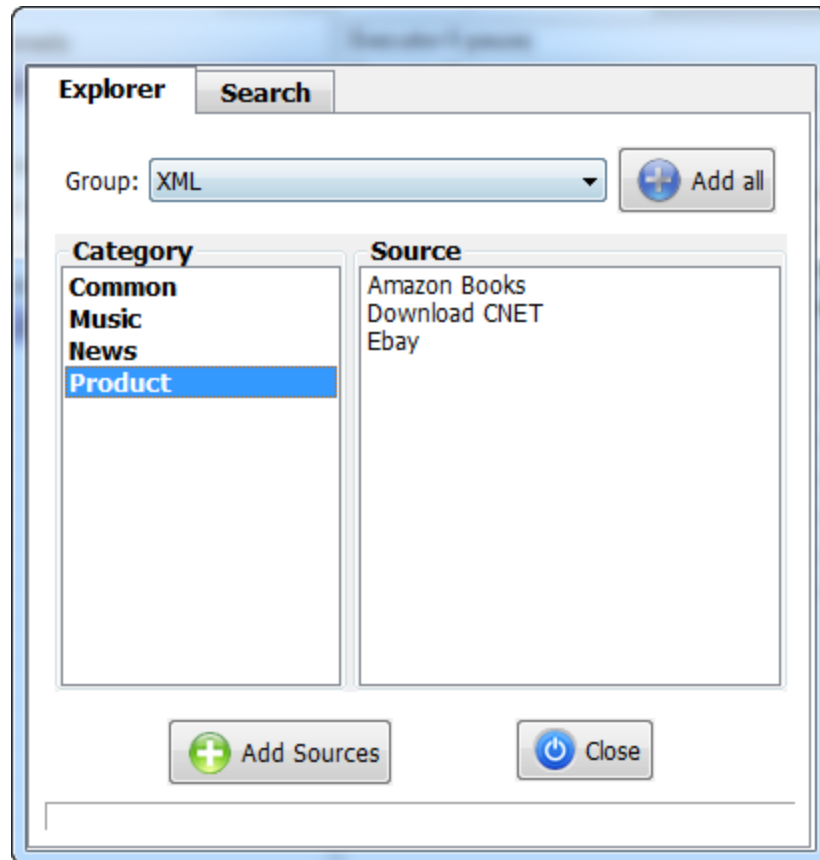
- **Crawl content, Stop crawl content** function for all channels in the scan list.
- On the interface system, the user selects the button **Crawl content**, the program will start downloading the data if it has not run or load data if the program is paused, it will continue to run.

### 3.20.3 Pause crawling

- Pause crawl function allows pausing crawling process.
- On the interface **Downloading content**, user click on Pause bottom to pause the channel in the list.

### 3.20.4 Add crawling channel

- Add crawl channel function is the ability to download one or several channels have been configured in the system to load the downloaded channel list.
- On the interface of **Crawl content** , user click Add channel bottom, a windows will be shown as below:



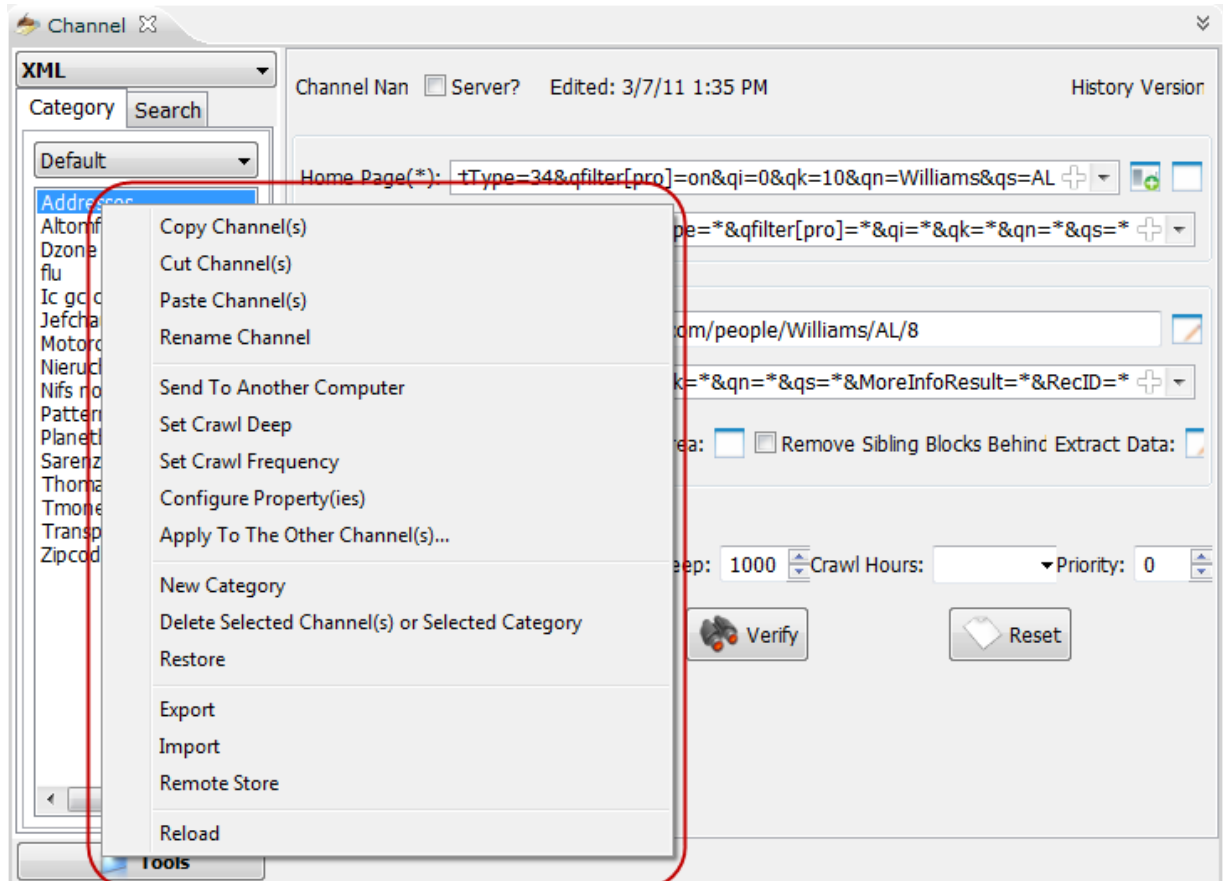
1. Users can select a channel or multiple channels in the same column using the **Ctrl** or **Shift** key and then click **Add** channels.
- **Add all** allows adding full configuration in channel list to the list for downloading. Once finished, click **Close**.

### 3.20.5 Remove channel from the list of scanned channels

- **Remove channel** function allows user remove channels from list.
- User can select a channel or multiple channels in the list using **Ctrl** or **Shift** key and then click **Remove Channel**

### 3.20.6 Right click menu.

- Users choose a channel in the channel list, then right click, a window will be shown as below:



- Users can select one of the functions displayed on the menu right. The utility functions which are described below:
  - **Download channel:** The program will download immediately the selected channel.
  - **Modify:** Edit configuration of selected channel.
  - **Add channel:** Add a channel to the list
  - **Remove channel:** Remove selected channel from the list.
  - **Remove all:** Remove all channels in the list.

- **Stop:** Stop downloading selected channel.
- **Update:** Update information and configuration of channel.

### 3.21 User management

- User interface allows to manage user in system including: User account, Group users
- On Tools menu, user click on **Users** on the left of menu, the interface will be shown as:

The screenshot displays the 'LogViewer2 Server' application window. On the left is a 'Configuration Menu' with options: System, Crawler, Data Group, **Users** (highlighted), Plugin, and Email. The main area is divided into three sections: 1. A 'Group' dropdown menu set to 'guest' with a green '+' icon. 2. A 'Users' section containing an empty list box and '+ Add' and '- Remove' buttons. 3. A 'Categories' section containing an empty list box with a '\*' icon and '+ Add' and '- Remove' buttons. On the right side, there are input fields for 'Username (\*)', 'Password 1x (\*)', and 'Password 2x (\*)', a 'Role' dropdown menu set to 'READ ONLY', and text boxes for 'Full Name:', 'Email:', and 'Groups:'. At the bottom right are 'Delete' and 'Save' buttons. A 'Tools' button is located at the bottom left of the main window.

### 3.21.1 User account

#### User account interface

- User account feature to manage the information about the user's account in the system.  
This feature is only available to administrators or managers.

#### Viewing User Accounts

1. Administrator selects a account to view information in detail as illustrated below:

The screenshot shows the 'LogViewer2' application window with the 'Server' tab selected. On the left is a 'Configuration Menu' with options: System, Crawler, Data Group, Users (highlighted), Plugin, and Email. The main area is divided into three sections:

- Group:** A dropdown menu set to 'admin' with a green '+' icon to its right.
- Users:** A list box containing one entry, 'Vietspider'. Below it are '+ Add' and '- Remove' buttons.
- Categories:** An empty list box with '+ Add' and '- Remove' buttons below it.

On the right side, there are input fields for user details:

- Username (\*):** 'Vietspider'
- Password 1x (\*):** Masked with dots
- Password 2x (\*):** Masked with dots
- Role:** A dropdown menu set to 'ADMINISTRATOR'
- Full Name:** 'Vietspider'
- Email:** 'Vietspider@gmail.com'
- Groups:** An empty list box

At the bottom right are 'Delete' and 'Save' buttons. At the bottom left is a 'Tools' button.

### **Edit user account**

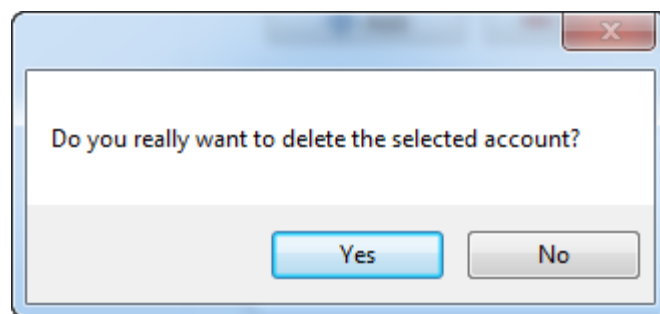
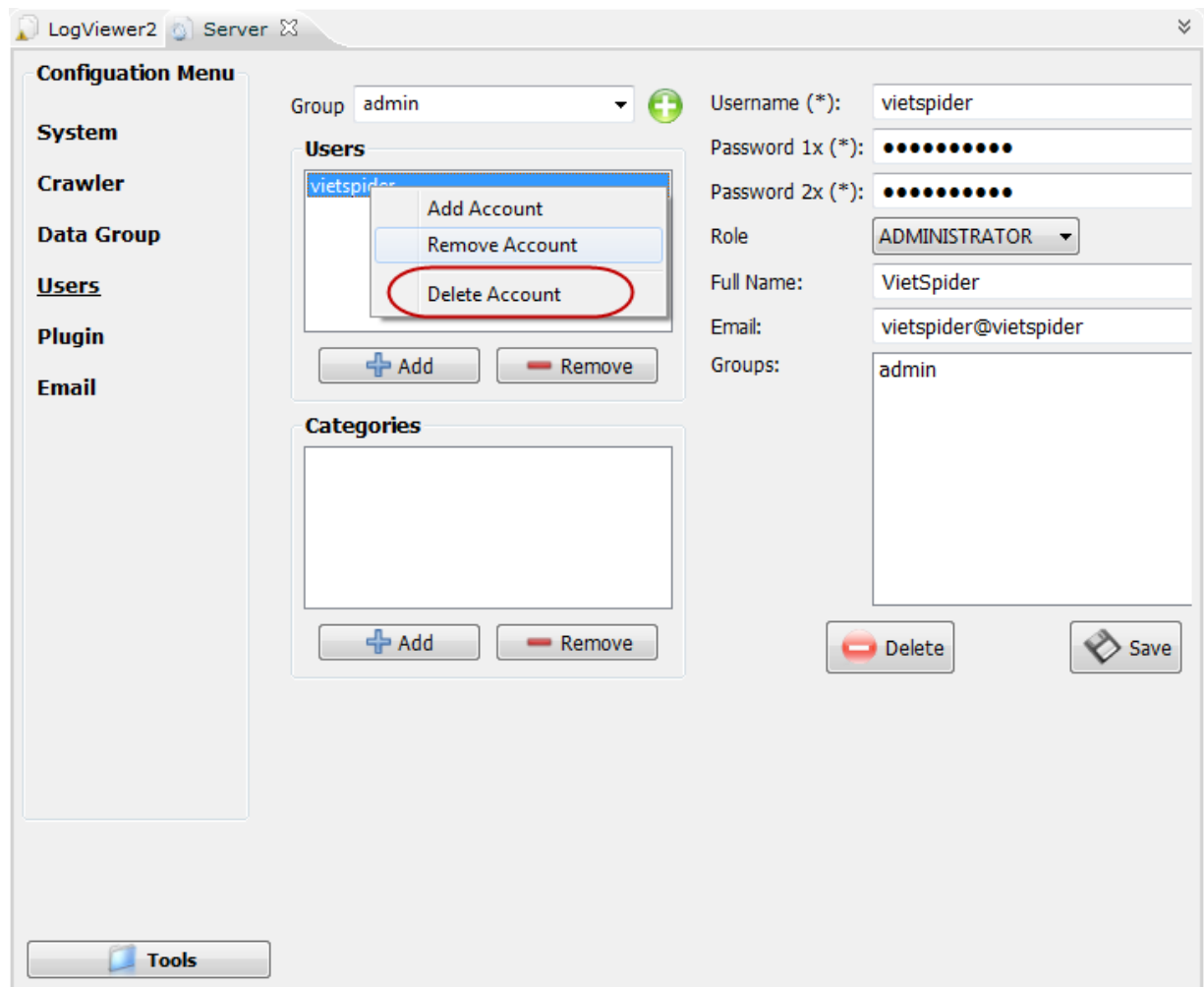
- On display the information of **user account**, administrator can change information such as usernames; passwords, powers ... then the administrator click **Save** to change new information.

### **Add new account**

- On the form displays user account information, the administrator click the **Add** button to create a new user account
- Then, admin fill account registration information in the form and click the **Save** button to save to create a new account.

### **Delete user account**

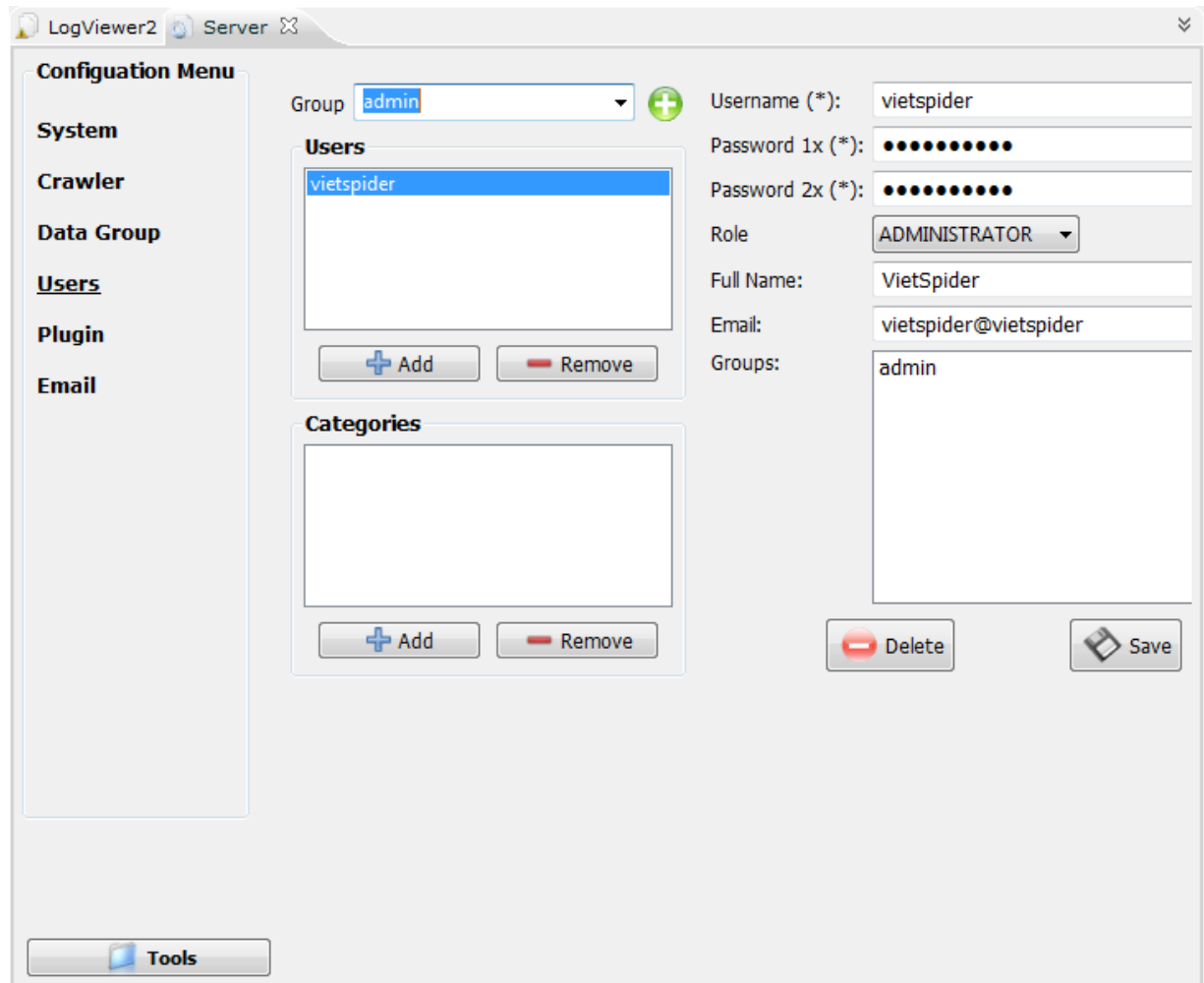
- In the list of users, administrators select a user account right-click and select **Delete account** to delete accounts.



1. Admin select **Yes** if really want to delete, otherwise, select **No** to cancel.

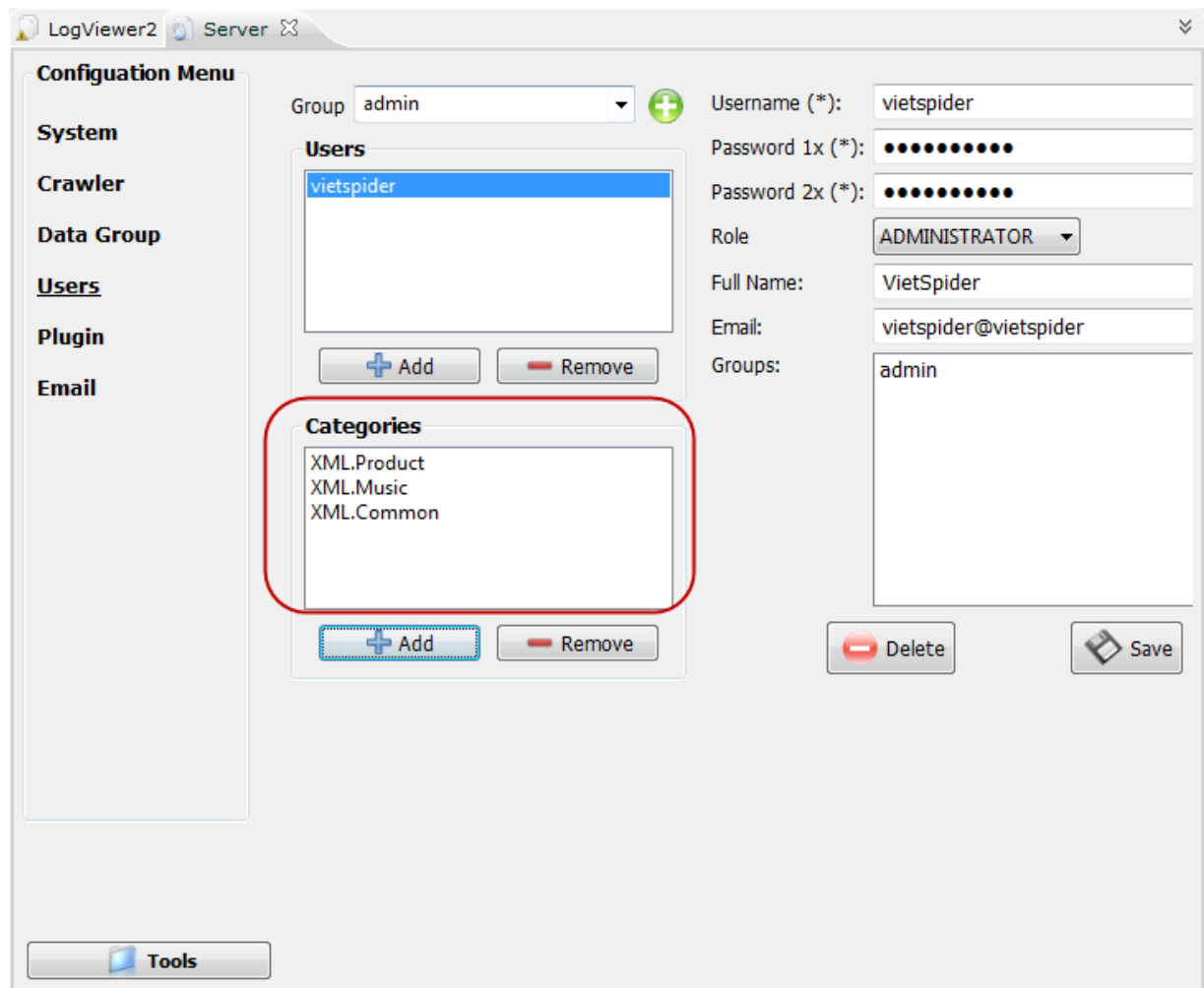
### 3.21.2 Manage categories in user group

- On the user group interface, the administrator can view information category on the user's program as shown below:

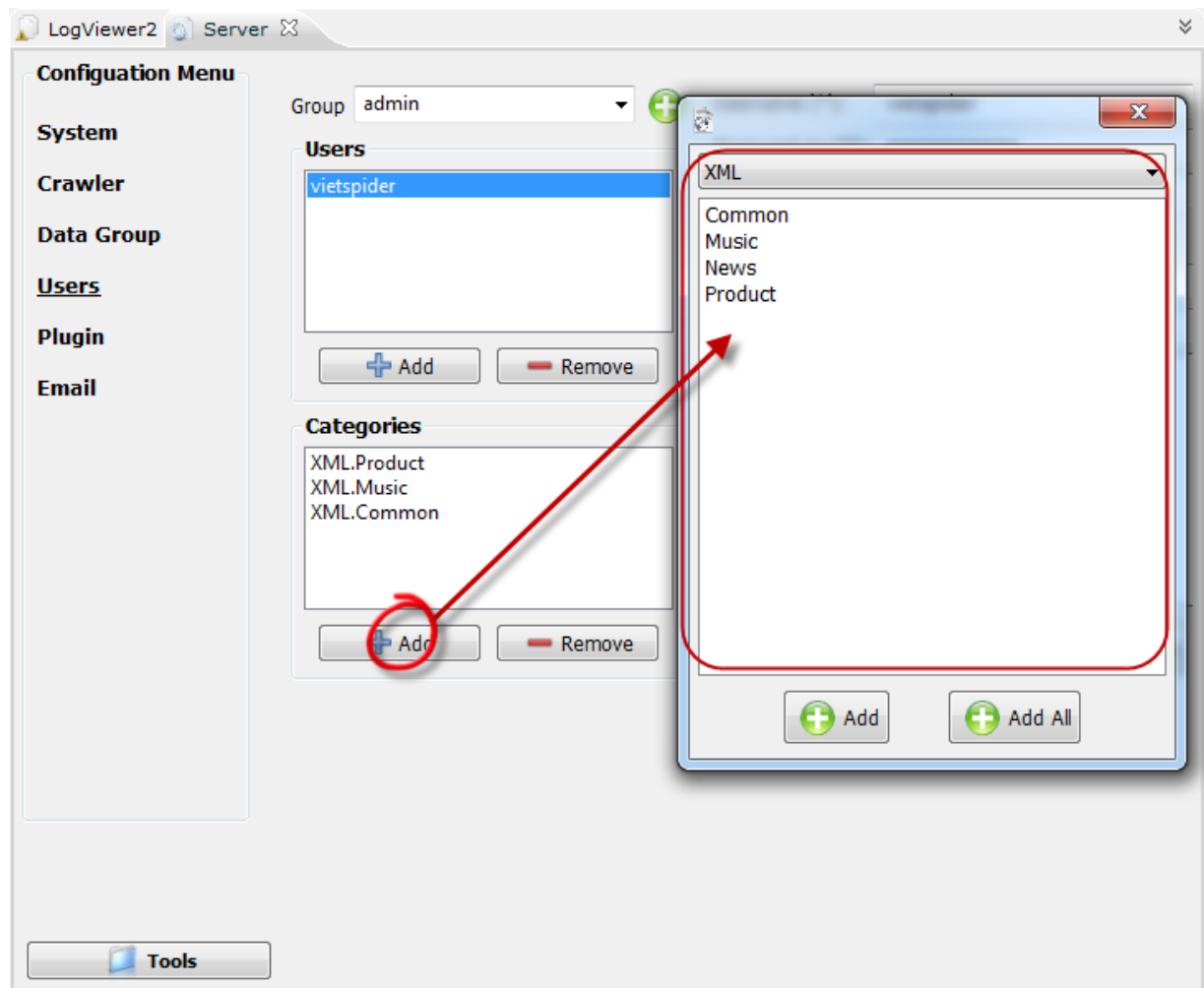


- The administrator can view the permission of each user or group can do in the categories.
- **Viewing category:** On User group interface, administrator select a group in system, select a user in that group, and automatically a windows will be shown a list of categories.

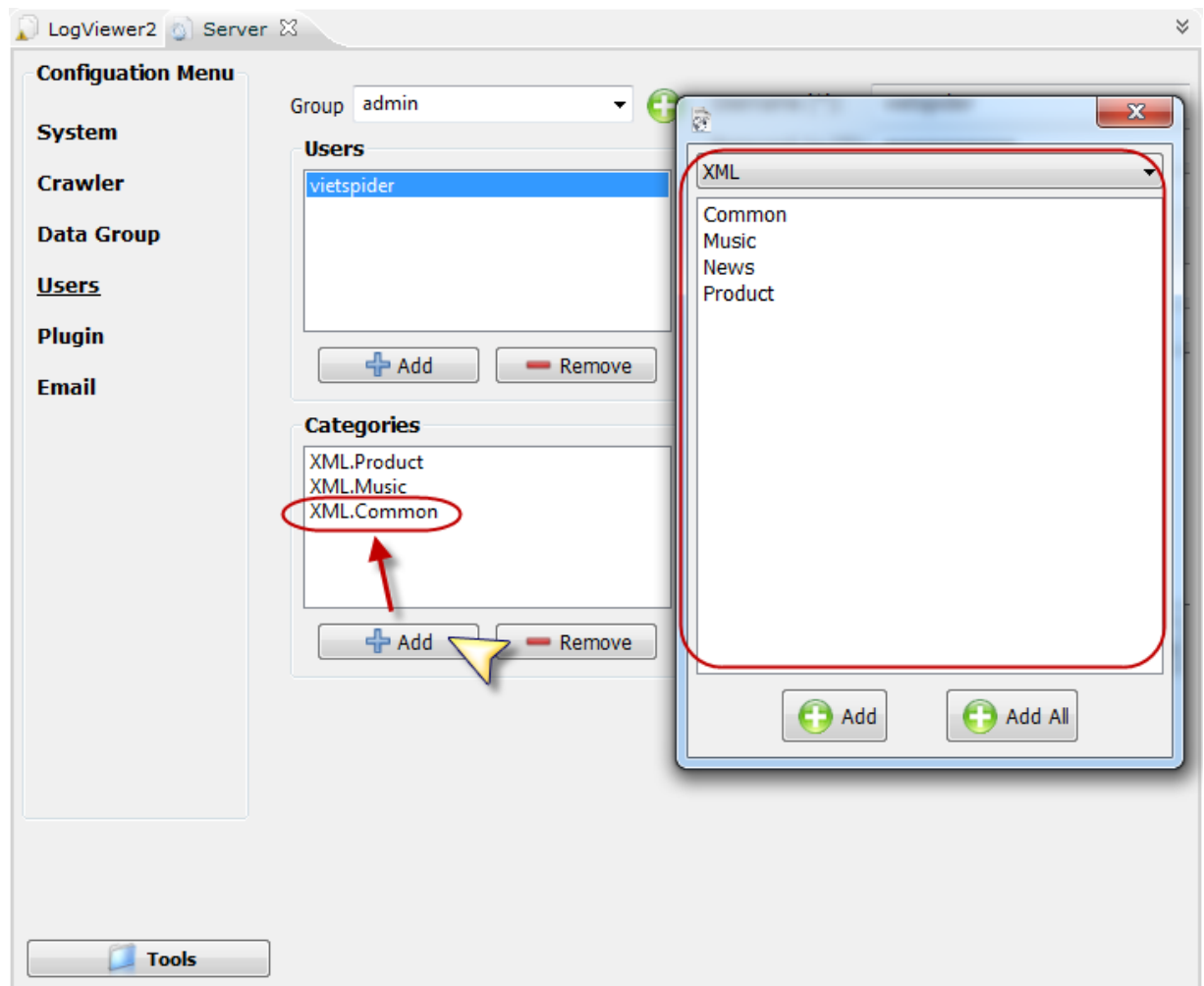




- **Add categories:** On user interface, administrator click on Add icon to add new categories to user, the interface as shown:



- Click on **Add category** button, a window will be appeared and ask administrator to select exist categories as figure bellow:



+ Administrator can select a category or multiple categories in the list and then click **Add** to add new categories to category user list.

- **Remove a category:** On category list, administrator select a category or multiple categories by press **Ctrl** or **Shift** and select Remove category.
- After changing information of user, administrator has to click on **Save** icon to save it.

## 3.22 Configuration

### 3.21.1 Configuration menu

- Graphical user interface for configuration allows administrators to configure the system parameters and loading programs.
- On the Tools menu, select the Configuration icon, the program interface as shown below:

The screenshot shows the 'LogViewer2' application window with the 'Server' tab selected. The 'Configuration Menu' on the left lists 'System', 'Crawler', 'Data Group', 'Users', 'Plugin', and 'Email'. The 'System' menu is active, displaying the 'Server Configuration' and 'Database Setting' sections.

**Server Configuration**

Host Name: Interboy-PC

App Port: 9245 [Restart] Web Port: 9245 [Restart]

Product Name: vietspider

Keep Data For(days): 10 [up/down] Clean Data At: 12 [up/down]

☒ Delete Data from Database

**Database Setting**

☐ Use RDBMs?

Database Type: MySQL [dropdown]

Database Host: jdbc Port: mysql://1:

Database Name: vietspiderdb

Username: interboy Password: xxxxxx

☐ Delete All Old Table? ☐ Create New Table

[Finish] [Cancel]

[Save] [Reset]

[Tools]

## System

**LogViewer2 Server**

**Configuration Menu**

- System**
- Crawler
- Data Group
- Users
- Plugin
- Email

**Server Configuration**

Host Name:

App Port:   Web Port:

Product Name:

Keep Data For(days):  Clean Data At:

☒ Delete Data from Database

**Database Setting**

☐ Use RDBMs?

Database Type:

Database Host:  Port:

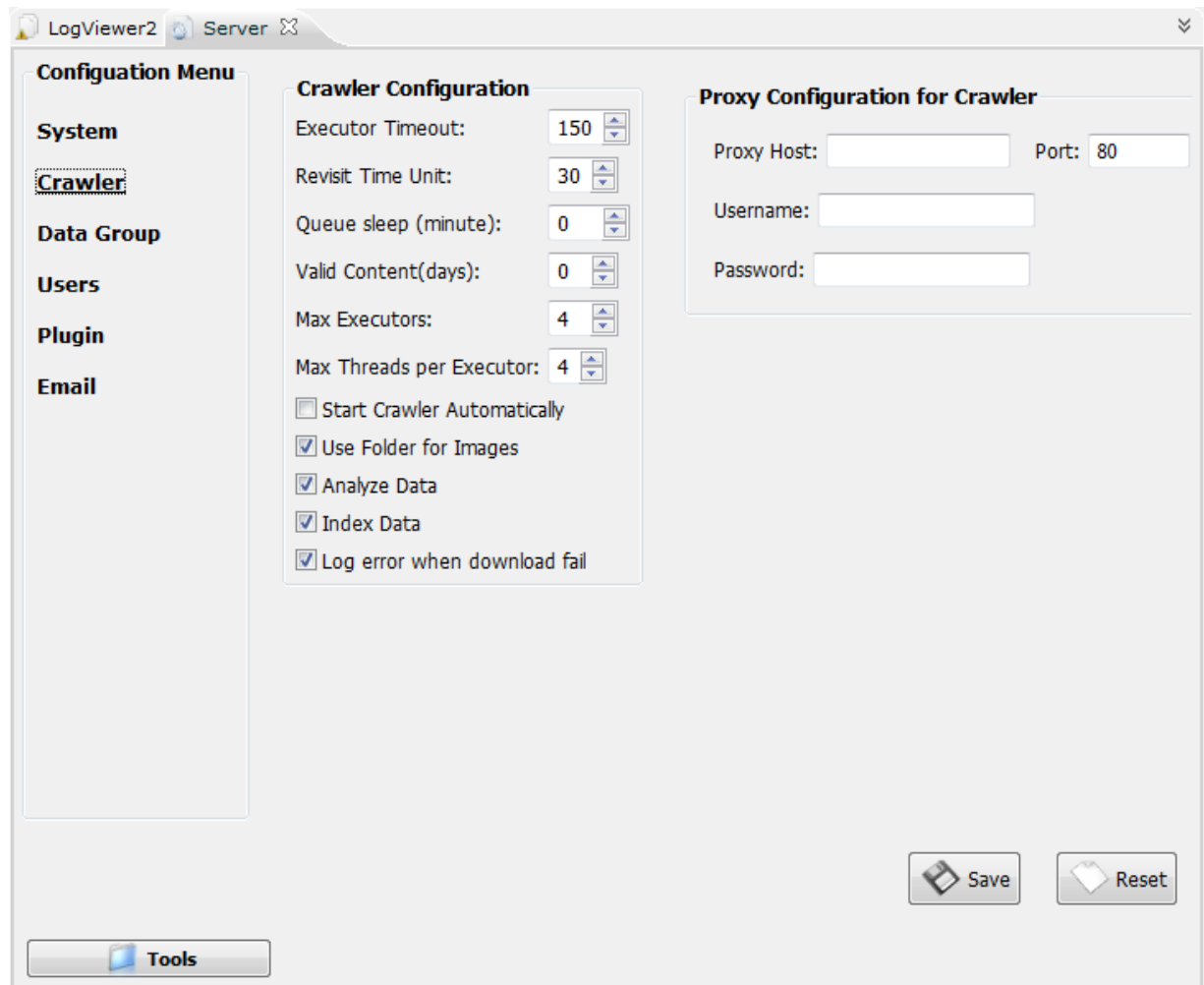
Database Name:

Username:  Password:

☐ Delete All Old Table? ☐ Create New Table

- Program requires information: **IP address, Port, Name of application, Keep data For, Clean Data At.**
- **Keep data for:** Unit is day. If data out of date will be removed base on the date of system minus this value.
- **Clean Data at:** Is a function allows to clean data in database at the time in the input form.

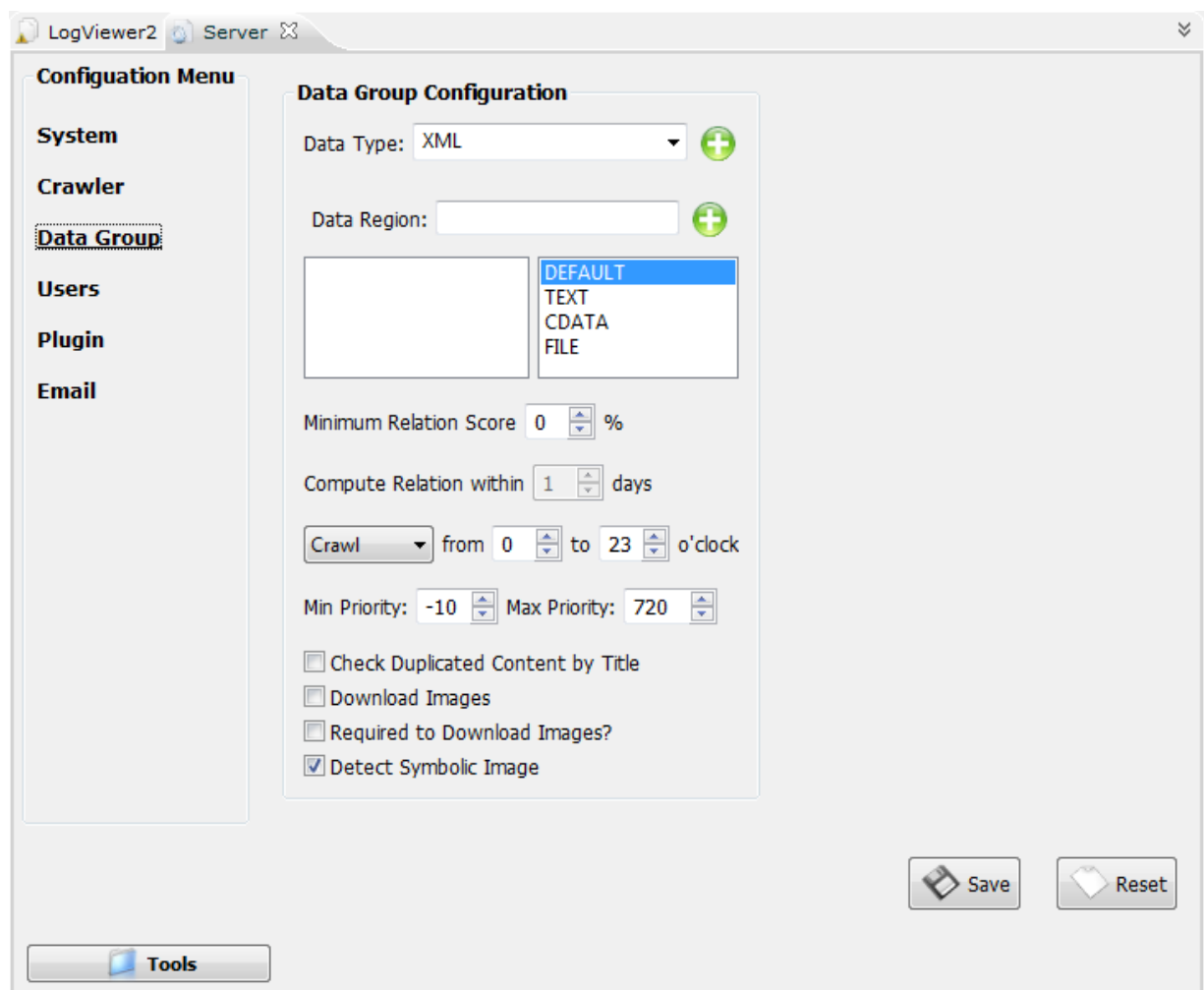
## Crawler



- **Configuration of Crawler:** Including configuration of what the crawler will get web data for example: Executor timeout, Revisit time unit, and queue sleep, User folder for images or Analyze data or so on.
  - **Queue sleep:** is the period of time crawler pause before next times to get web data. The unit is minute.
  - **Executor timeout:** request timeout of default browser. Unit is minute.
  - **Max executor:** Limit the number of channel when downloading at the same time.
  - **Max threads per executor:** Limit the flow per channel when downloading.
  - **Start Crawler automatically:** Automatic download channels when program execute.

- **User folder for images:** Not using database to store image, instead of using folder to store them.
- **Analyze data:** The program can analyze the duplicated data when downloading.
- **Index data:** The program can index the data.
- After finishing all configuration, Administrator click **Save** to save it, otherwise, click **Reset** to cancel.

### Data Group.



- **Configuring Data Group** including configuration of crawl area.
  - **Minimum relation score:** Is a percentage of related data each other.
  - **Compute Relation within:** Is the value of program index related data.

- **Time crawl:** User can set time for crawl or not for 0 to 23 hour in day.
- **Check duplicated content by title:** the program will check duplicate item by title. It avoid same item in database.
- **Download image:** Download image form source or not.
- **Required to download images:** If not select, program download images faille, the program will not download again. Otherwise, the program will force to download images.
- **Detect symbolic image:** Put information of image to Meta tag.
- After finishing, administrator click on **Save** to save configuration.