USER GUIDE

MANTRA
WEB EXTRACTOR

www.altiliagroup.com

ALTILIA
we make data smart

MANTRA WEB EXTRACTOR

USER GUIDE

TABLE OF CONTENTS

CONVENTIONS 2

CHAPTER 2 – BASICS 6

CHAPTER 3 - WORKSPACE 7

Menu bar 7

Toolbar 9

Status bar 9

Web view 10

Task area 10

Preview area 11

CHAPTER 4 – RECORD A TASK 12

Creating a new task 12

Record navigation actions 13

Record extraction actions 19

CHAPTER 5 – TASK EDITING/HOW TO EDIT A TASK 39

CHAPTER 5 – TASK RUNNING/HOW TO RUN A TASK 54

CHAPTER 6 – KEYBOARD SHORTCUT 55

CHAPTER 7 - TROUBLESHOOTING 56

CONVENTIONS

To help you when reading this manual, the following conventions have been adopted:

Convention	Meaning
Extraction actions	The actions that allow you to extract information from a web page (extract a list of elements, refine a list, define a pager, etc.…)
Firebug inspector	A tool used to inspect the HTML, CSS, Script and DOM of a web page
Navigation actions	The actions that a human performs when browsing a website (click on a link, form filling, etc.…)
Page action	A block in the task area representing a web page
Task	A web extraction process
Web extraction process	A set of navigation and extraction actions
MWE	MANTRA Web Extractor
Wep file	The file where a task is stored

CHAPTER 1 – GETTING STARTED

MANTRA Web Extractor is a visual tool that allows you to define extraction processes from the web. For example, you can use the Web Extractor to get product information from an ecommerce site or users posts from a forum.

You can download the correct version of the MWE, based on your operating system (Microsft Windows, Linux or Mac OS) from ALTILIA's website www.altiliagroup.com. To launch the MWE, unzip the "MWE.zip" file that you have downloaded and then run the executable file called "MWE". The first time you launch the MWE, the tool will ask you for the username and password of your license account. You can purchase your username and password by contacting us, or you can obtain a 30 days free trial license by signing up through the appropriate form available on our website. If an error occurs during the login, please consult the troubleshooting section at the end of this document.

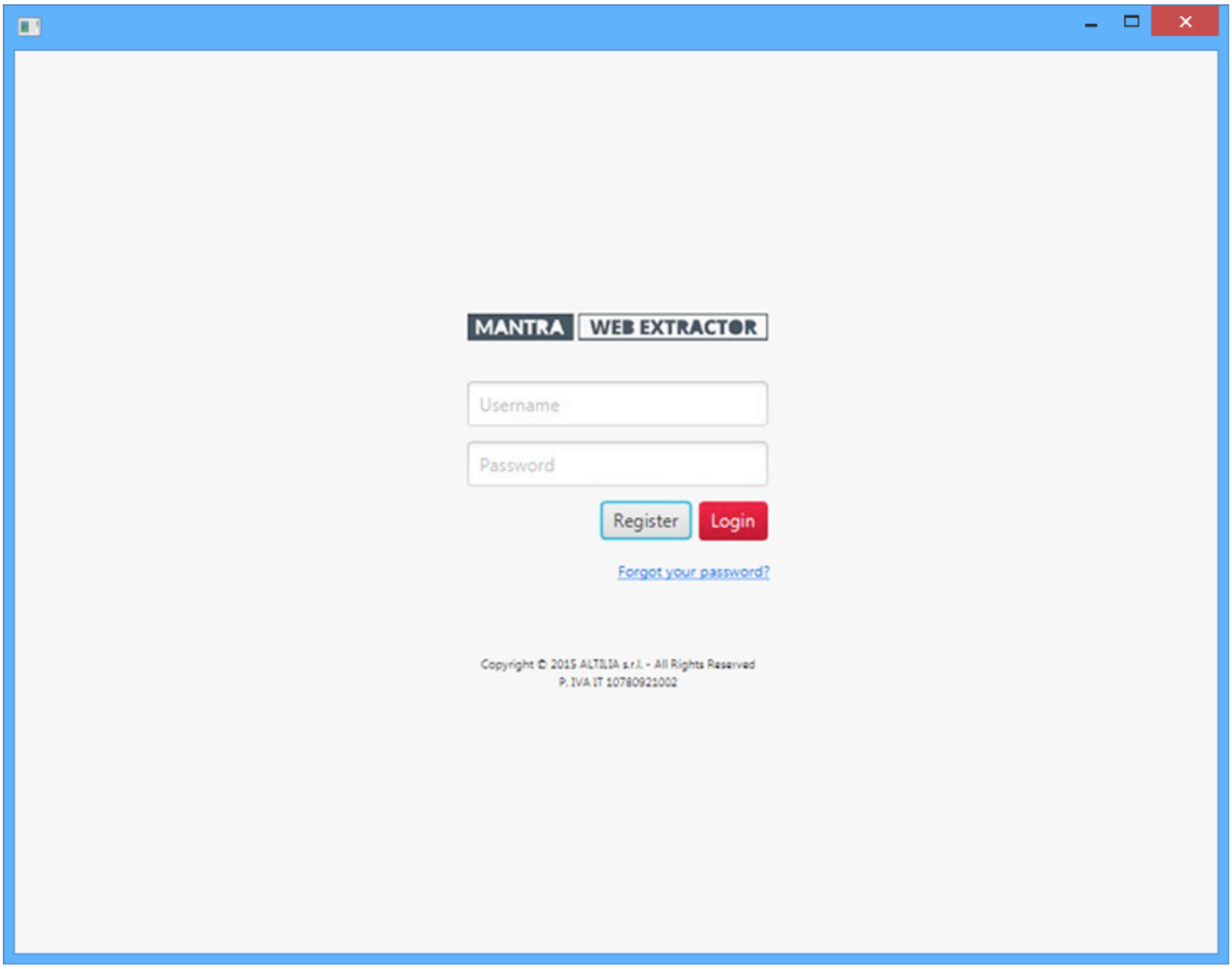

MANTRA WEB EXTRACTOR
Username
Password
Register
Login
Forgot your password?
Copyright © 2015 ALTILIA s.r.l. - All Rights Reserved
P. IVA IT 10780921002

Figure 1: Login form

To register you in order to obtain a 30 days free trial license, click on the *Register* button from the Login form. The register form (Figure 2) requires the following information:

1. First name
2. Last name
3. Company name/name of the company you work for
4. Position held in the company
5. An email address that will be used as your username and to send you the password.
6. A phone number (optional)

Before starting the registration, you need to accept our terms & conditions, by selecting the appropriate/dedicated checkbox. If the form is correctly completed, an email containing the username and password to access the MWE will be sent to your email address. If an error occurs, please consult the troubleshooting section at the end of this document.

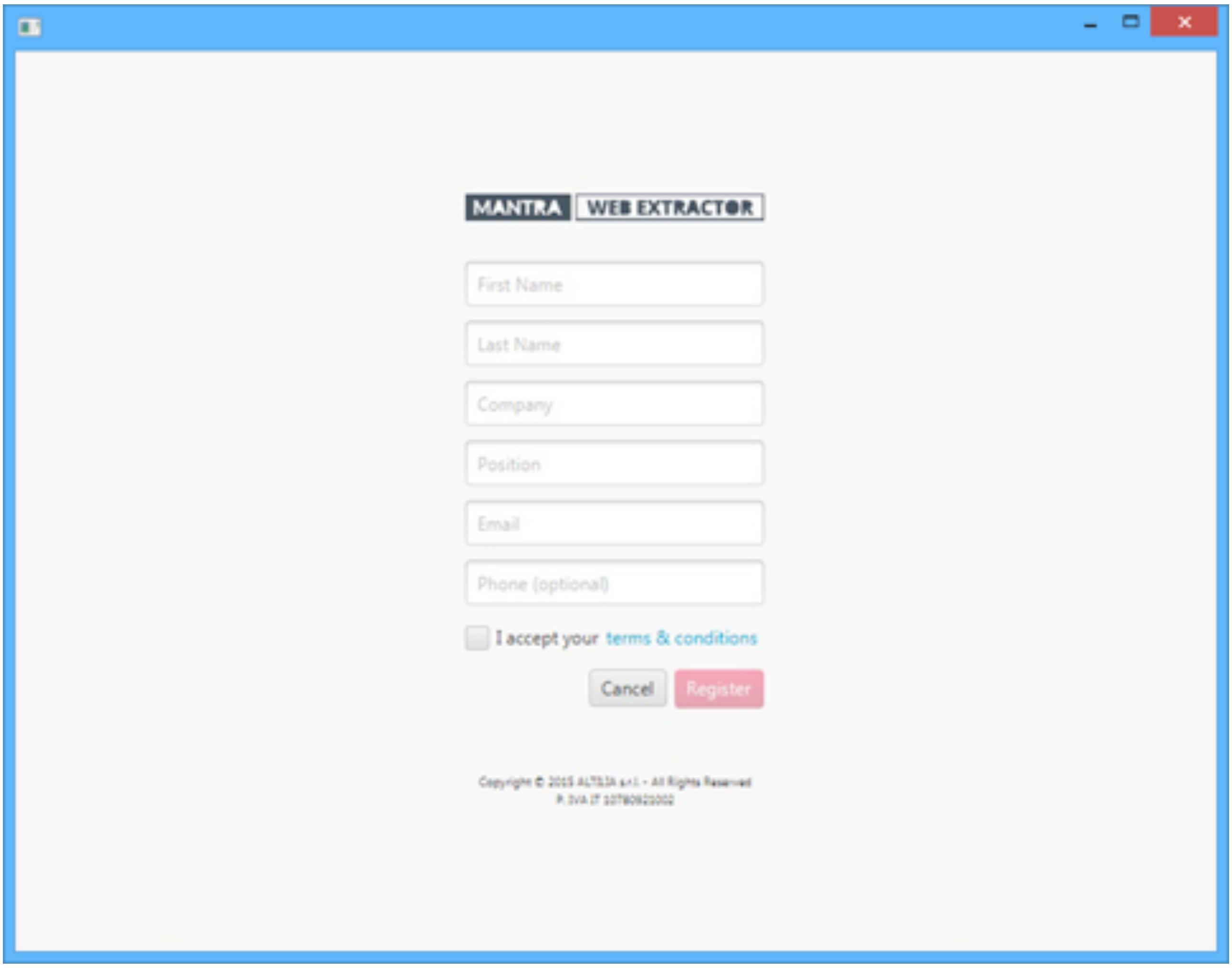

MANTRA WEB EXTRACTOR
First Name
Last Name
Company
Position
Email
Phone (optional)
☐ I accept your terms & conditions
Cancel
Register

Figure 2: Register form

If you forgot your password, you can recover it by clicking on the dedicated/ appropriate button from the login form. A form will be displayed and you will have to enter the e-mail address associated with your MWE account, then click on *Continue*. We will send you an email containing your new password.

CHAPTER 2 – BASICS

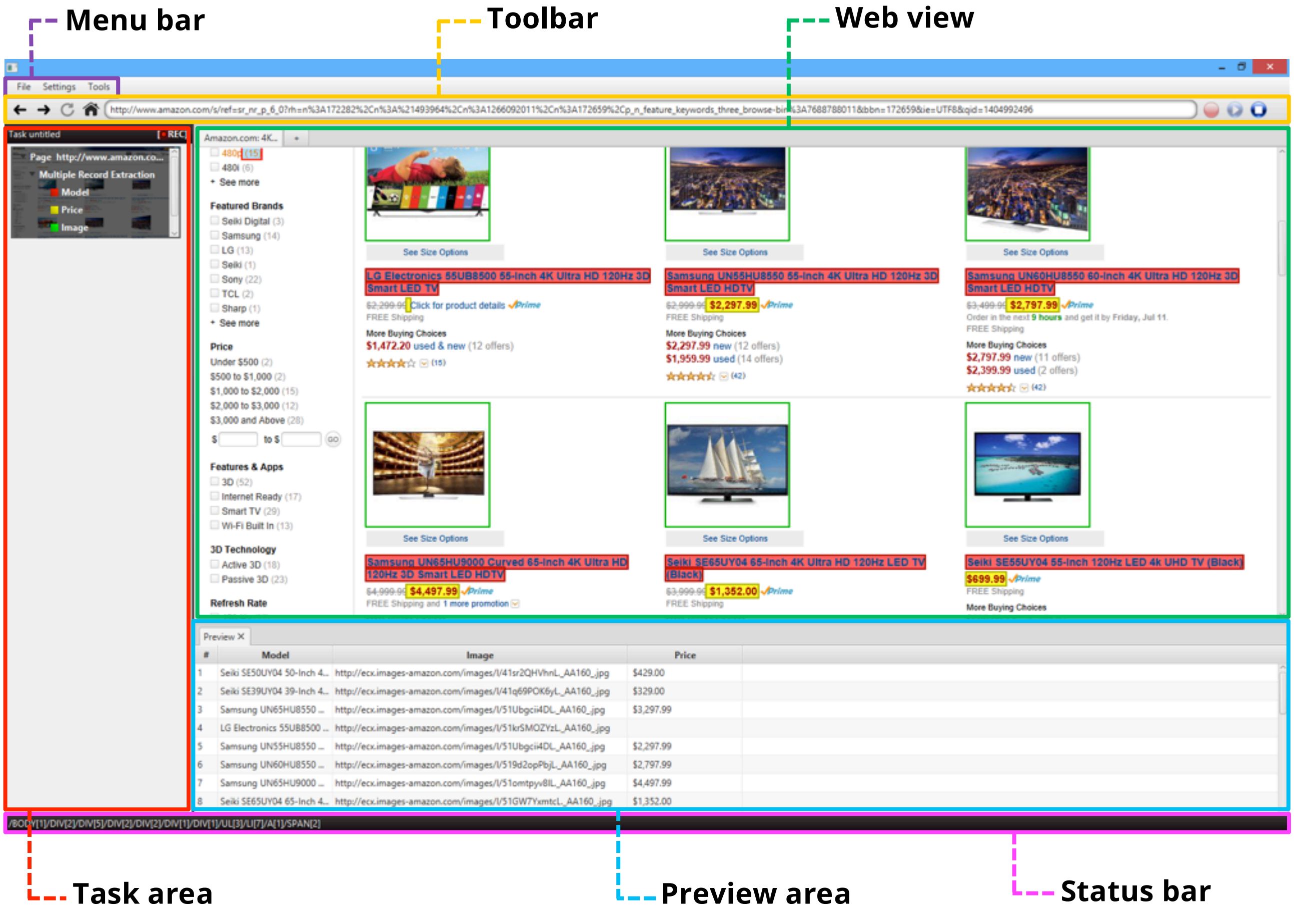
MANTRA Web Extractor looks and works like a common browser with addictional specific functions that allow you to record and run processes of extraction from the web. There are two distinct phases of operation:

1. **Recording:** during this phase the MWE automatically records all navigation actions that you may perform browsing a website. In addition, you can define specific extraction actions through a graphical selection of the contents you want to extract. A web extraction process is a set of these actions and can be stored in a file with ".wep" extension.
2. **Running:** during this phase the MWE automatically performs both navigation and extraction actions recorded in the recording phase. The results of the extraction are stored in a XML file.

In the next sections, the tool will be explainedin detail in all its aspects. In Chapter 3 we will talk about the workspace. In Chapter 4 all the functions concerning the record of a task will be introduced. Although these functions are very intuitive, we used an example-driven approach to explain them for a more immediate understanding. In Chapter 5 we will show you you the steps you have to follow if you want to edit a previously recorded task. In Chapter 6 we will explain to you how to run a task and we will give you an example of extraction result. Finally, the last two chapters respectively list the keyboard shortcuts that you can invoke and the possible solutions to the main problems that may occur.

CHAPTER 3 - WORKSPACE

The MWE workspace is thought to help you focus on creating, editing and running web extraction processes. The workspace includes a menu bar, a toolbar, a web view, a status bar, a task area and a preview area.


Menu bar
Toolbar
Web view
File
Settings
Tools
Task untitled
REC
Page http://www.amazon.co...
Multiple Record Extraction
Model
Price
Image
Preview
Task area
Preview area
Status bar

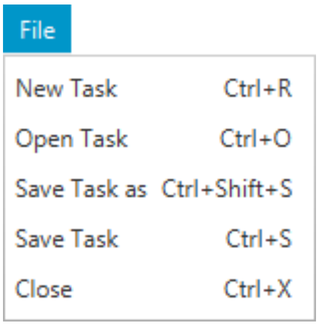
#	Model	Image	Price
1	Seiki SE50UY04 50-Inch 4...	http://ecx.images-amazon.com/images/I/41sr2QHVhnL._AA160_.jpg	$429.00
2	Seiki SE39UY04 39-Inch 4...	http://ecx.images-amazon.com/images/I/41q69POK6yL._AA160_.jpg	$329.00
3	Samsung UN65HU8550 ...	http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg	$3,297.99
4	LG Electronics 55UB8500 ...	http://ecx.images-amazon.com/images/I/51krSMOZYzL._AA160_.jpg	
5	Samsung UN55HU8550 ...	http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg	$2,297.99
6	Samsung UN60HU8550 ...	http://ecx.images-amazon.com/images/I/519d2opPbjL._AA160_.jpg	$2,797.99
7	Samsung UN65HU9000 ...	http://ecx.images-amazon.com/images/I/51omtpyv8IL._AA160_.jpg	$4,497.99
8	Seiki SE65UY04 65-Inch 4...	http://ecx.images-amazon.com/images/I/51GW7YxmtcL._AA160_.jpg	$1,352.00

/BODY[1]/DIV[2]/DIV[5]/DIV[2]/DIV[2]/DIV[1]/DIV[1]/UL[3]/LI[7]/A[1]/SPAN[2]

Figure 3: MWE workspace

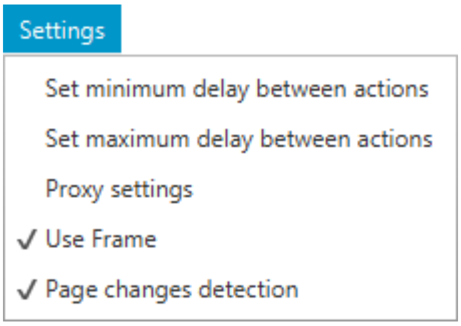
MENU BAR

The menu bar includes:


File
New Task Ctrl+R
Open Task Ctrl+O
Save Task as Ctrl+Shift+S
Save Task Ctrl+S
Close Ctrl+X

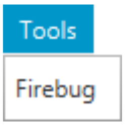
- **Menu File** This menu contains the necessary commands to create a new task, to open an existing one, to save a task to disk and to close the software.
- **Menu Settings** This menu contains the following commands:

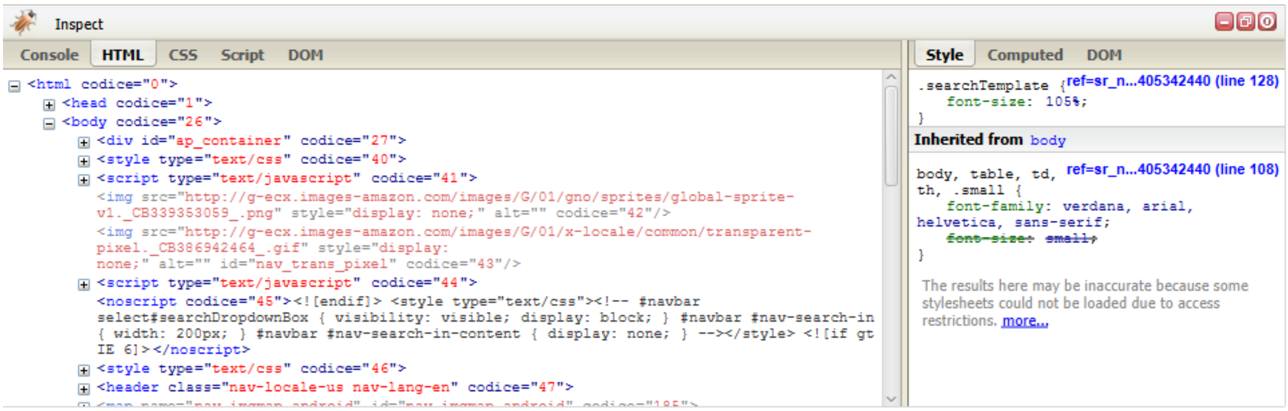
- *Set minimum/maximum delay between actions* that allow you to set, respectively, a minimum and a maximum time, in seconds, that must be waited before performing an action. The real delay will be a random time between the minimum and maximum delay value. This might be useful when the website from which you want to extract adopts mechanisms to avoid automatic extractions. The set values are persistent, so you do not need to set them at every restart.


Settings
Set minimum delay between actions
Set maximum delay between actions
Proxy settings
✓ Use Frame
✓ Page changes detection

- *Proxy settings* that allow you to set a proxy. The set values are persistent, so you do not need to set them at every restart.
- *Use Frame* that, if selected, allows you to extract elements from frames of a web page. The set option is persistent, so you do not need to set them at every restart. It is selected by default.
- *Page changes detection* that, if selected, allows you to detect the possible web page changes that might occur during the task recording. It should always be selected. Uncheck it only if you notice performance problems.It is selected by default . The set option is persistent, so you do not need to set them at every restart.

❖ **Menu Tools** This menu only contains the command *Firebug* that allow you to show the Firebug inspector.





```
<html codice="0">
    <head codice="1">
    <body codice="26">
        <div id="ap_container" codice="27">
        <style type="text/css" codice="40">
        <script type="text/javascript" codice="41">
        <img src="http://g-ecx.images-amazon.com/images/G/01/gno/sprites/global-sprite-
        v1._CB339353059_.png" style="display: none;" alt="" codice="42"/>
        <img src="http://g-ecx.images-amazon.com/images/G/01/x-locale/common/transparent-
        pixel._CB386942464_.gif" style="display:
        none;" alt="" id="nav_trans_pixel" codice="43"/>
        <script type="text/javascript" codice="44">
        <noscript codice="45"><![endif]> <style type="text/css"><!-- #navbar
        select#searchDropdownBox { visibility: visible; display: block; } #navbar #nav-search-in
        { width: 200px; } #navbar #nav-search-in-content { display: none; } --></style> <![if gt
        IE 6]></noscript>
        <style type="text/css" codice="46">
        <header class="nav-locale-us nav-lang-en" codice="47">
```

```
.searchTemplate {ref=sr_n...405342440 (line 128)
    font-size: 105%;
}
```

Inherited from body

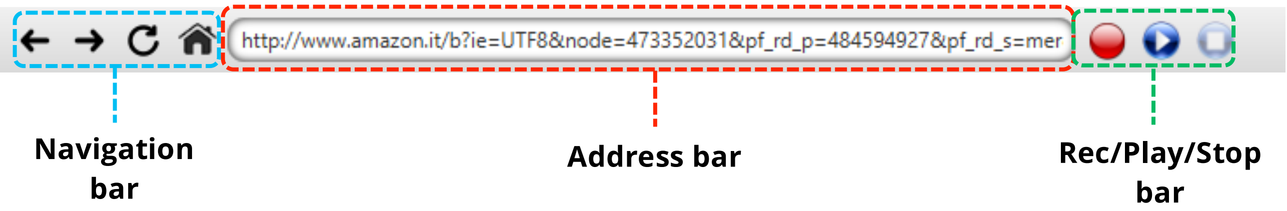
```
body, table, td, ref=sr_n...405342440 (line 108)
th, .small {
    font-family: verdana, arial,
helvetica, sans-serif;
    font-size: small;
}
```

The results here may be inaccurate because some stylesheets could not be loaded due to access restrictions. more...

Figure 4: Firebug inspector

TOOLBAR

The toolbar includes the typical commands of a browser web with in addition three commands used to record, run or stop a web extraction process.



http://www.amazon.it/b?ie=UTF8&node=473352031&pf_rd_p=484594927&pf_rd_s=mer

Navigation bar:

- ← **Back button** To navigate backwards the history of visited web pages.
- → **Forward button** To navigate forward the history of visited web pages.
- ↻ **Refresh button** To reload the current web page.
- 🏠 **Home button** To load the home web page, which is google by default.

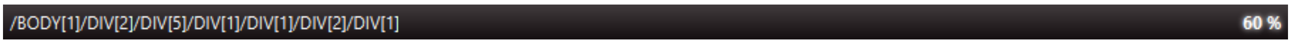
Address bar: shows the current URL and accepts a typed URL that navigates you to a chosen website.

Rec/Play/Stop bar:

- ● **Rec button** To start a task recording.
- ▶ **Play button** To run a task.
- ⏸ **Pause button** To pause the running task.
- ■ **Stop button** To stop the recording/running task.

STATUS BAR

The status bar is located at the bottom of the main window. It shows the xpath of the item in the web page over which the mouse is pointed. During the page load, it also shows the loadingpercentage.



/BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV[2]/DIV[1] 60 %

Figure 5: Status bar

WEB VIEW

The web view displays the loaded web page. In addition, when compared to a common browser, in this area you can graphically select the contents to be extracted/ you want to extract.

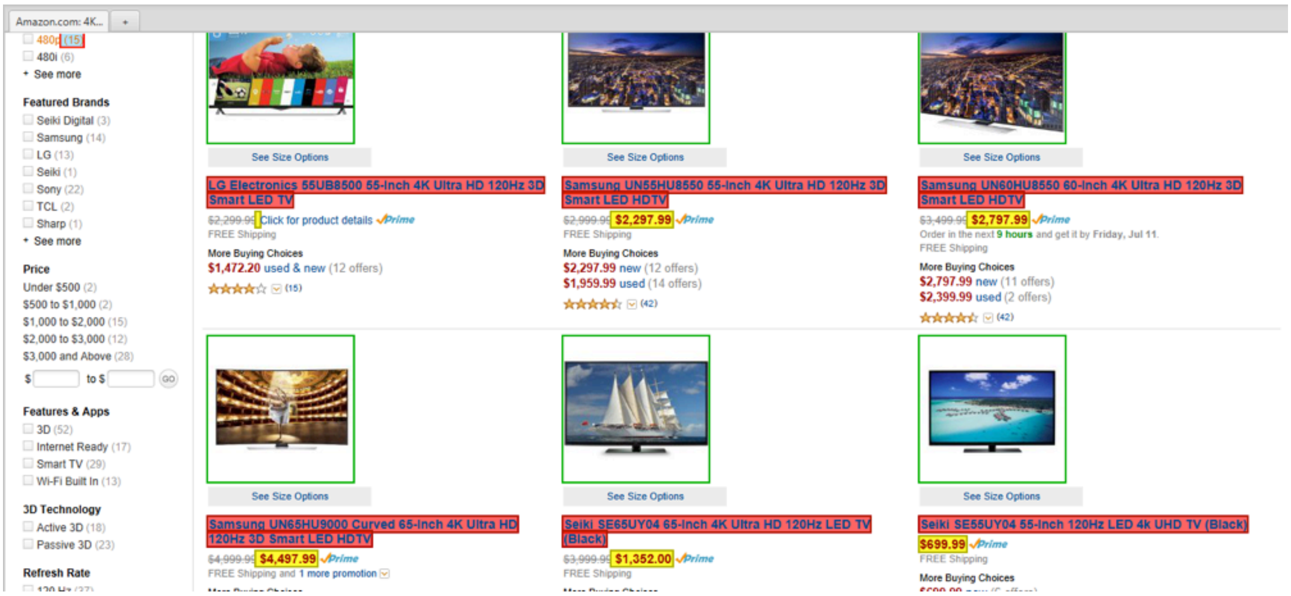

Amazon.com: 4K...
480p (15)
480i (6)
See more
Featured Brands
Seiki Digital (3)
Samsung (14)
LG (13)
Seiki (1)
Sony (22)
TCL (2)
Sharp (1)
See more
Price
Under $500 (2)
$500 to $1,000 (2)
$1,000 to $2,000 (15)
$2,000 to $3,000 (12)
$3,000 and Above (28)
Features & Apps
3D (52)
Internet Ready (17)
Smart TV (29)
Wi-Fi Built In (13)
3D Technology
Active 3D (18)
Passive 3D (23)
Refresh Rate
See Size Options
LG Electronics 55UB8500 55-Inch 4K Ultra HD 120Hz 3D Smart LED TV
$2,299.99 Click for product details Prime
Samsung UN55HU8550 55-Inch 4K Ultra HD 120Hz 3D Smart LED HDTV
$2,999.99 $2,297.99 Prime
Samsung UN60HU8550 60-Inch 4K Ultra HD 120Hz 3D Smart LED HDTV
$3,499.99 $2,797.99 Prime
Samsung UN65HU9000 Curved 65-Inch 4K Ultra HD 120Hz 3D Smart LED HDTV
$4,999.99 $4,497.99 Prime
Seiki SE65UY04 65-Inch 4K Ultra HD 120Hz LED TV (Black)
$3,999.99 $1,352.00 Prime
Seiki SE55UY04 55-Inch 120Hz LED 4k UHD TV (Black)
$699.99 Prime

Figure 6: Web view with some items selected to be extracted

TASK AREA

This area shows the current task loaded or in recording. A task is displayed as a sequence of blocks. A block represents a web page (so we will call it page action) and contains the extraction actions recorded on that page, organized as a tree. For example, Figure 7 shows a task recorded on two pages of Amazon.com. The first page action contains the actions recorded in order to fill a search form and start the search. The second page action contains the actions recorded in order to extract the model name, the price and the image of the products from the resulting page.

The top bar contains the task name (*untitled* by default) and a label that indicates when a task is in recording.

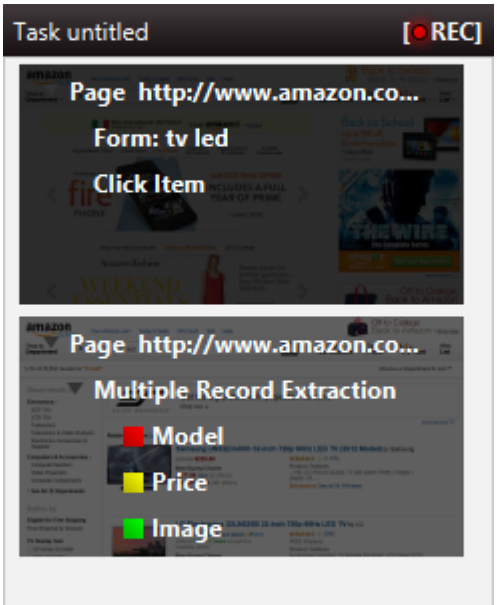
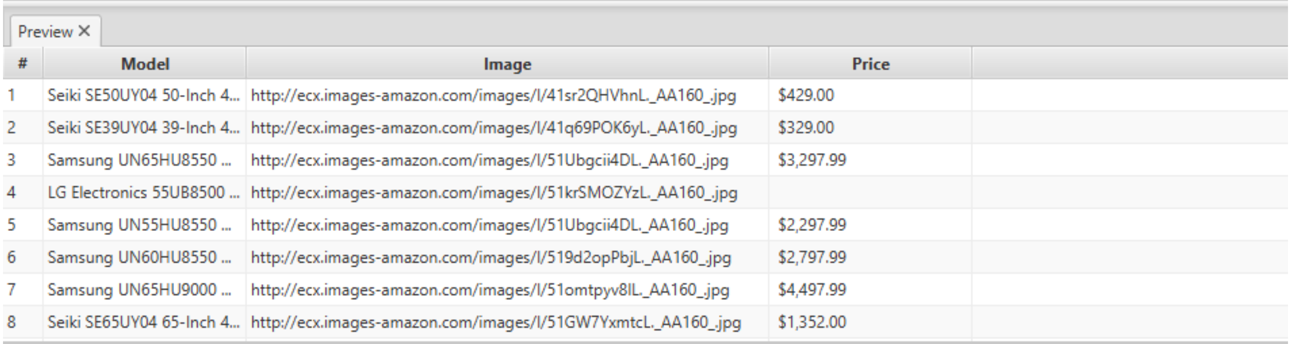

Task untitled
[● REC]
Page http://www.amazon.co...
Form: tv led
Click Item
Page http://www.amazon.co...
Multiple Record Extraction
🟥 Model
🟨 Price
🟩 Image

Figure 7: Task area

PREVIEW AREA

This area shows in a table a preview of the information extracted from the current web page. The columns represent the attributes defined during the recording of the task, while the rows are the values extracted. This values are properly aligned, so a row contains the data items relative to the same date record.



#	Model	Image	Price
1	Seiki SE50UY04 50-Inch 4...	http://ecx.images-amazon.com/images/I/41sr2QHVhnL._AA160_.jpg	$429.00
2	Seiki SE39UY04 39-Inch 4...	http://ecx.images-amazon.com/images/I/41q69POK6yL._AA160_.jpg	$329.00
3	Samsung UN65HU8550 ...	http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg	$3,297.99
4	LG Electronics 55UB8500 ...	http://ecx.images-amazon.com/images/I/51krSMOZYzL._AA160_.jpg	
5	Samsung UN55HU8550 ...	http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg	$2,297.99
6	Samsung UN60HU8550 ...	http://ecx.images-amazon.com/images/I/519d2opPbjL._AA160_.jpg	$2,797.99
7	Samsung UN65HU9000 ...	http://ecx.images-amazon.com/images/I/51omtpyv8IL._AA160_.jpg	$4,497.99
8	Seiki SE65UY04 65-Inch 4...	http://ecx.images-amazon.com/images/I/51GW7YxmtcL._AA160_.jpg	$1,352.00

Figure 8: Preview area

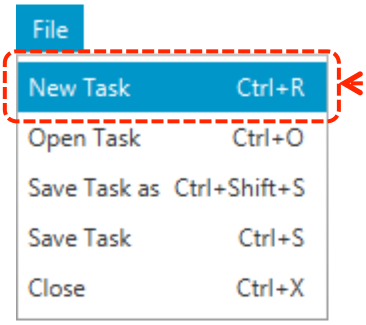
CHAPTER 4 – RECORDING A TASK

In this section, we will use an example-driven approach to introduce you all the functions concerning the record of a task.

CREATE A NEW TASK

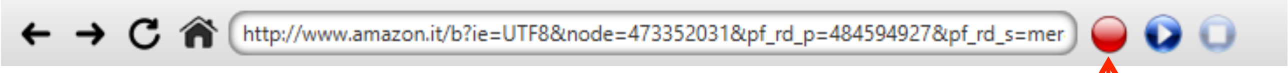
The task creation coincides with the start of the record phase. There are three ways to do this:

1. File -> New Task


File
New Task Ctrl+R
Open Task Ctrl+O
Save Task as Ctrl+Shift+S
Save Task Ctrl+S
Close Ctrl+X

Create a new task and start the record phase

2. Click on the *record button* on the toolbar

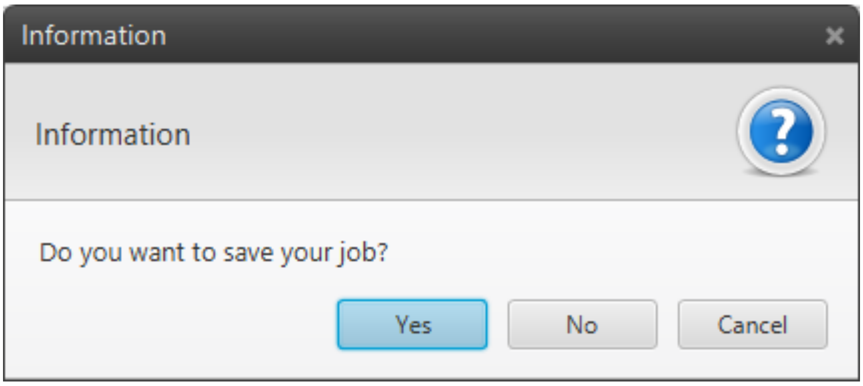


http://www.amazon.it/b?ie=UTF8&node=473352031&pf_rd_p=484594927&pf_rd_s=mer

Record button

3. Press Ctrl + R

If a task is already loaded a message will be displayed to ask you if you want to save the current work.


Information
Information
Yes
No
Cancel

Do you want to save your job?

Choose:

- *Yes* to save the current task and start the new task recording.
- *No* to discard the current task and start the new task recording.
- *Cancel* to abort the operation of task creation.

After starting the recording of a new task an empty page action is added in the task area, representing the current web page. Now, you are able to record both navigation and extraction actions. In the next sections, you will learn what you can do.

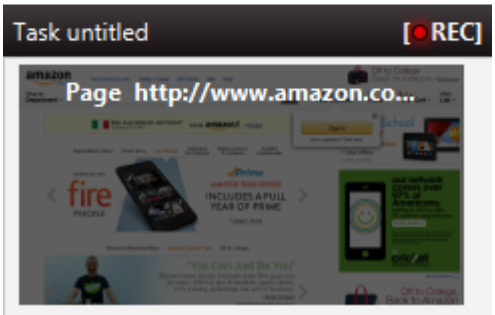


Figure 9: The empty page action recorded after the task creation

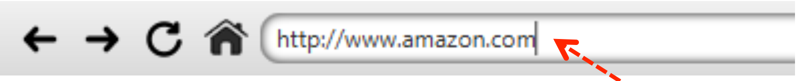
RECORD NAVIGATION ACTIONS

A navigation action is an interaction with a web page (for example, a log on, a mouse click or a form filling). The MWE offers you the opportunity to record such a series of actions and then to replay them later to reach the same web page destination. This recording is simply achieved with a standard browsing experience.

Record a web page load

To record a web page load:

1. Type the URL in the address bar



Example

Type the amazon.com home page address

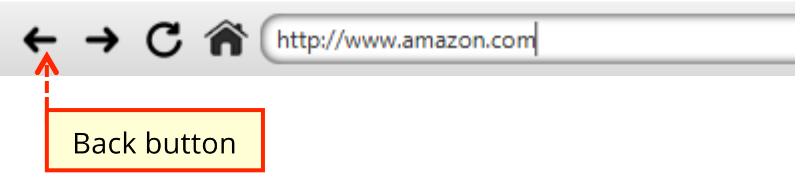
2. Press

At the end of the web page loading, a new empty page action is added in the task area, representing the new web page.

Record a back action

To record a back action:

1. Click the back button on the toolbar



As a result, a Back action is recorded in the page action of the starting web page. In this case no page action will be added in the task area, but the controls will return to the previous page action.

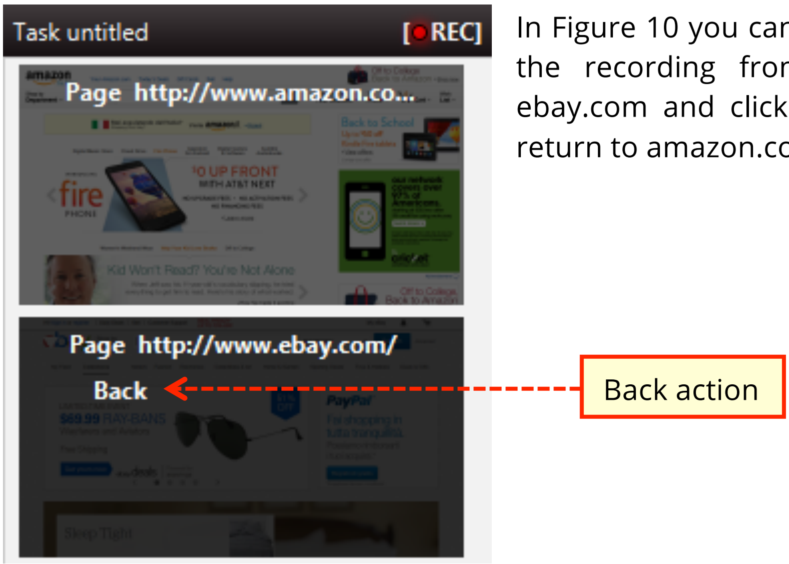


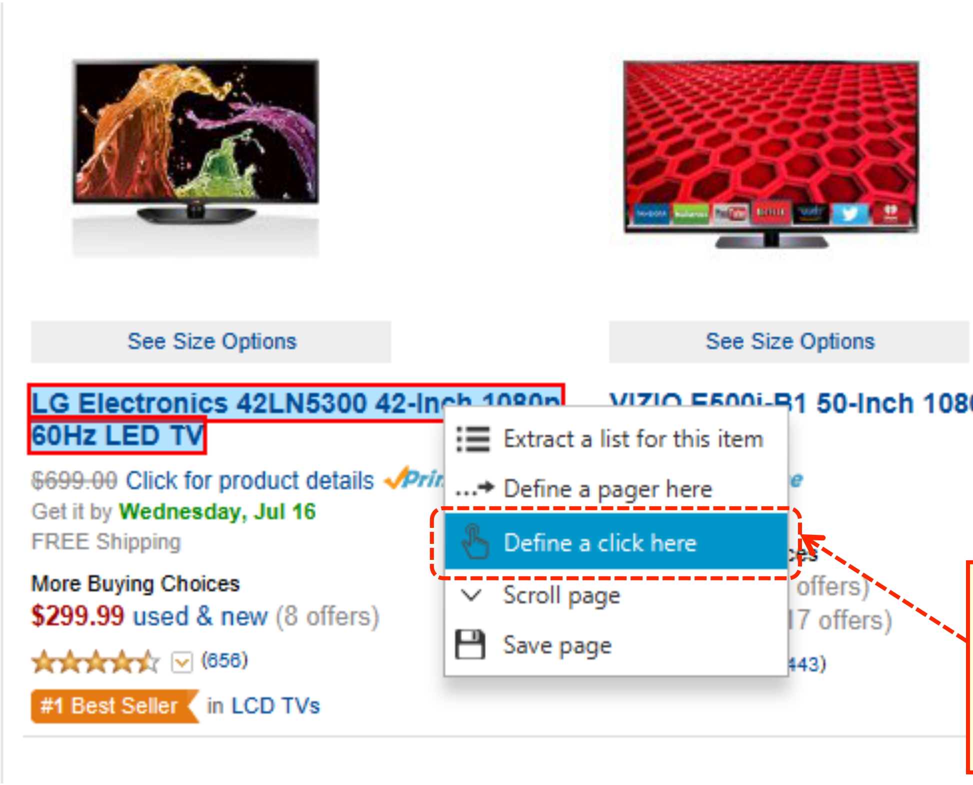
Figure 10: Back action recording example

In Figure 10 you can see the resulting task after starting the recording from amazon.com home page, load ebay.com and click the back button in the toolbar to return to amazon.com.

Record a click action

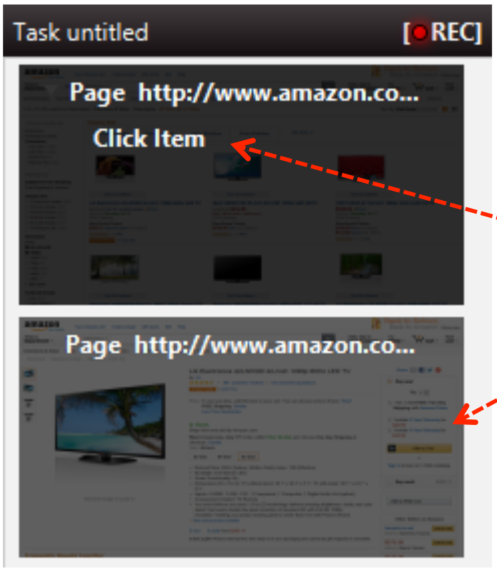
There are two ways to record a click:

1. Simple click on a web page element.
2. Right click on the element of the web page on which you want to record a click and select the command *Define a click here*.


See Size Options
See Size Options
LG Electronics 42LN5300 42-Inch 1080p 60Hz LED TV
VIZIO E500i-B1 50-Inch 1080p Smart LED
$699.00 Click for product details
Get it by Wednesday, Jul 16
FREE Shipping
More Buying Choices
$299.99 used & new (8 offers)
(656)
#1 Best Seller in LCD TVs
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

Select this command to define a click manually

The first method is automatic, the MWE will detect your click and record it. The second method is manual. You can use it if the first method fails in the click recording. In both cases, a click action is recorded in the page action representing the web page containing the element clicked. In addition, if the click will load a new web page, a new empty page action is added in the task area.

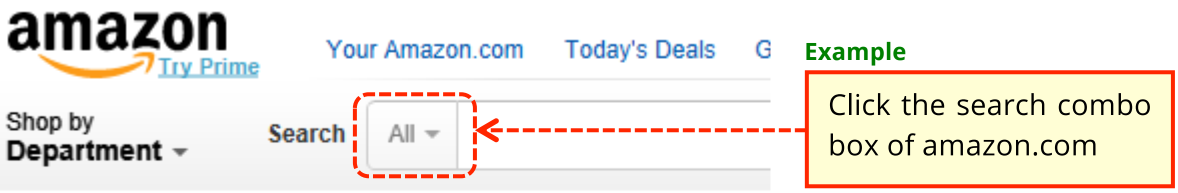

Task untitled
[REC]
Page http://www.amazon.co...
Click Item
Page http://www.amazon.co...

The click action recorded and the new empty page action added that represents the web page result of the click

Record an element selection from a combo box

To record an element selection from a combo box:

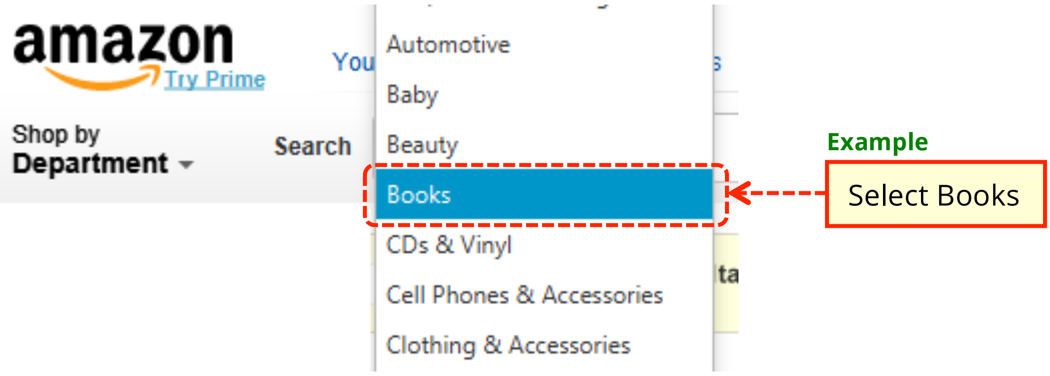
1. Click the combo box button


amazon
Try Prime
Your Amazon.com
Today's Deals
G
Shop by
Department
Search
All

Example

Click the search combo box of amazon.com

2. Select an element


amazon
Try Prime
You
Shop by
Department
Search
Automotive
Baby
Beauty
Books
CDs & Vinyl
Cell Phones & Accessories
Clothing & Accessories

Example

Select Books

Both the click and combo selection actions are added in the page action, as shown in Figure 11. If the selection will load a new web page, a new empty page action is added in the task area.

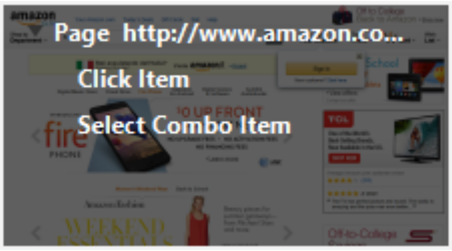
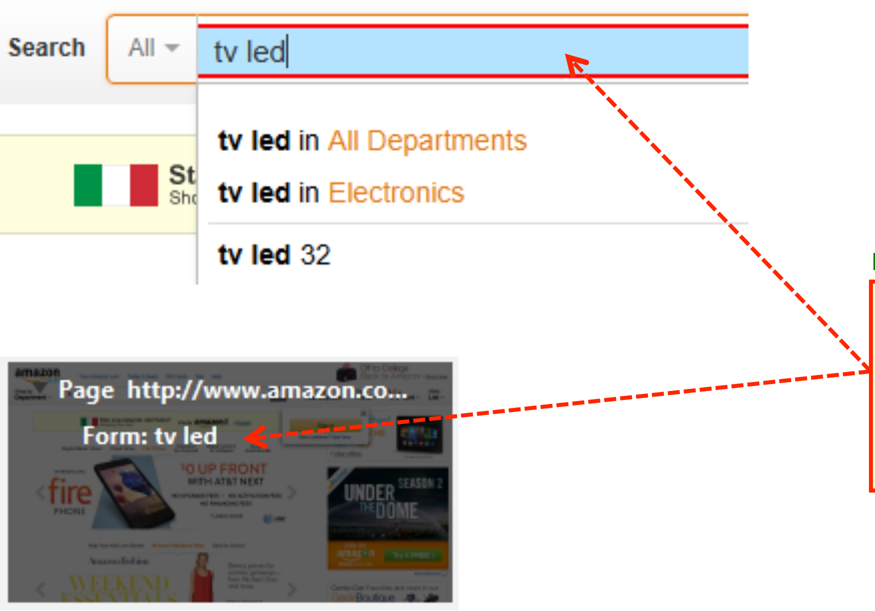

Page http://www.amazon.co...
Click Item
Select Combo Item

Figure 11: Combo box selection action

Record a form filling

To record a form filling type a text in a form, the MWE automatically detects the input element where you are writing and the text inserted.


Search
All
tv led
tv led in All Departments
tv led in Electronics
tv led 32
Page http://www.amazon.co...
Form: tv led

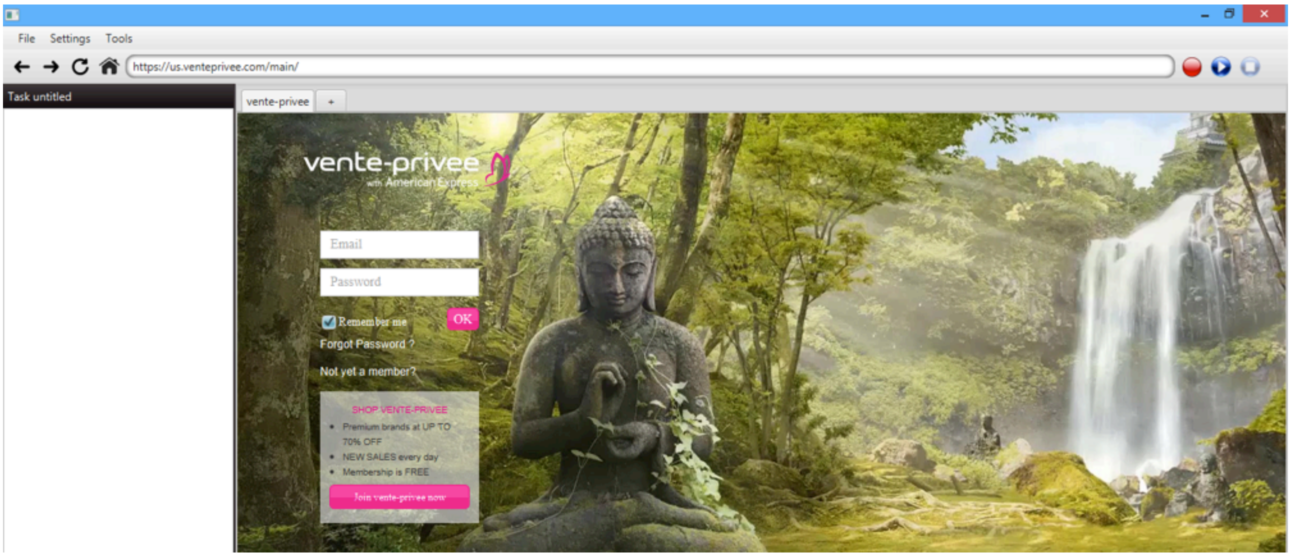
Example

Type “tv led” in the search form of amazon. A form filling action is added in the page action

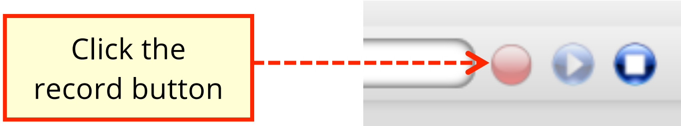
Example of use: record a login

In this section, you can see how to combine the previous actions to record a complete login to a website. Suppose for example, you want to extract data from an outlet website like *venteprivee.com* where the login is required to see the offered products. This is the procedure you have to follow:

1. Type the address *https://us.venteprivee.com/main/* in the address bar and press ENTER.

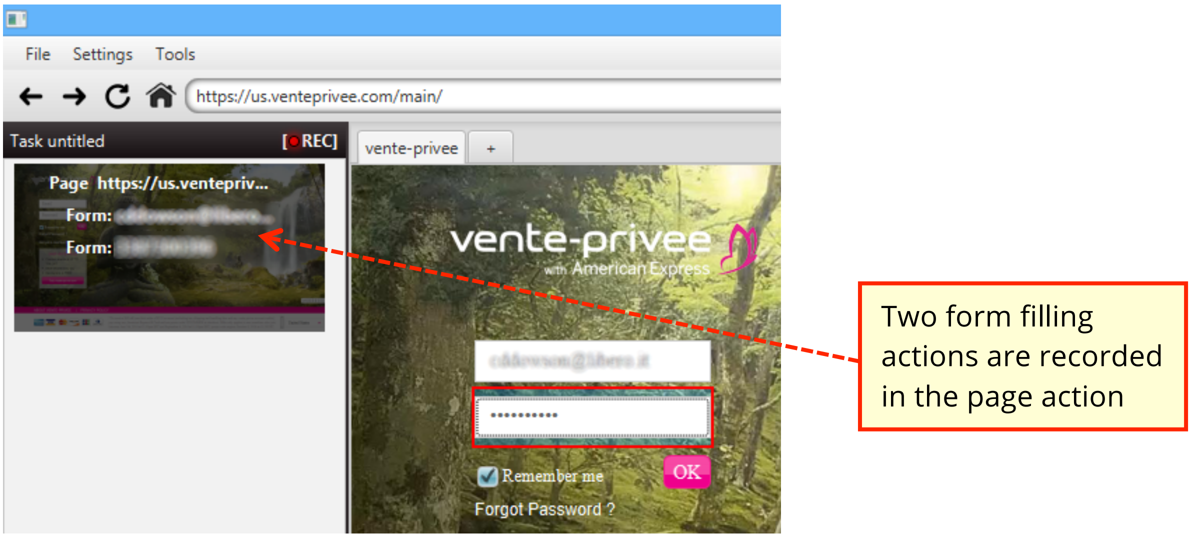

File
Settings
Tools
https://us.venteprivee.com/main/
Task untitled
vente-privee
+
vente-privee
with American Express
Email
Password
Remember me
OK
Forgot Password ?
Not yet a member?
SHOP VENTE-PRIVEE
Premium brands at UP TO 70% OFF
NEW SALES every day
Membership is FREE
Join vente-privee now

2. Start recording



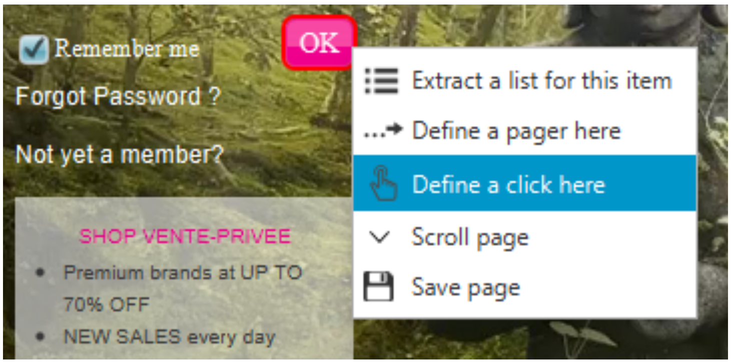
Click the record button

3. Insert username e password


File
Settings
Tools
https://us.venteprivee.com/main/
Task untitled
[REC]
vente-privee
+
Page https://us.ventepriv...
Form: [redacted]
Form: [redacted]
vente-privee
with American Express
Remember me
OK
Forgot Password ?

Two form filling actions are recorded in the page action

4. Record a click on OK button


Remember me
OK
Forgot Password ?
Not yet a member?
SHOP VENTE-PRIVEE
Premium brands at UP TO 70% OFF
NEW SALES every day
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

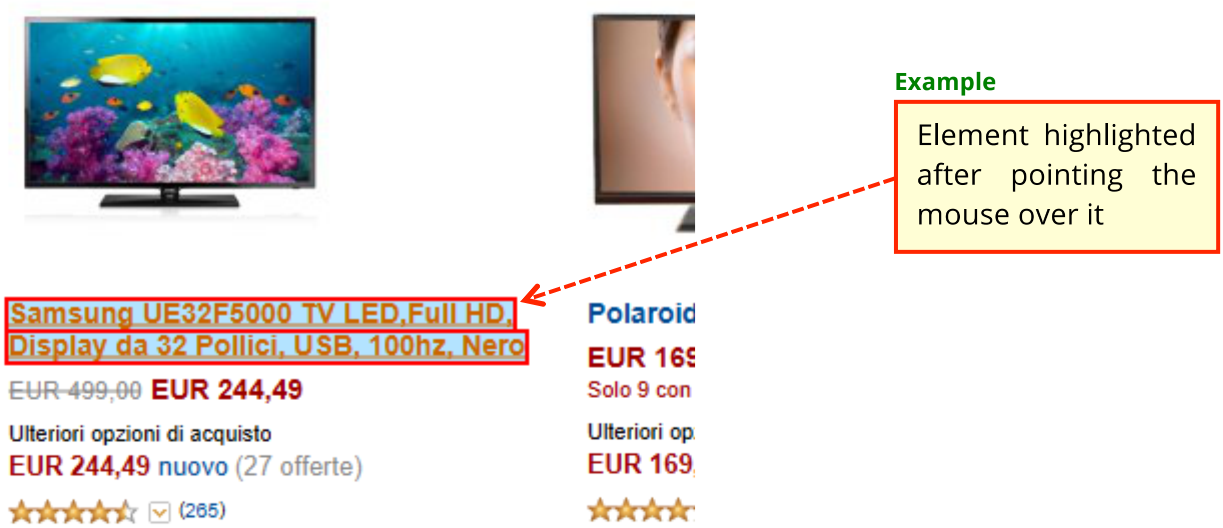
If you stop the recording now, logout and press the play button, the MWE will perform the login automatically.

RECORD EXTRACTION ACTIONS

In order to extract data from a website, the tool offers you some features to record complete extraction patterns.

Extraction of a list of items

This function allows you to select a list of items to be extracted from a web page. In recording, when you point the mouse over an element, this will be highlighted in a blue rectangle with red edges.

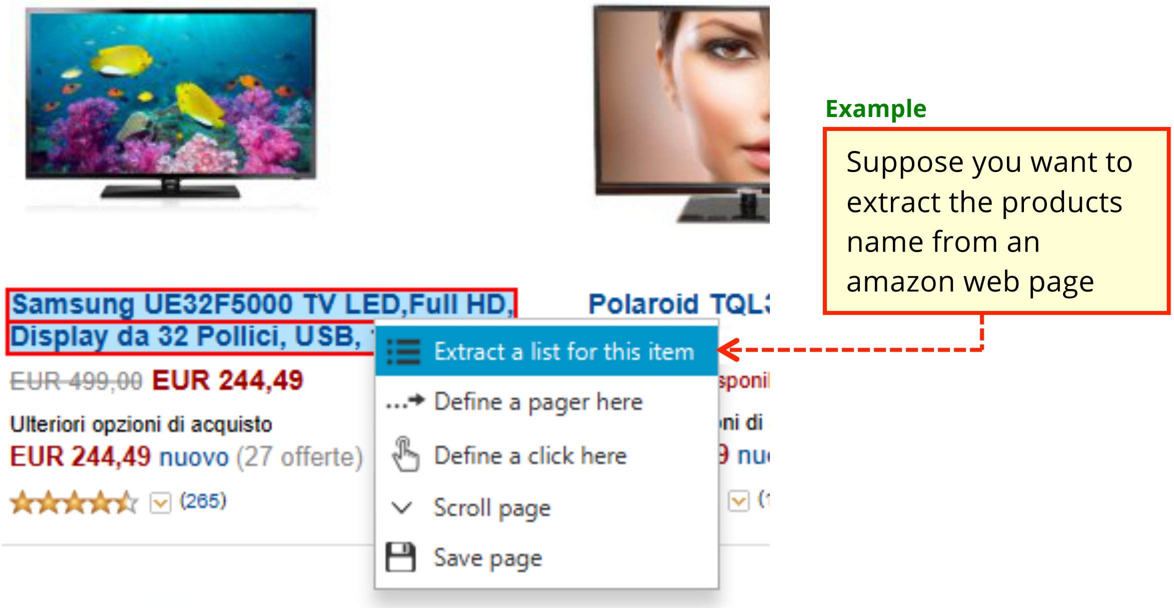

Samsung UE32F5000 TV LED,Full HD, Display da 32 Pollici, USB, 100hz, Nero
EUR 499,00 EUR 244,49
Ulteriori opzioni di acquisto
EUR 244,49 nuovo (27 offerte)
(265)
Polaroid
EUR 169
Solo 9 con
Ulteriori op
EUR 169

Example

Element highlighted after pointing the mouse over it

To extract a list from the selected current element, :

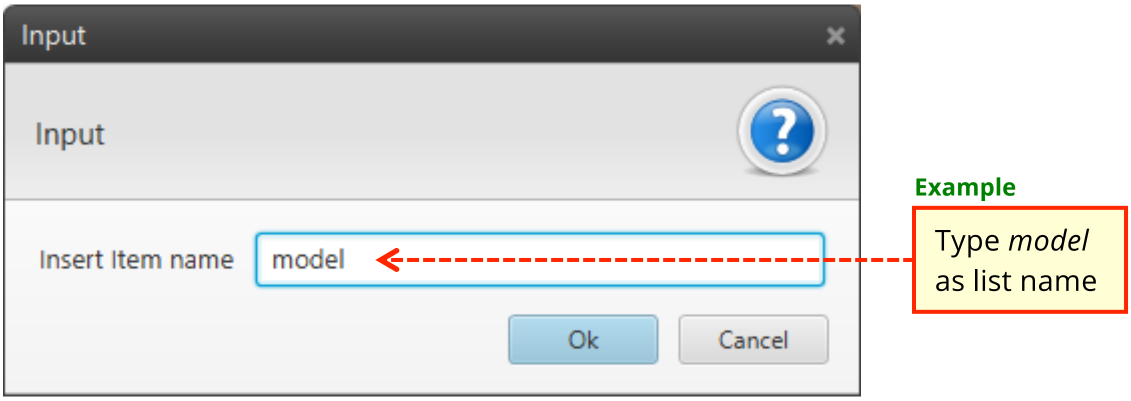
1. Right click on the element from which you want to extract a list and select *Extract a list for this item*


Samsung UE32F5000 TV LED,Full HD, Display da 32 Pollici, USB,
EUR 499,00 EUR 244,49
Ulteriori opzioni di acquisto
EUR 244,49 nuovo (27 offerte)
(265)
Polaroid TQL
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

Example

Suppose you want to extract the products name from an amazon web page

2. Insert a name to identify the list



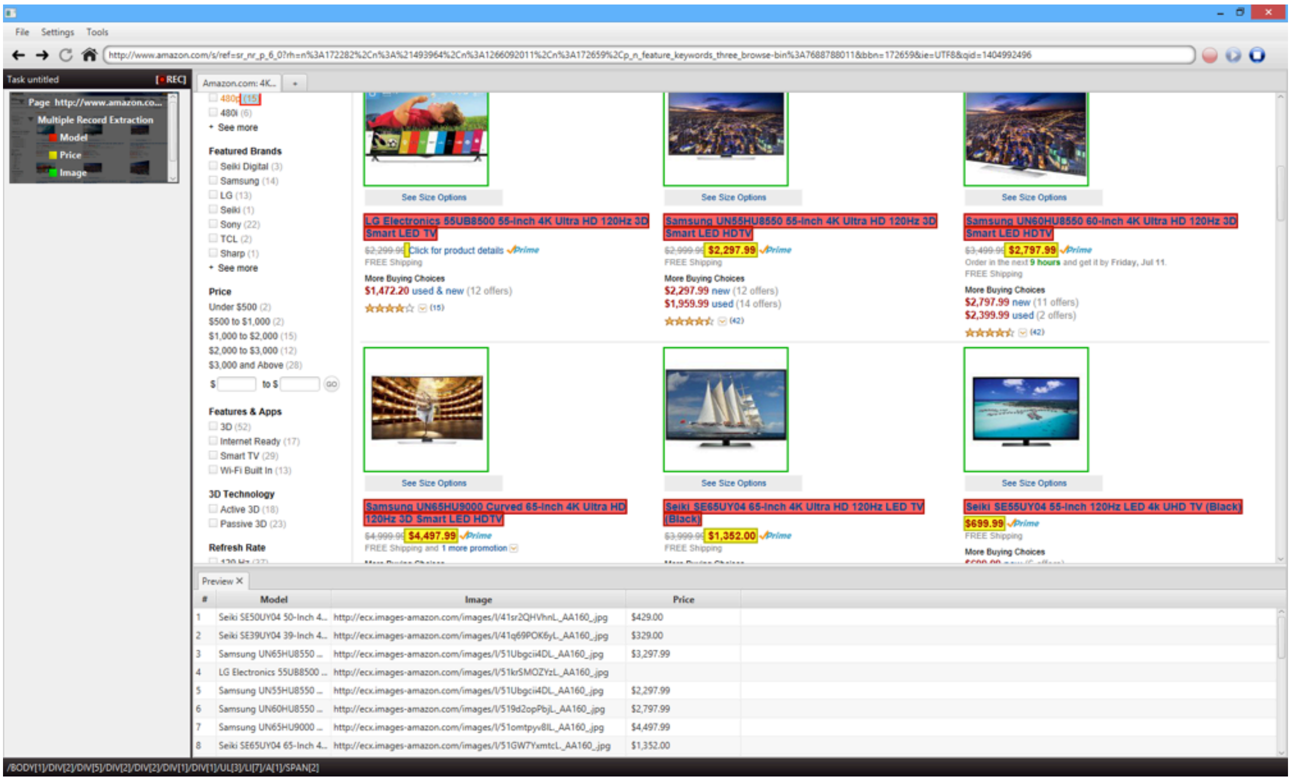
Insert Item name: model

Example

Type *model* as list name

The tool automatically recognizes all other analogue elements in the web page and shows a preview table where each row contain a record and each column represents items of the same type.

Figure 12 shows what you obtain after recording the extraction of three list of items from a TV web page: the names of the models, the prices and the image URLs. Each list is marked with a different colour. In this specific case, red for the models, yellow for the prices, and green for the image URLs.



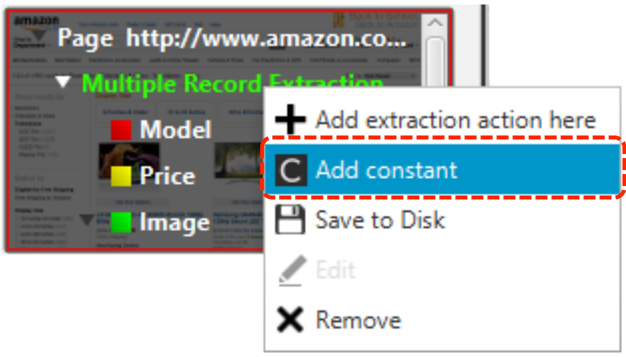
#	Model	Image	Price
1	Seiki SE50UY04 50-Inch 4...	http://ecx.images-amazon.com/images/I/41sr2QHVhnL._AA160_.jpg	$429.00
2	Seiki SE39UY04 39-Inch 4...	http://ecx.images-amazon.com/images/I/41q69POK6yL._AA160_.jpg	$329.00
3	Samsung UN65HU8550 ...	http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg	$3,297.99
4	LG Electronics 55UB8500 ...	http://ecx.images-amazon.com/images/I/51krSMOZYzL._AA160_.jpg	
5	Samsung UN55HU8550 ...	http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg	$2,297.99
6	Samsung UN60HU8550 ...	http://ecx.images-amazon.com/images/I/519d2opPbjL._AA160_.jpg	$2,797.99
7	Samsung UN65HU9000 ...	http://ecx.images-amazon.com/images/I/51omtpyv8IL._AA160_.jpg	$4,497.99
8	Seiki SE65UY04 65-Inch 4...	http://ecx.images-amazon.com/images/I/51GW7YxmtcL._AA160_.jpg	$1,352.00

/BODY[1]/DIV[2]/DIV[5]/DIV[2]/DIV[2]/DIV[1]/DIV[1]/UL[3]/LI[7]/A[1]/SPAN[2]

Figure 12: Extract three list of elements from an amazon web page

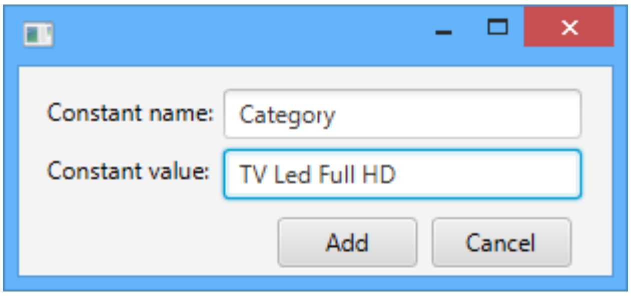
In the preview table each column coincides with a list, while the row contains the values from each list related to the same element. In the example above, a row contains the model name, the price and the URL image of a specific television. If an element is not present in the web page, for example a TV without price, the corresponding cell will be left blank. A particular list that you can record is a list consisting of constant values. This may be useful when you want to add some information that aren't present in the web pages. To add a list of constants to a Multiple Record Extraction:

1. Right click on the Multiple Record Extraction and select the command *Add constant*


Page http://www.amazon.co...
Multiple Record Extraction
Model
Price
Image
Add extraction action here
Add constant
Save to Disk
Edit
Remove

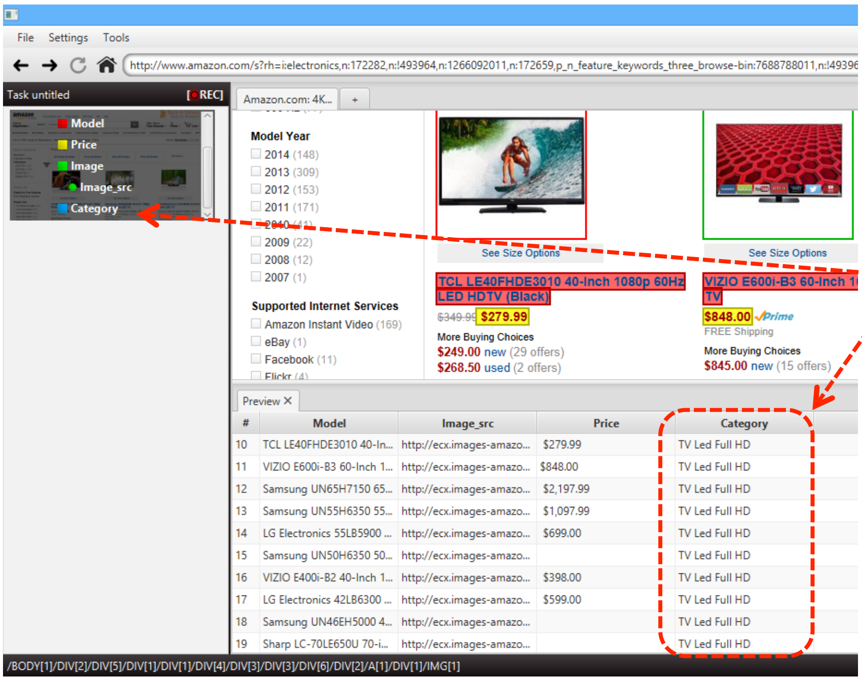
Add list of constants

2. In the dialog that will appear type the constant name and the constant value



Constant name: Category
Constant value: TV Led Full HD

A column of constant values will be added In the preview area

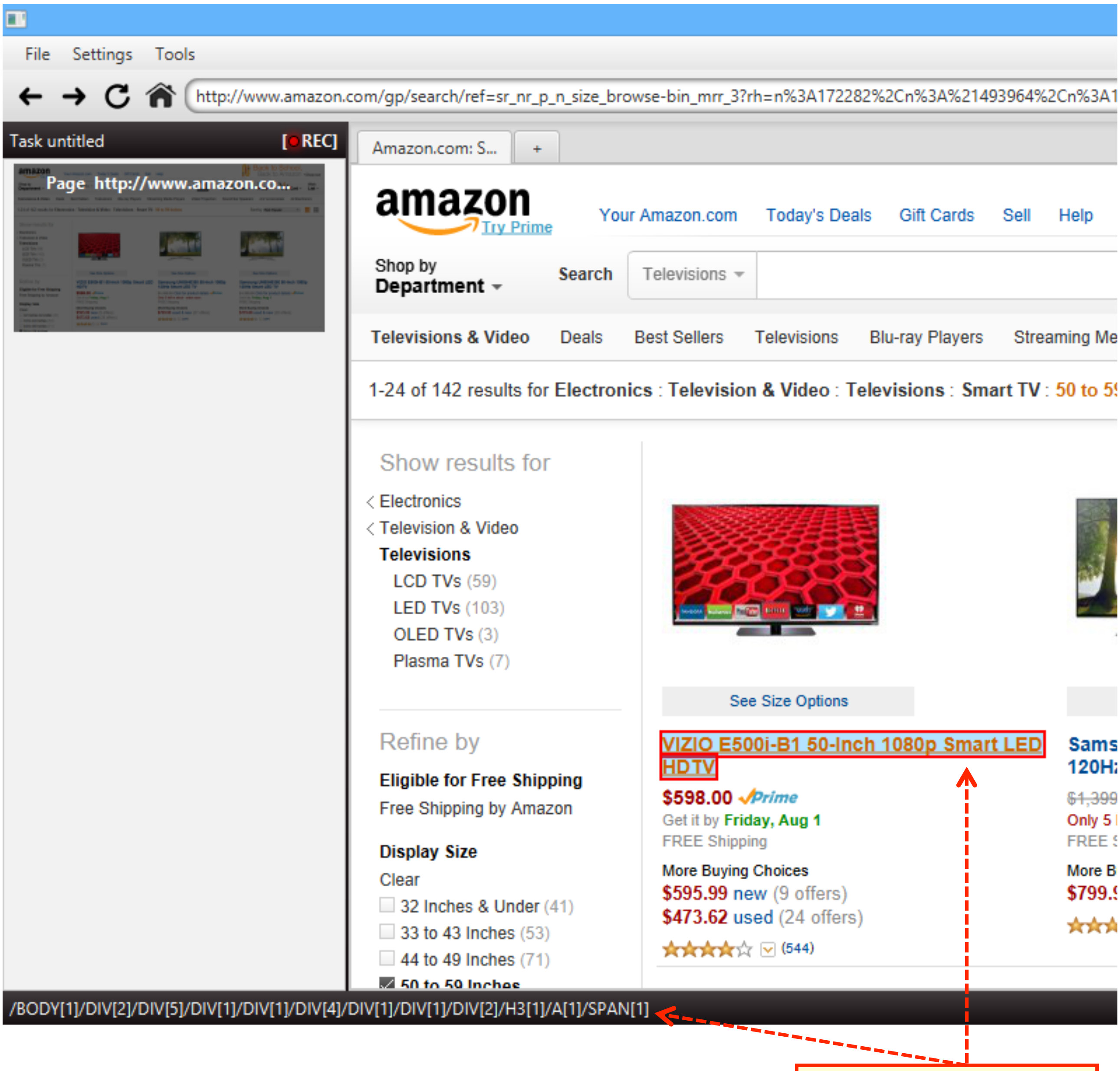

File Settings Tools
http://www.amazon.com/s?rh=i:electronics,n:172282,n:493964,n:1266092011,n:172659,p_n_feature_keywords_three_browse-bin:7688788011,n:49396
Task untitled
REC
Amazon.com: 4K...
Model
Price
Image
Image_src
Category
Model Year
2014 (148)
2013 (309)
2012 (153)
2011 (171)
2010 (41)
2009 (22)
2008 (12)
2007 (1)
Supported Internet Services
Amazon Instant Video (169)
eBay (1)
Facebook (11)
Flickr (4)
See Size Options
TCL LE40FHDE3010 40-Inch 1080p 60Hz LED HDTV (Black)
$349.99 $279.99
More Buying Choices
$249.00 new (29 offers)
$268.50 used (2 offers)
VIZIO E600i-B3 60-Inch 1 TV
$848.00 Prime
FREE Shipping
More Buying Choices
$845.00 new (15 offers)
Preview
/BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV[4]/DIV[3]/DIV[3]/DIV[6]/DIV[2]/A[1]/DIV[1]/IMG[1]

#	Model	Image_src	Price	Category
10	TCL LE40FHDE3010 40-In...	http://ecx.images-amazo...	$279.99	TV Led Full HD
11	VIZIO E600i-B3 60-Inch 1...	http://ecx.images-amazo...	$848.00	TV Led Full HD
12	Samsung UN65H7150 65...	http://ecx.images-amazo...	$2,197.99	TV Led Full HD
13	Samsung UN55H6350 55...	http://ecx.images-amazo...	$1,097.99	TV Led Full HD
14	LG Electronics 55LB5900 ...	http://ecx.images-amazo...	$699.00	TV Led Full HD
15	Samsung UN50H6350 50...	http://ecx.images-amazo...		TV Led Full HD
16	VIZIO E400i-B2 40-Inch 1...	http://ecx.images-amazo...	$398.00	TV Led Full HD
17	LG Electronics 42LB6300 ...	http://ecx.images-amazo...	$599.00	TV Led Full HD
18	Samsung UN46EH5000 4...	http://ecx.images-amazo...		TV Led Full HD
19	Sharp LC-70LE650U 70-i...	http://ecx.images-amazo...		TV Led Full HD

The result of the addition of a constant

Select specific elements to be extracted

In some cases, it may be difficult to select an item just by pointing the mouse over it. For example, suppose you want to select the anchor item that contains the link to the product from an amazon web page. This it's impossible because, when pointing the mouse over it, the underlying span item will be selected as shown below.

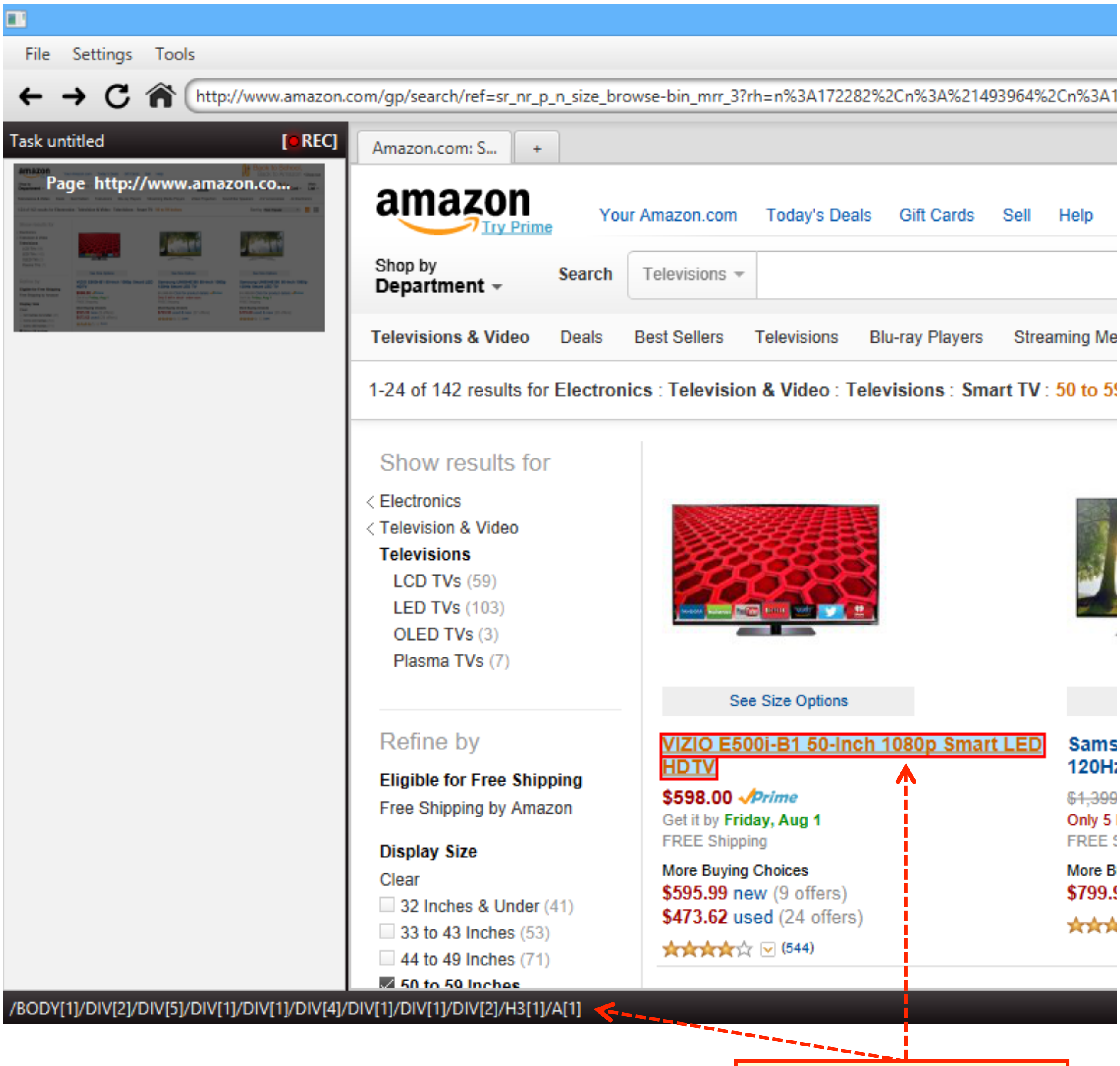

File
Settings
Tools
http://www.amazon.com/gp/search/ref=sr_nr_p_n_size_browse-bin_mrr_3?rh=n%3A172282%2Cn%3A%21493964%2Cn%3A1
Task untitled
[● REC]
Page http://www.amazon.co...
Amazon.com: S...
amazon
Try Prime
Your Amazon.com
Today's Deals
Gift Cards
Sell
Help
Shop by Department
Search
Televisions
Televisions & Video
Deals
Best Sellers
Televisions
Blu-ray Players
Streaming Me
1-24 of 142 results for Electronics : Television & Video : Televisions : Smart TV : 50 to 5
Show results for
Electronics
Television & Video
Televisions
LCD TVs (59)
LED TVs (103)
OLED TVs (3)
Plasma TVs (7)
Refine by
Eligible for Free Shipping
Free Shipping by Amazon
Display Size
Clear
32 Inches & Under (41)
33 to 43 Inches (53)
44 to 49 Inches (71)
50 to 59 Inches
See Size Options
VIZIO E500i-B1 50-Inch 1080p Smart LED HDTV
$598.00 Prime
Get it by Friday, Aug 1
FREE Shipping
More Buying Choices
$595.99 new (9 offers)
$473.62 used (24 offers)
(544)

/BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV[4]/DIV[1]/DIV[1]/DIV[2]/H3[1]/A[1]/SPAN[1]

The element selected is the SPAN[1] instead of the parent A[1]

To navigate up in the path you can use the key combination:

Ctrl + ↑ (CTRL + ARROW UP)


File
Settings
Tools
http://www.amazon.com/gp/search/ref=sr_nr_p_n_size_browse-bin_mrr_3?rh=n%3A172282%2Cn%3A%21493964%2Cn%3A1
Task untitled
[REC]
Page http://www.amazon.co...
Amazon.com: S...
+
amazon
Try Prime
Your Amazon.com
Today's Deals
Gift Cards
Sell
Help
Shop by Department
Search
Televisions
Televisions & Video
Deals
Best Sellers
Televisions
Blu-ray Players
Streaming Me
1-24 of 142 results for Electronics : Television & Video : Televisions : Smart TV : 50 to 5
Show results for
Electronics
Television & Video
Televisions
LCD TVs (59)
LED TVs (103)
OLED TVs (3)
Plasma TVs (7)
Refine by
Eligible for Free Shipping
Free Shipping by Amazon
Display Size
Clear
32 Inches & Under (41)
33 to 43 Inches (53)
44 to 49 Inches (71)
50 to 59 Inches
See Size Options
VIZIO E500i-B1 50-Inch 1080p Smart LED HDTV
$598.00 Prime
Get it by Friday, Aug 1
FREE Shipping
More Buying Choices
$595.99 new (9 offers)
$473.62 used (24 offers)
(544)
Sams 120H
$1,399
Only 5
FREE
More B
$799.

```
/BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV[4]/DIV[1]/DIV[1]/DIV[2]/H3[1]/A[1]
```

The anchor element A[1] selected after pressing CTRL+ARROW UP

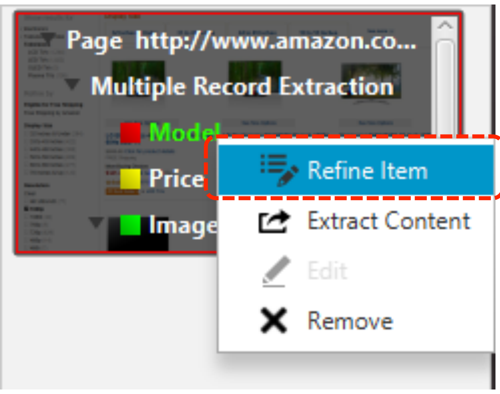
Similarly, you can use the key combination CTRL + ARROW DOWN, CTRL + ARROW LEFT and CTRL + ARROW RIGHT respectively to select a child (random) of the current element, its left sibling and its right sibling.

Refining a list

Sometimes the list obtained applying the procedure described in the previous section may not contain all and only the required items. It may include less items than the necessary or include unwanted items. In these cases, it is possible to refine the list by adding or removing an item. Typically, specifying one or two examples of items that you would like to be included or excluded from the list, the tool is able to return a more accurate list.

To refine a list:

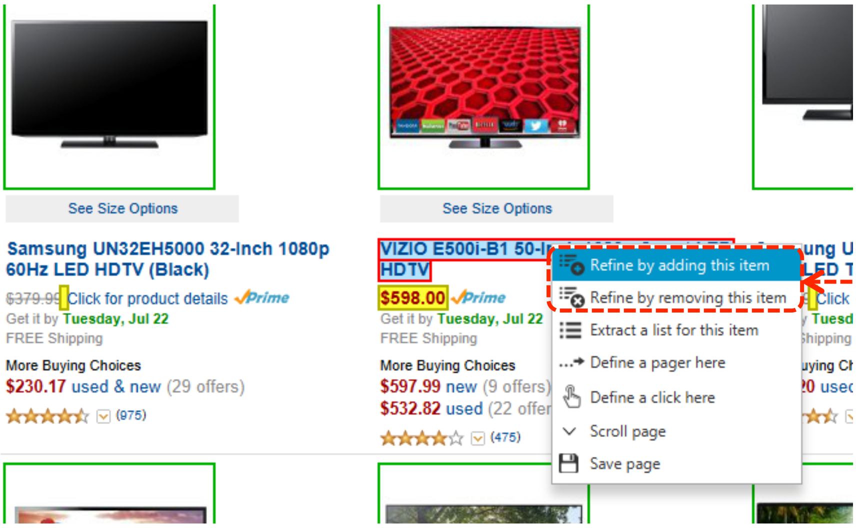
1. Right click on it and select the command *Refine Item*


Page http://www.amazon.co...
Multiple Record Extraction
Model
Price
Image
Refine Item
Extract Content
Edit
Remove

Example

Refine the list *Model*

2. Right click on an item in the web page and select the command *Refine by adding this item* or *Refine by removing this item*.


See Size Options
See Size Options
Samsung UN32EH5000 32-Inch 1080p 60Hz LED HDTV (Black)
$379.99 Click for product details Prime
Get it by Tuesday, Jul 22
FREE Shipping
More Buying Choices
$230.17 used & new (29 offers)
(975)
VIZIO E500i-B1 50-I
HDTV
$598.00 Prime
Get it by Tuesday, Jul 22
FREE Shipping
More Buying Choices
$597.99 new (9 offers)
$532.82 used (22 offer
(475)
Refine by adding this item
Refine by removing this item
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

Example

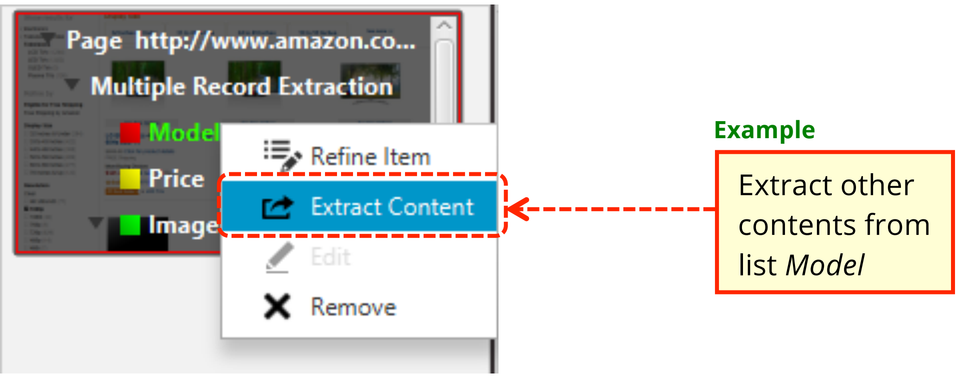
Select refine by adding to generalize the list adding this item

3. Repeat step 2 if necessary.

Extracting content

By default, the content extracted from an item is its text. It is possible, however, to select other contents. To do this:

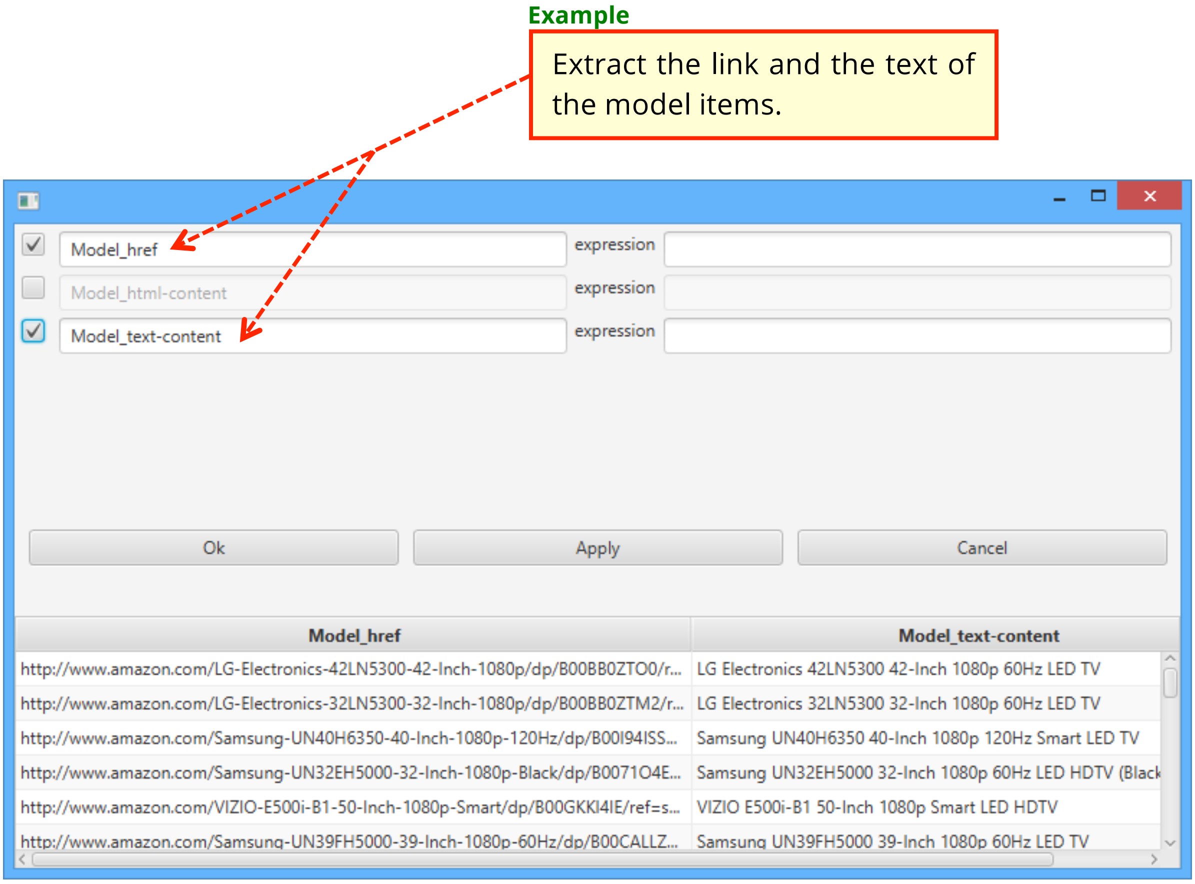
1. Right click on a list in the page action and select the command *Extract content*



Example

Extract other contents from list *Model*

2. , select, From the dialog that will appear, the contents to be extracted



Example

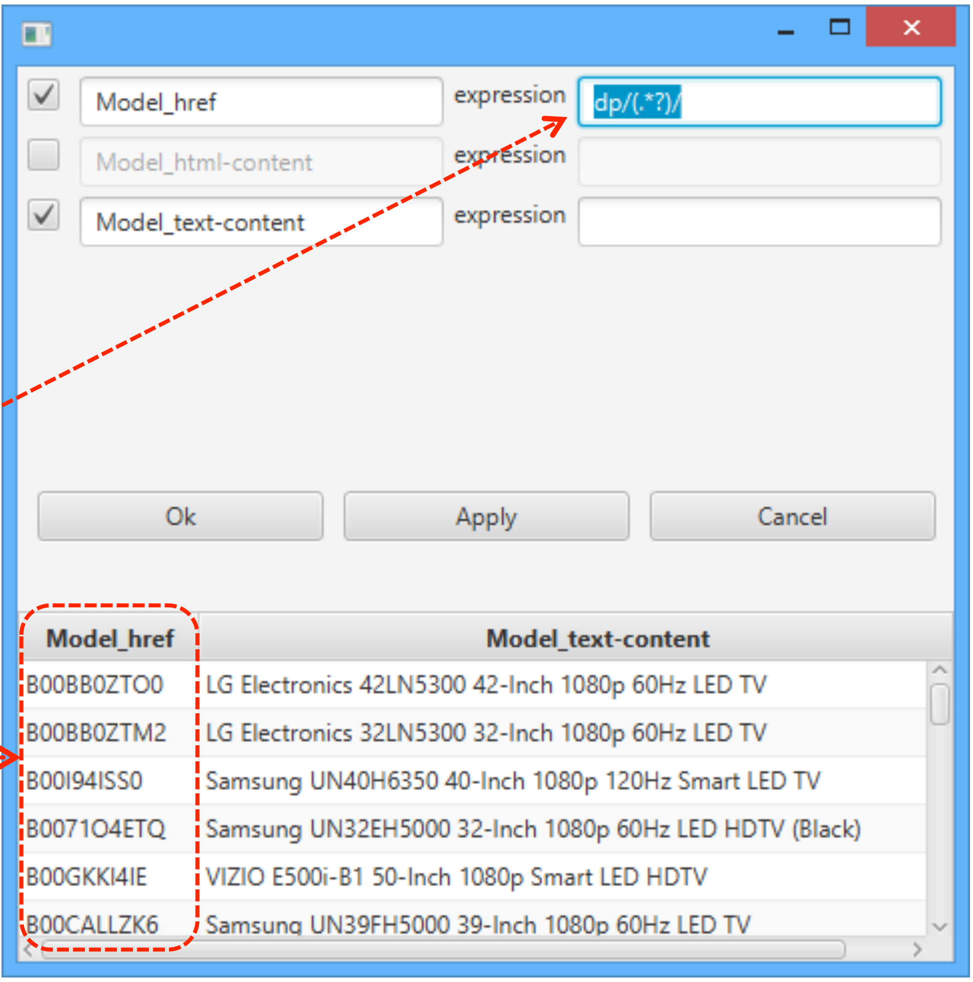
Extract the link and the text of the model items.

☑ Model_href expression
☐ Model_html-content expression
☑ Model_text-content expression

Model_href	Model_text-content
http://www.amazon.com/LG-Electronics-42LN5300-42-Inch-1080p/dp/B00BB0ZTO0/r...	LG Electronics 42LN5300 42-Inch 1080p 60Hz LED TV
http://www.amazon.com/LG-Electronics-32LN5300-32-Inch-1080p/dp/B00BB0ZTM2/r...	LG Electronics 32LN5300 32-Inch 1080p 60Hz LED TV
http://www.amazon.com/Samsung-UN40H6350-40-Inch-1080p-120Hz/dp/B00I94ISS...	Samsung UN40H6350 40-Inch 1080p 120Hz Smart LED TV
http://www.amazon.com/Samsung-UN32EH5000-32-Inch-1080p-Black/dp/B0071O4E...	Samsung UN32EH5000 32-Inch 1080p 60Hz LED HDTV (Black
http://www.amazon.com/VIZIO-E500i-B1-50-Inch-1080p-Smart/dp/B00GKKI4IE/ref=s...	VIZIO E500i-B1 50-Inch 1080p Smart LED HDTV
http://www.amazon.com/Samsung-UN39FH5000-39-Inch-1080p-60Hz/dp/B00CALLZ...	Samsung UN39FH5000 39-Inch 1080p 60Hz LED TV

3. Click on the *Ok* button

In the *expression* text field, you can write a regular expression to extract a portion of text. For example, each link in the screen above contains the id of the product, represented by the alphanumeric string after *dp/*. To extract this string instead of the link, you can write a regular expression like *dp/(.*?)/* and press ENTER or *Apply* button. The result is showed below:



☑ Model_href expression dp/(.*?)/
☐ Model_html-content expression
☑ Model_text-content expression

Model_href	Model_text-content
B00BB0ZTO0	LG Electronics 42LN5300 42-Inch 1080p 60Hz LED TV
B00BB0ZTM2	LG Electronics 32LN5300 32-Inch 1080p 60Hz LED TV
B00I94ISS0	Samsung UN40H6350 40-Inch 1080p 120Hz Smart LED TV
B0071O4ETQ	Samsung UN32EH5000 32-Inch 1080p 60Hz LED HDTV (Black)
B00GKKI4IE	VIZIO E500i-B1 50-Inch 1080p Smart LED HDTV
B00CALLZK6	Samsung UN39FH5000 39-Inch 1080p 60Hz LED TV

Example

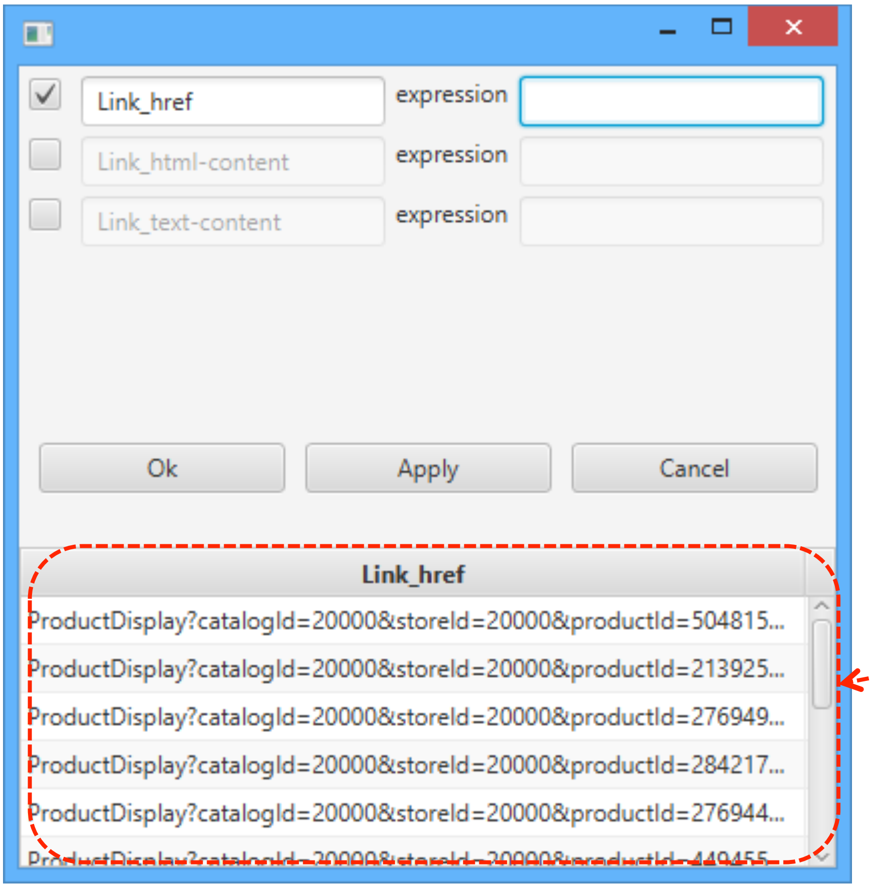
Regular expression to extract the product id from the link

A particular expression is **abs_link[<regex>][-e]**.

The **abs_link** command transforms a relative link to an absolute link. The **regex** option indicates to apply the regular expression denoted by *regex* before the link transformation. The **–e** option indicates to unescapes the result string containing entity escapes to a string containing the actual Unicode characters corresponding to the escapes.

Possible examples of use:

Transform relative links to absolute links

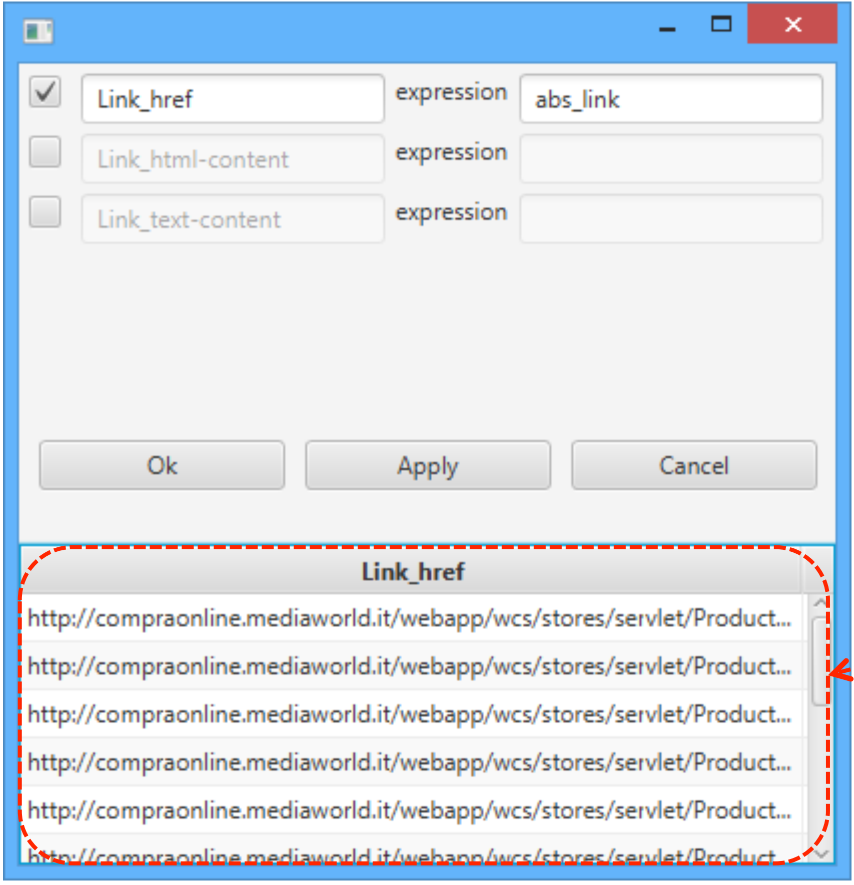


☑ Link_href expression
☐ Link_html-content expression
☐ Link_text-content expression

Link_href
ProductDisplay?catalogId=20000&storeId=20000&productId=504815...
ProductDisplay?catalogId=20000&storeId=20000&productId=213925...
ProductDisplay?catalogId=20000&storeId=20000&productId=276949...
ProductDisplay?catalogId=20000&storeId=20000&productId=284217...
ProductDisplay?catalogId=20000&storeId=20000&productId=276944...

Example

Relative links



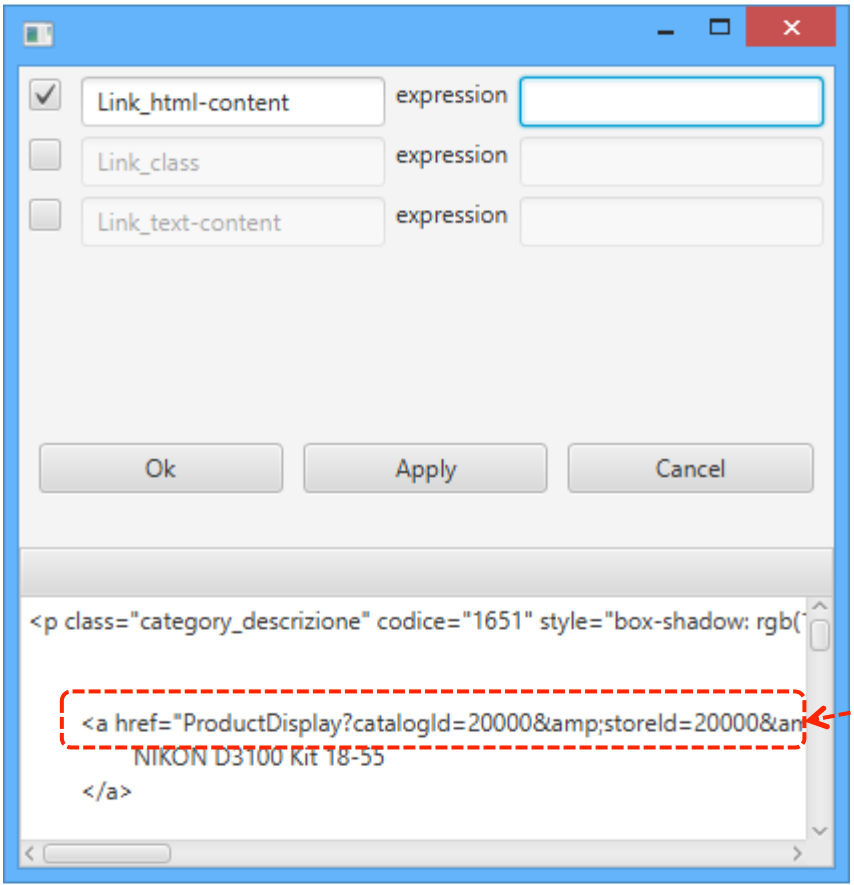
☑ Link_href expression abs_link
☐ Link_html-content expression
☐ Link_text-content expression

Link_href
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/Product...
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/Product...
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/Product...
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/Product...
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/Product...

Example

The absolute links result of the *abs_link* expression

Extract a relative link from a text and transform it in an absolute link



☑ Link_html-content expression

☐ Link_class expression

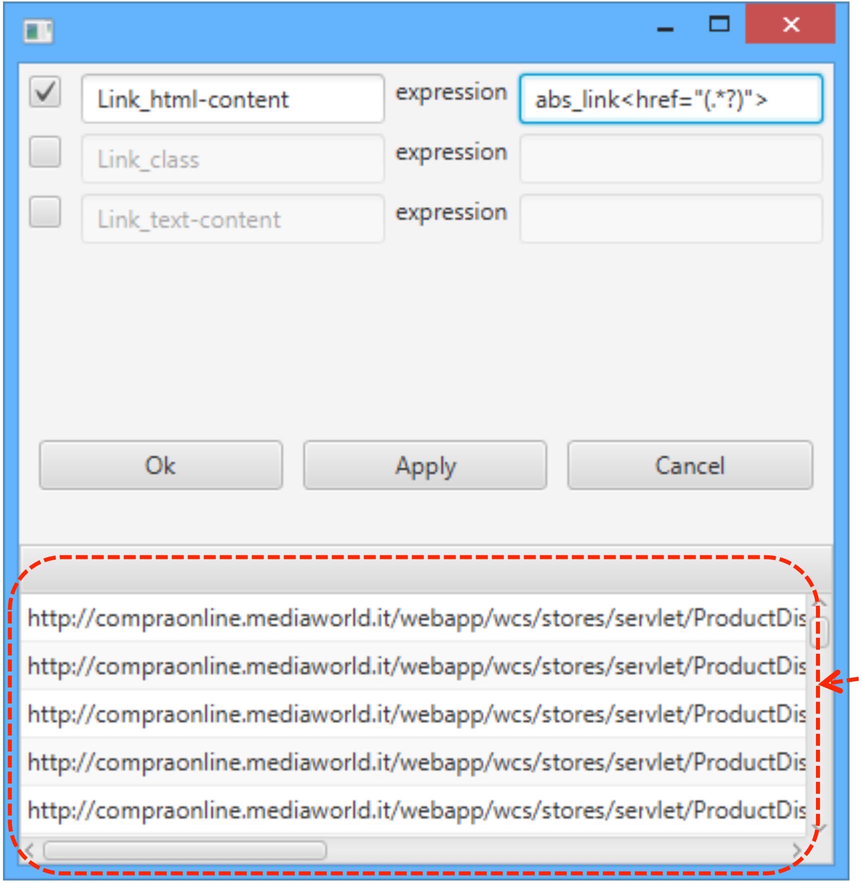
☐ Link_text-content expression

```
<p class="category_descrizione" codice="1651" style="box-shadow: rgb(

    <a href="ProductDisplay?catalogId=20000&amp;storeId=20000&an
        NIKON D3100 Kit 18-55
    </a>
```

Example

The link to extract



☑ Link_html-content expression abs_link<href="(.*?)">

☐ Link_class expression

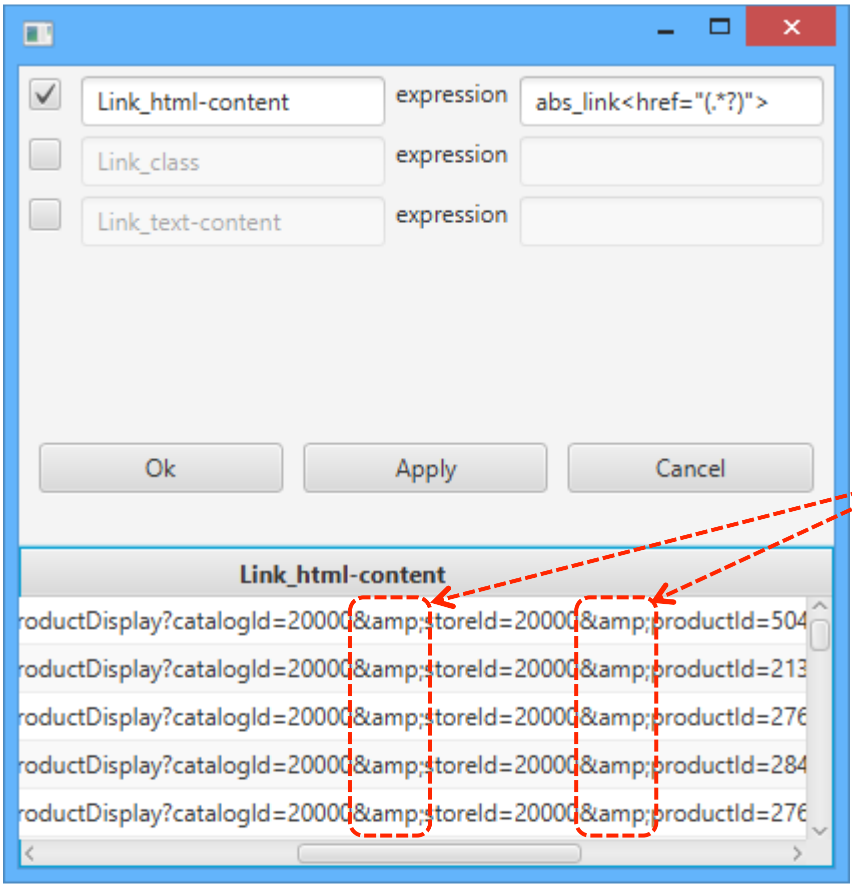
☐ Link_text-content expression

```
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/ProductDis
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/ProductDis
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/ProductDis
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/ProductDis
http://compraonline.mediaworld.it/webapp/wcs/stores/servlet/ProductDis
```

Example

The absolute links obtained after the application of the regular expression *href="(.*?)"* and the *abs_link* command

Extract a relative link from a text, transform it in an absolute link and convert entity escapes to Unicode characters



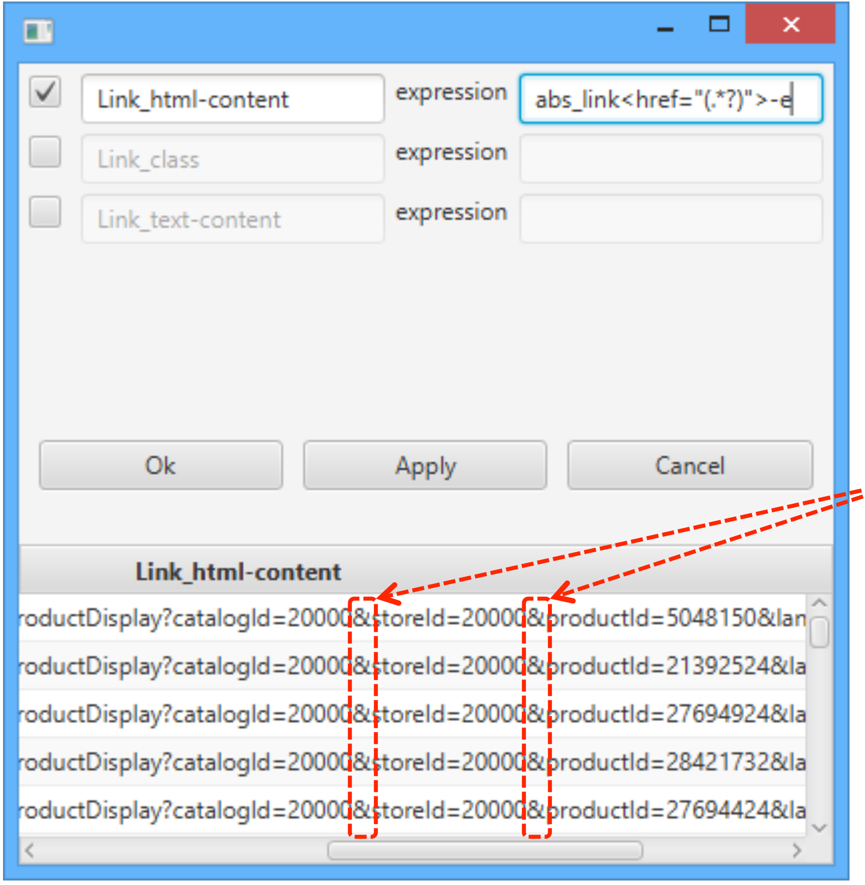
☑ Link_html-content expression abs_link<href="(.*?)">
☐ Link_class expression
☐ Link_text-content expression

Link_html-content

```
roductDisplay?catalogId=20000&amp;storeId=20000&amp;productId=504
roductDisplay?catalogId=20000&amp;storeId=20000&amp;productId=213
roductDisplay?catalogId=20000&amp;storeId=20000&amp;productId=276
roductDisplay?catalogId=20000&amp;storeId=20000&amp;productId=284
roductDisplay?catalogId=20000&amp;storeId=20000&amp;productId=276
```

Example

Entity escapes to convert (&amp; corresponds to &)



☑ Link_html-content expression abs_link<href="(.*?)">-e
☐ Link_class expression
☐ Link_text-content expression

Link_html-content

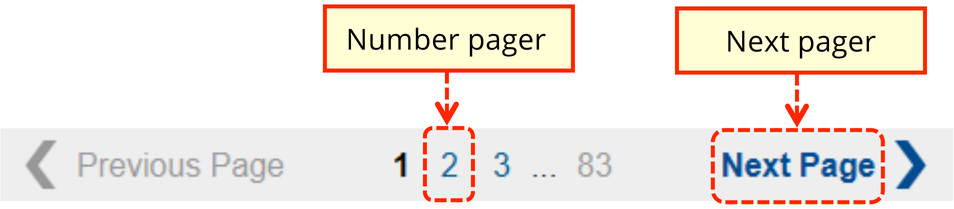
```
roductDisplay?catalogId=20000&storeId=20000&productId=5048150&lan
roductDisplay?catalogId=20000&storeId=20000&productId=21392524&la
roductDisplay?catalogId=20000&storeId=20000&productId=27649224&la
roductDisplay?catalogId=20000&storeId=20000&productId=28421732&la
roductDisplay?catalogId=20000&storeId=20000&productId=27694224&la
```

Example

Entity escapes converted after –*e* option usage

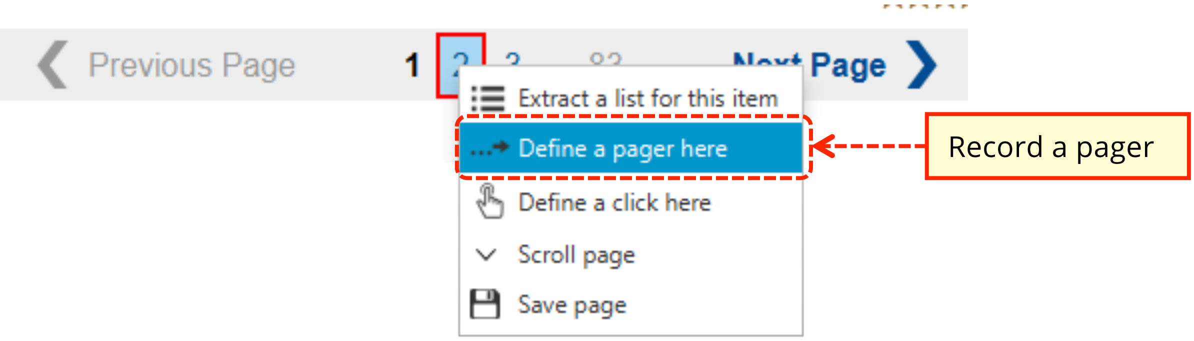
Defining a pager

Usually, the data are organized in more than one web page. In this case, it is possible to instruct the MWE to apply the extraction actions recorded on the first web page to the next pages. To do this, you must indicate where is the pager, that is a web page element on which the user must click to load a next page. There are two kind of pager: next and number. In the example below, you can see two possible pagers from an amazon.com web page.

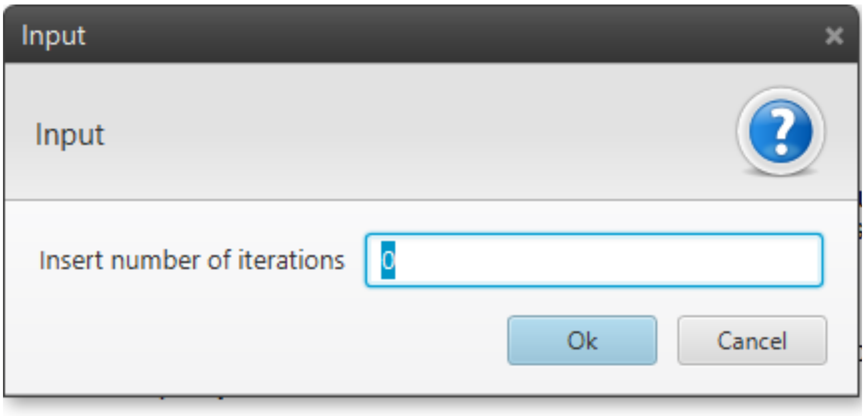

Number pager
Next pager
Previous Page
1 2 3 ... 83
Next Page

To define a pager:

1. Right click on a pager (next or number) and select *Define a pager here*

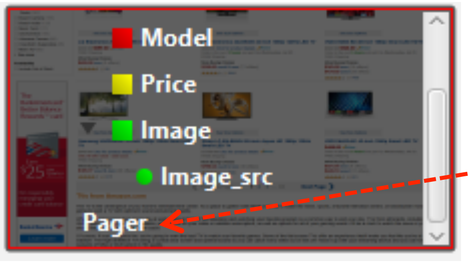

Previous Page
1 2 3 ... 83
Next Page
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page
Record a pager

2. In the input dialog insert the number of pages on which iterate. The default values is 0 that indicates to iterate until the last page.


Input
Input
Ok
Cancel

Insert number of iterations: 0

As a result, a pager action is added in the page action.


Model
Price
Image
Image_src
Pager

Pager action

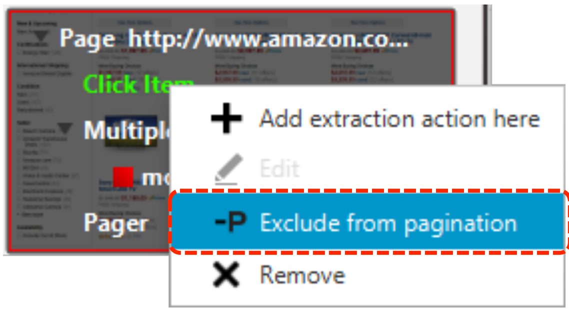
Warning: a pager should be defined as the last action so that first will be extracted the data and then will be loaded the next page and so on.

Exclude from pagination

In some cases, you may want to avoid that specific actions are performed at each pagination. Suppose for example you want to extract data from an ecommerce web site where to reach a specific category of products, you first need to click a menu option. You want to perform the click only once during the running phase on multiple pages, because the site remembers the option selection and a second click will have the effect to deselect the option. To obtain this behaviour, you need to exclude the click action from pagination.

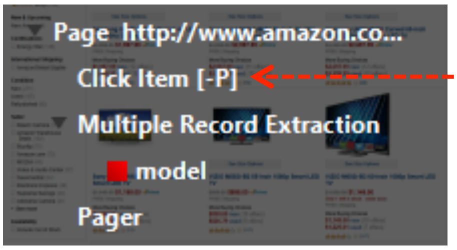
The steps to be followed are:

1. Right click on an action


Page http://www.amazon.co...
Click Item
Multipl
mo
Pager
Add extraction action here
Edit
-P Exclude from pagination
Remove

Select this command to exclude from pagination the click action, so it will be performed only once

2. A label **[-P]** will appear near the action to indicate that the option is selected. To disable it right click again on the action and deselect the option *Exclude from pagination*.


Page http://www.amazon.co...
Click Item [-P]
Multiple Record Extraction
model
Pager

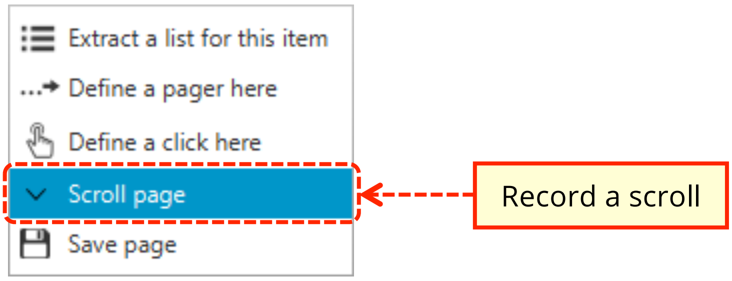
The click action excluded from pagination

Scroll page

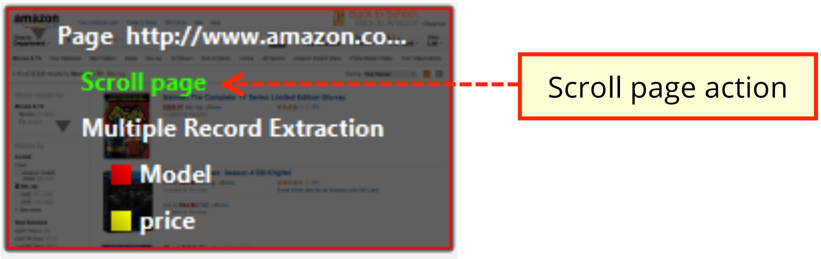
Sometimes the data to be extracted are not organized in more than one web page with a pager to navigate them, but they are in a unique web page and only a part of them is visible. To load the rest you need to scroll the web page. The MWE offers you the opportunity to handle this case recording a scroll page action. Recording it as first action in a page action, during the running phase, the tool will scroll the web page until there will be contents to be loaded and then it will execute the other extraction actions recorded.

To record a scroll action:

3. Right click on any point of the page and select *Scroll page*


Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page
Record a scroll

In the page action a scroll page action is recorded


Page http://www.amazon.co...
Scroll page
Multiple Record Extraction
Model
price
Scroll page action

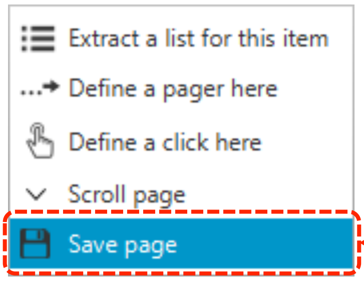
Warning: a scroll should be defined as the first action so that first will perform the scroll and then will be extracted the data.

Saving a page

It is possible to save complete web pages just by recording a save page action. During the running phase the tool will download the HTML code of the page and both image and css files, so you can load it offline.

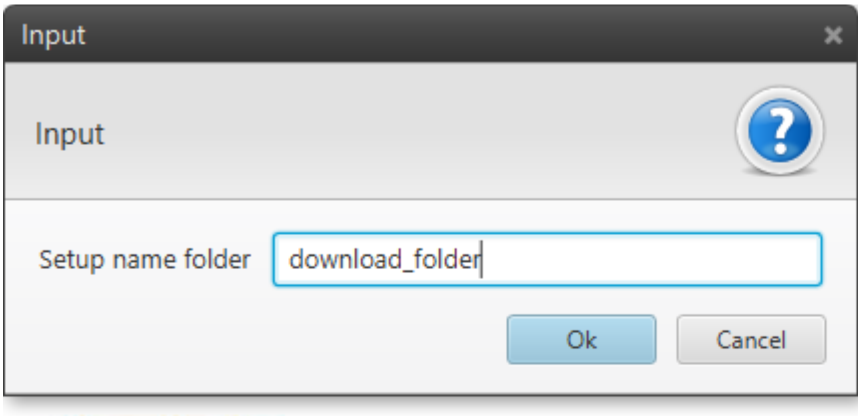
To record this action:

1. Right click on any point of the page and select *Save page*


Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

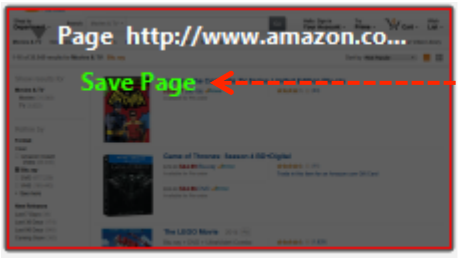
Record a save page action

2. Insert the name of the folder in which the files will be donwloaded


Input
Input
Ok
Cancel

Setup name folder: download_folder

In the page action a save page action is recorded


Page http://www.amazon.co...
Save Page

Save page action

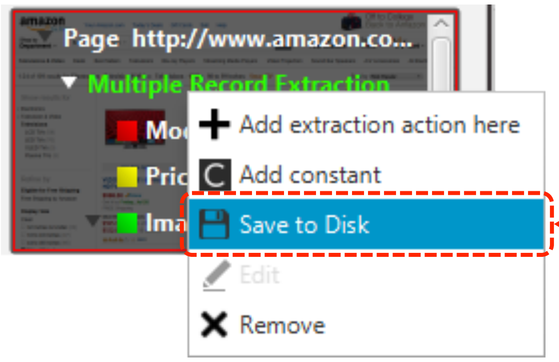
Save a Multiple Record Extraction

As explained above, a multiple record extraction recorded on a web page, can be applied to the next pages by defining a pager. This is possible because these pages maintain the same structure of the page where the actions were recorded.

In many cases even different portions of a web site share the same structure, so you can think to use a single multiple record extraction to extract data from them. Consider for example an ecommerce site where there are different categories of products. Typically, they have all the same layout, so you can record a multiple record extraction on a category and reuse it to extract the same data from another category.

To do this, the MWE allows you to save a multiple record extraction to disk and load it in another process. The step to be followed are shown below:

1. Right click on the Multiple Record Extraction in the page action and select *Save to Disk*

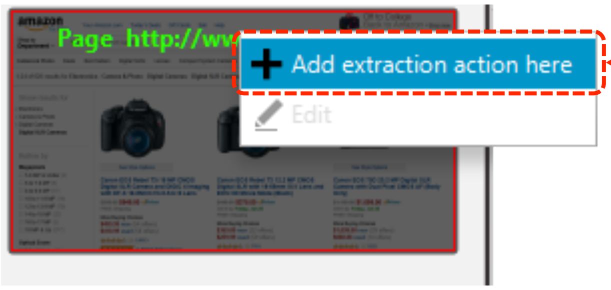

Page http://www.amazon.co...
Multiple Record Extraction
Mo
Pric
Ima
Add extraction action here
Add constant
Save to Disk
Edit
Remove

Example

Save the MRE recoded on amazon.com TV category

2. Specify where to save the multiple record extraction

3. Recording a new process on another category, right click on the action after wich you want to add the previously saved multiple record extraction and select *Add extraction action here*



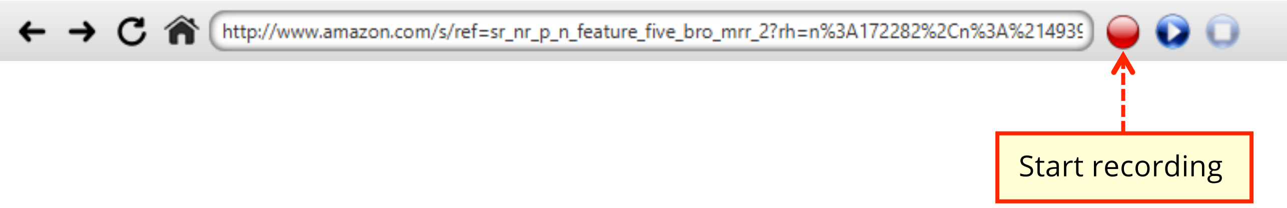
Example

Load the MRE to extract the same data also from the reflex category

Example of use: recording a complete task of extraction

In this section, you can see how to combine the previous actions to record a complete task of extraction. Suppose for example, you want to extract data from amazon about a category of products like TV. The steps to be followed are:

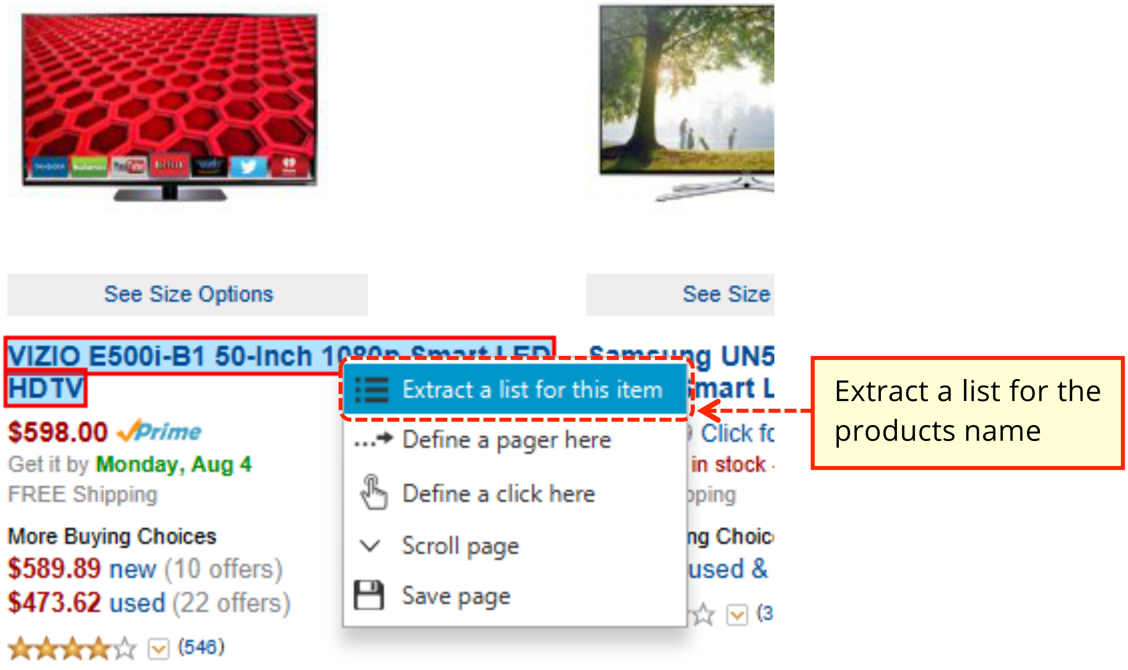
1. Load the amazon.com web site and navigate to the TV category
2. Click the record button to start the task recording



http://www.amazon.com/s/ref=sr_nr_p_n_feature_five_bro_mrr_2?rh=n%3A172282%2Cn%3A%214939

Start recording

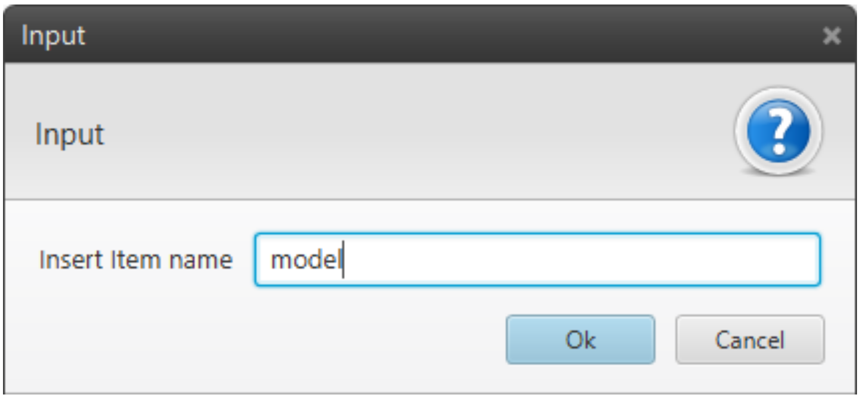
3. Extract the products name by right clicking on a product name and selecting the command *Extract a list for this item*



See Size Options
See Size
VIZIO E500i-B1 50-Inch 1080p Smart LED HDTV
Samsung UN5
Smart L
$598.00 Prime
Get it by Monday, Aug 4
FREE Shipping
More Buying Choices
$589.89 new (10 offers)
$473.62 used (22 offers)
(546)
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

Extract a list for the products name

4. Type the list name

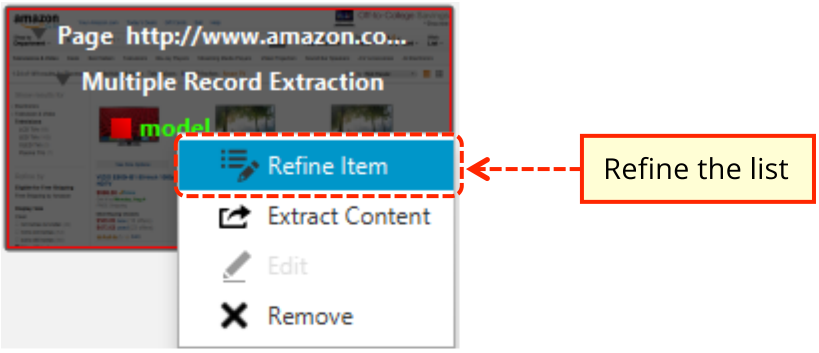


Input

Insert Item name: model

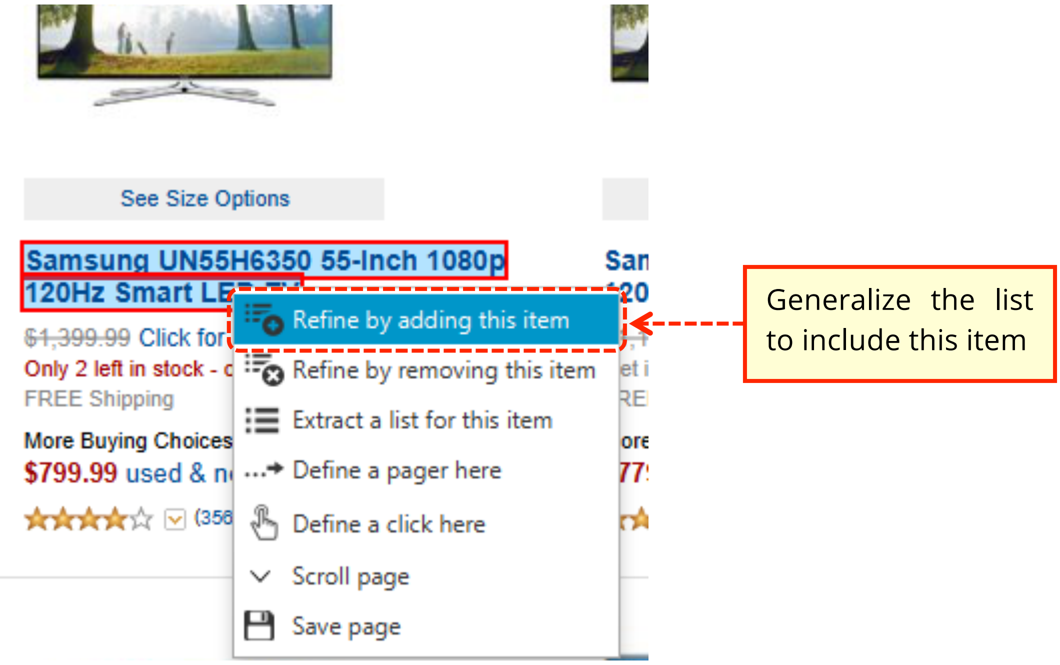
Input
Ok
Cancel

5. If necessary, refine a list by right clicking on a list in the page action and selecting the command Refine Item


Page http://www.amazon.co...
Multiple Record Extraction
model
Refine Item
Extract Content
Edit
Remove

Refine the list

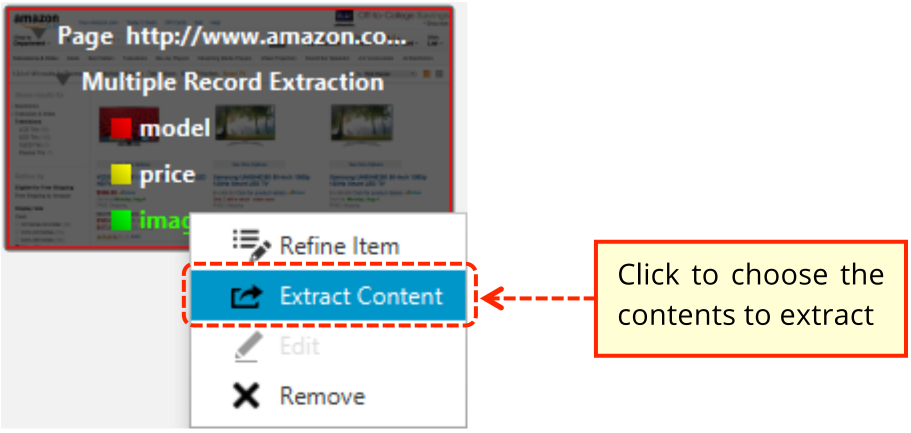
6. Right click to another product name and choose the command *Refine by adding this item* (repeat this step if necessary)


See Size Options
Samsung UN55H6350 55-Inch 1080p
120Hz Smart LE
$1,399.99 Click for
Only 2 left in stock - c
FREE Shipping
More Buying Choices
$799.99 used & n
(356
Refine by adding this item
Refine by removing this item
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page

Generalize the list to include this item

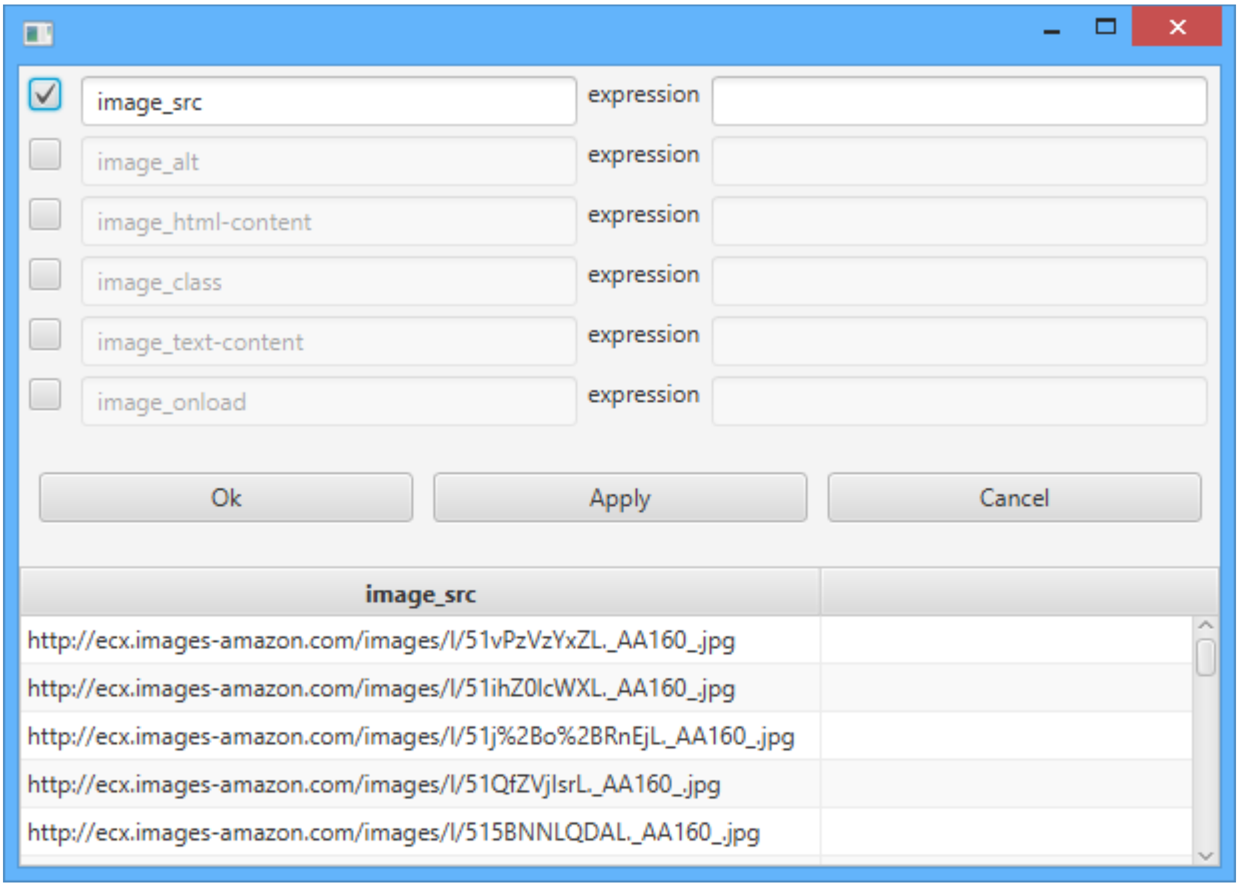
7. Repeat the step 3, 4 and if necessary 5 and 6 to extract the TV price list and the TV image list

8. To extract the URL address from each image, right click to the list and select the command *Extract Content*


Page http://www.amazon.co...
Multiple Record Extraction
model
price
imag
Refine Item
Extract Content
Edit
Remove

Click to choose the contents to extract

9. In the extract content dialog check the attribute src and press Ok

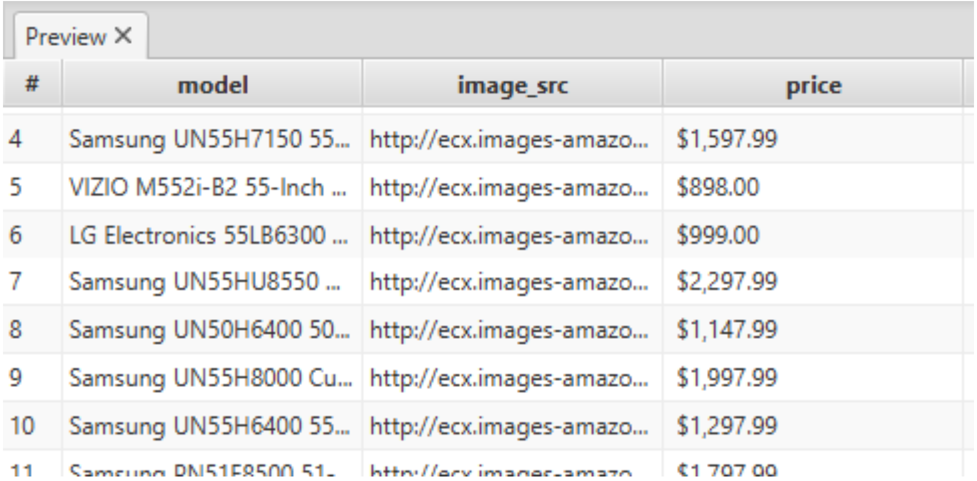


☑ image_src expression
☐ image_alt expression
☐ image_html-content expression
☐ image_class expression
☐ image_text-content expression
☐ image_onload expression

Ok
Apply
Cancel

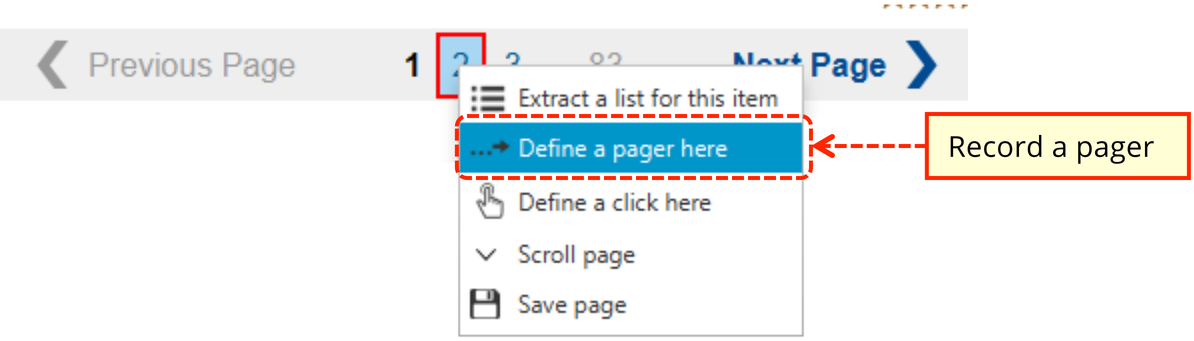
image_src
http://ecx.images-amazon.com/images/I/51vPzVzYxZL._AA160_.jpg
http://ecx.images-amazon.com/images/I/51ihZ0lcWXL._AA160_.jpg
http://ecx.images-amazon.com/images/I/51j%2Bo%2BRnEjL._AA160_.jpg
http://ecx.images-amazon.com/images/I/51QfZVjIsrL._AA160_.jpg
http://ecx.images-amazon.com/images/I/515BNNLQDAL._AA160_.jpg

10. The contents extracted are shown In the preview area



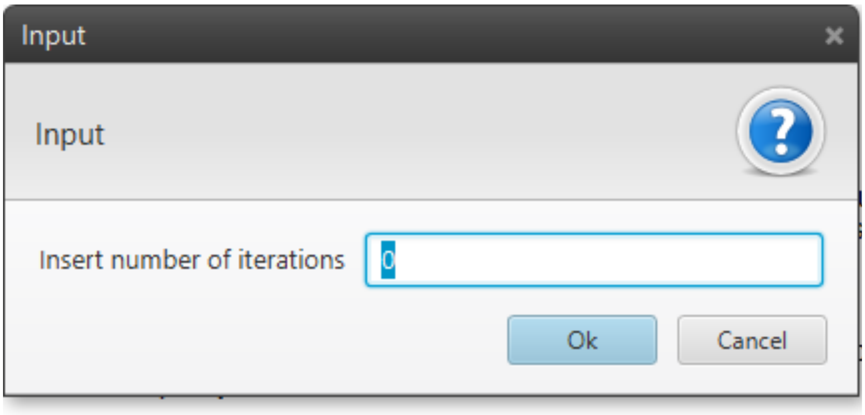
#	model	image_src	price
4	Samsung UN55H7150 55...	http://ecx.images-amazo...	$1,597.99
5	VIZIO M552i-B2 55-Inch ...	http://ecx.images-amazo...	$898.00
6	LG Electronics 55LB6300 ...	http://ecx.images-amazo...	$999.00
7	Samsung UN55HU8550 ...	http://ecx.images-amazo...	$2,297.99
8	Samsung UN50H6400 50...	http://ecx.images-amazo...	$1,147.99
9	Samsung UN55H8000 Cu...	http://ecx.images-amazo...	$1,997.99
10	Samsung UN55H6400 55...	http://ecx.images-amazo...	$1,297.99
11	Samsung PN51F8500 51...	http://ecx.images-amazo...	$1,797.99

11. Right click to a pager and select the command *Define a pager here* to apply the extraction to the next pages.



Previous Page
1
2
3
83
Next Page
Extract a list for this item
Define a pager here
Define a click here
Scroll page
Save page
Record a pager

12. Insert the number of pages on which iterate In the input dialog



Input
Input
Ok
Cancel

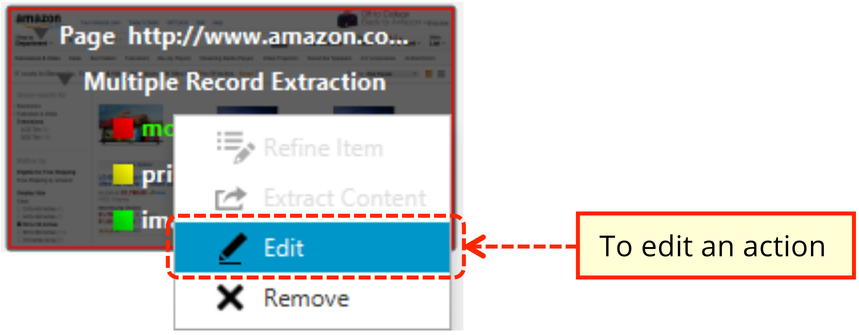
Insert number of iterations: 0

13. Press the stop button to end the task recording.

CHAPTER 5 – EDITING A TASK

The MWE offers you a set of functions to edit a previously recorded task. In this section, we will show you the steps you have to follow to edit the different actions that may be present in a task.

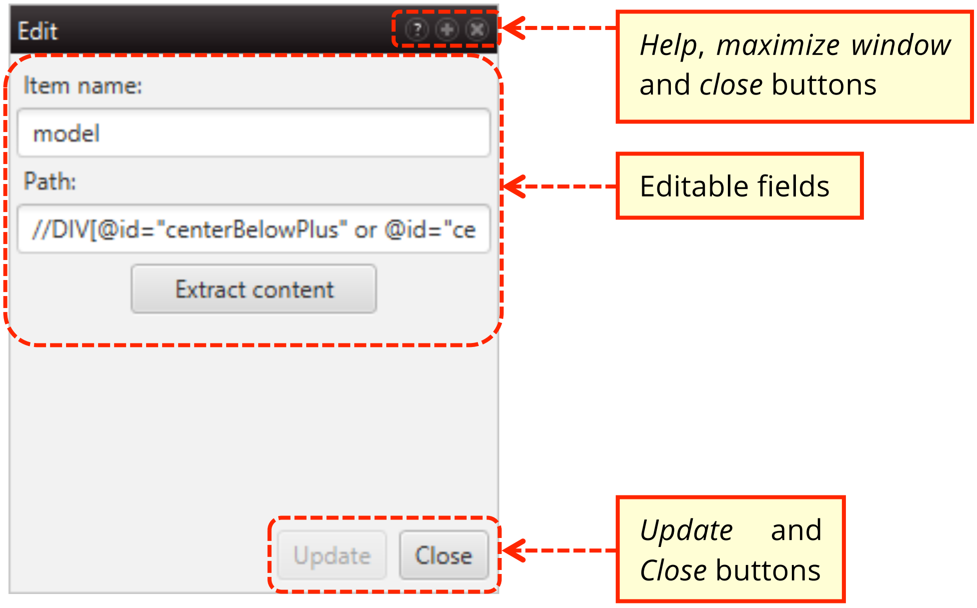
To edit an action, right click on it and select the command *Edit*


Page http://www.amazon.co...
Multiple Record Extraction
Refine Item
Extract Content
Edit
Remove

To edit an action

The *Edit* button is available only if the tool is not in recording or playing mode.

In the task area will appear an edit panel like this:


Edit
Extract content
Update
Close

Item name: model

Path: //DIV[@id="centerBelowPlus" or @id="ce

Help, maximize window and *close* buttons

Editable fields

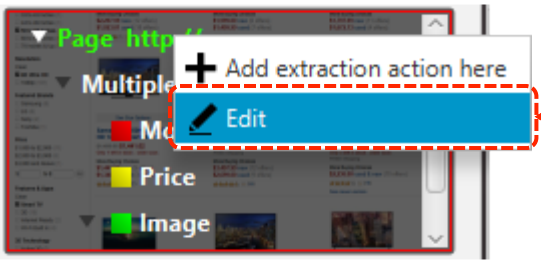
Update and *Close* buttons

Editing a page action

To edit a page action means editing the web page address where it was recorded. This is useful when you want to apply a previously recorded task on a web page to another. To work it's necessary that the new web page maintains the same layout of the original web page. If not a message will appear to ask you what to do with the actions recognized invalid.

To edit a page action:

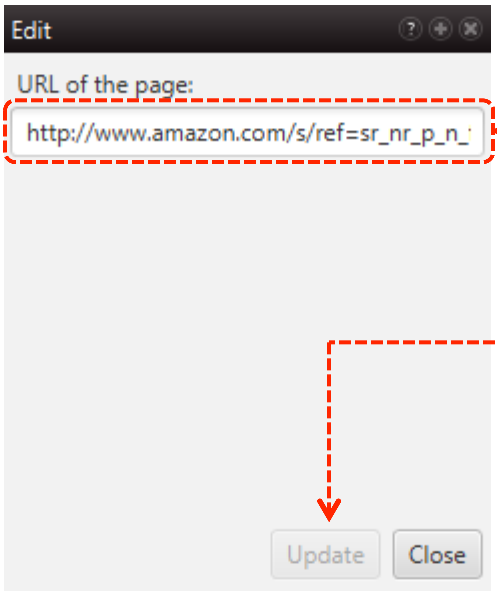
1. Right click on the page action and select the command *Edit*



Example

Edit a page action recorded on the amazon TV category

2. In the edit panel the URL of the web page where the action was recorded is shown.



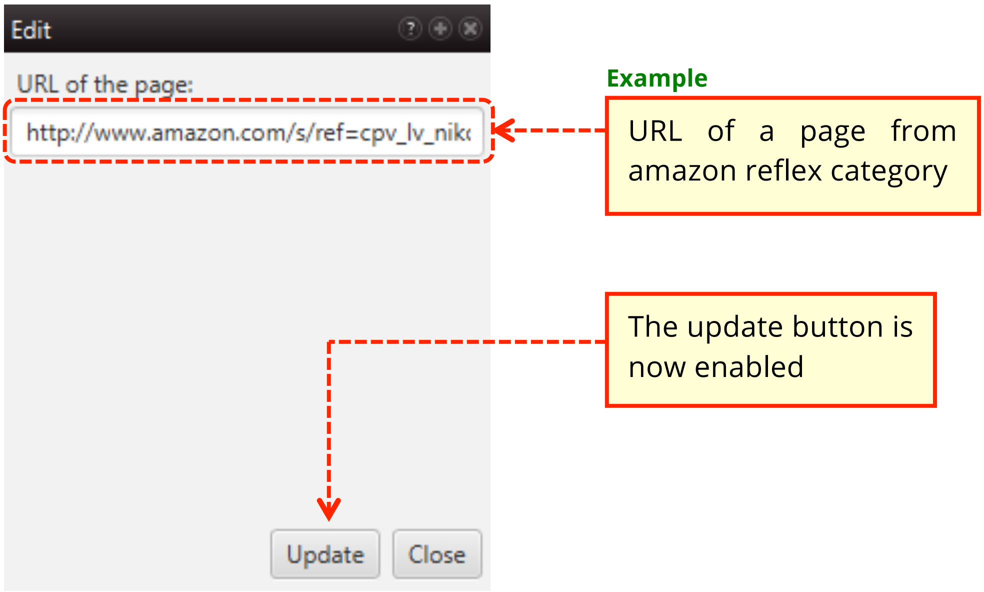
URL of the page: http://www.amazon.com/s/ref=sr_nr_p_n_

Example

URL of a page from amazon TV category

The update button will remain disabled until you load a new URL

3. To load the new URL, you can use the tool as a common browser and navigate the site to reach the desired URL or you can paste an address in the editable filed and press the ENTER button to load it.


Edit
URL of the page:
Update
Close

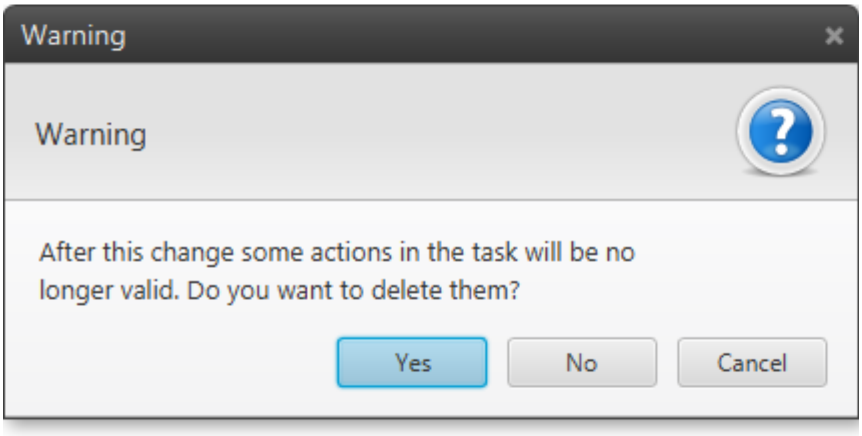
http://www.amazon.com/s/ref=cpv_lv_nik

Example

URL of a page from amazon reflex category

The update button is now enabled

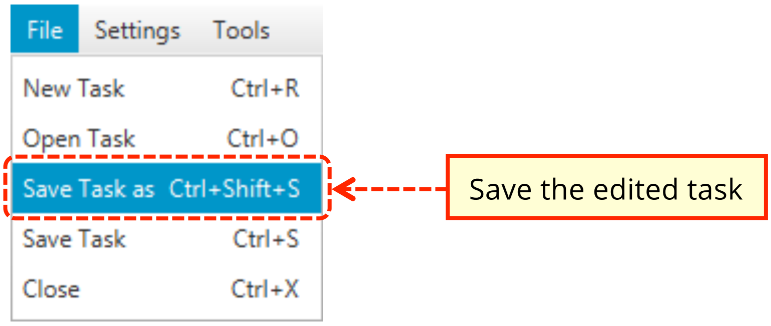
4. Click the *Update* button and then *Close*.
    4.1. If some actions are no longer valid on this new page a warning message will appear


Warning
Warning
Yes
No
Cancel

After this change some actions in the task will be no longer valid. Do you want to delete them?

    4.2. Click *Yes* if you want to delete the invalid actions, *No* to keep them even if invalid, Cancel to abort the update.

5. To make the changes persistent, save the task from the File menu.

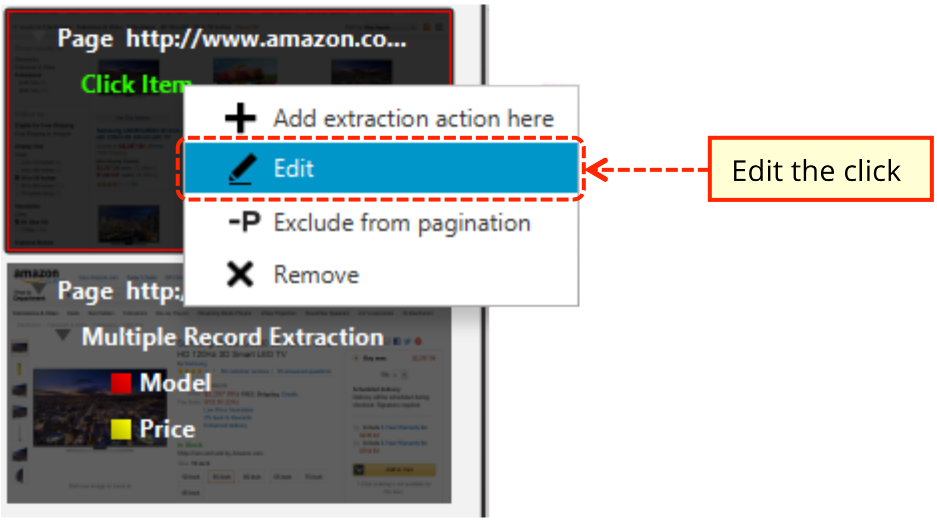

File
Settings
Tools
New Task Ctrl+R
Open Task Ctrl+O
Save Task as Ctrl+Shift+S
Save Task Ctrl+S
Close Ctrl+X

Save the edited task

Editing a click

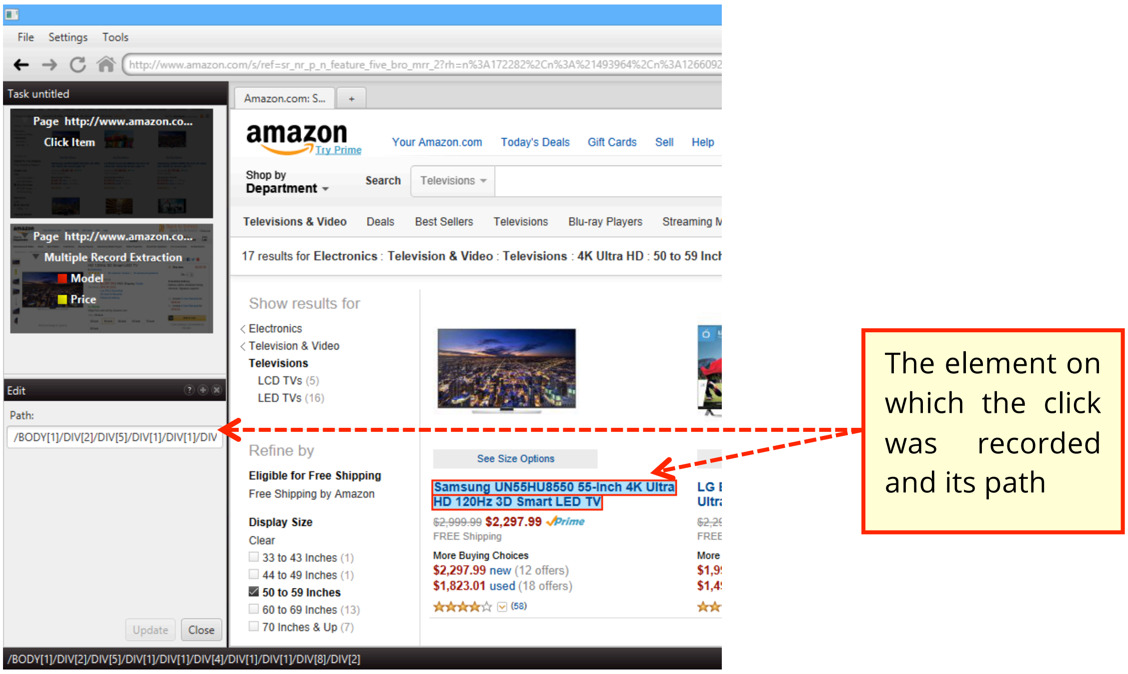
The procedure to edit a click is very simple, but needs attention because a click typically loads a new web page where presumably other actions were recorded. The steps to be followed are:

1. Right click on the click action and select the command *Edit*


Page http://www.amazon.co...
Click Item
Add extraction action here
Edit
Exclude from pagination
Remove
Page http:
Multiple Record Extraction
Model
Price

Edit the click

2. In the edit panel, the element path where the click was recorded is shown while in the web view the corresponding element is highlighted to allow an easier identification.

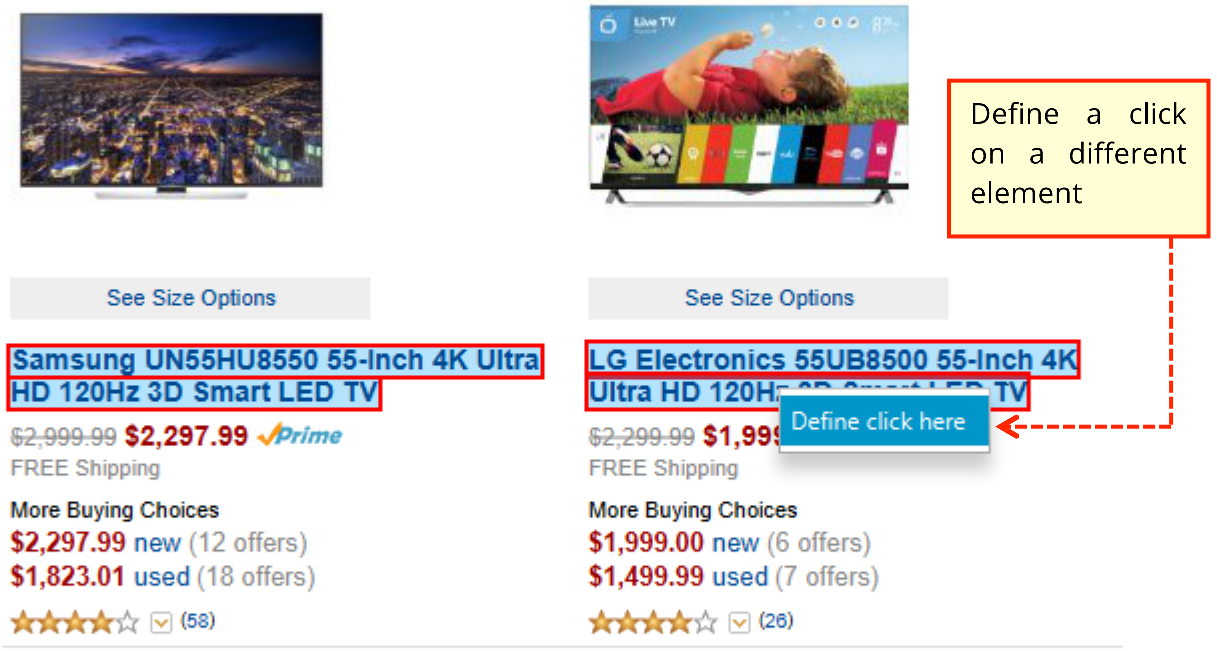

File
Settings
Tools
http://www.amazon.com/s/ref=sr_nr_p_n_feature_five_bro_mrr_2?rh=n%3A172282%2Cn%3A%21493964%2Cn%3A1266092
Task untitled
Page http://www.amazon.co...
Click Item
Page http://www.amazon.co...
Multiple Record Extraction
Model
Price
Edit
Update
Close
Amazon.com: S...
amazon
Try Prime
Your Amazon.com
Today's Deals
Gift Cards
Sell
Help
Shop by Department
Search
Televisions
Televisions & Video
Deals
Best Sellers
Televisions
Blu-ray Players
Streaming M
17 results for Electronics : Television & Video : Televisions : 4K Ultra HD : 50 to 59 Inch
Show results for
Electronics
Television & Video
Televisions
LCD TVs (5)
LED TVs (16)
Refine by
Eligible for Free Shipping
Free Shipping by Amazon
Display Size
Clear
33 to 43 Inches (1)
44 to 49 Inches (1)
50 to 59 Inches
60 to 69 Inches (13)
70 Inches & Up (7)
See Size Options
Samsung UN55HU8550 55-Inch 4K Ultra HD 120Hz 3D Smart LED TV
$2,999.99 $2,297.99 Prime
FREE Shipping
More Buying Choices
$2,297.99 new (12 offers)
$1,823.01 used (18 offers)
(58)

Path: /BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV

/BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV[4]/DIV[1]/DIV[1]/DIV[8]/DIV[2]

The element on which the click was recorded and its path

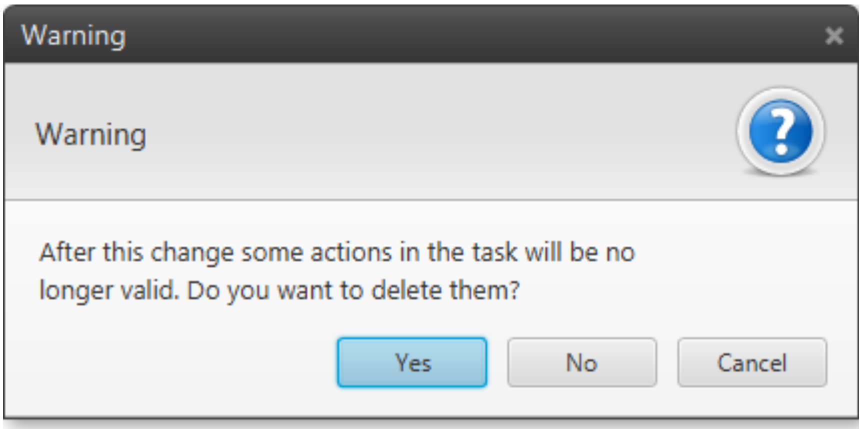
3. Right click to a different web page element and select Define click here


See Size Options
Samsung UN55HU8550 55-Inch 4K Ultra HD 120Hz 3D Smart LED TV
$2,999.99 $2,297.99 Prime
FREE Shipping
More Buying Choices
$2,297.99 new (12 offers)
$1,823.01 used (18 offers)
(58)
See Size Options
LG Electronics 55UB8500 55-Inch 4K Ultra HD 120H
$2,299.99 $1,99
Define click here
FREE Shipping
More Buying Choices
$1,999.00 new (6 offers)
$1,499.99 used (7 offers)
(26)

Define a click on a different element

4. Click *Update*.

5. A warning message will appear to ask you what to do with the next actions that may no longer be valid.


Warning
×
Warning
Yes
No
Cancel

After this change some actions in the task will be no longer valid. Do you want to delete them?

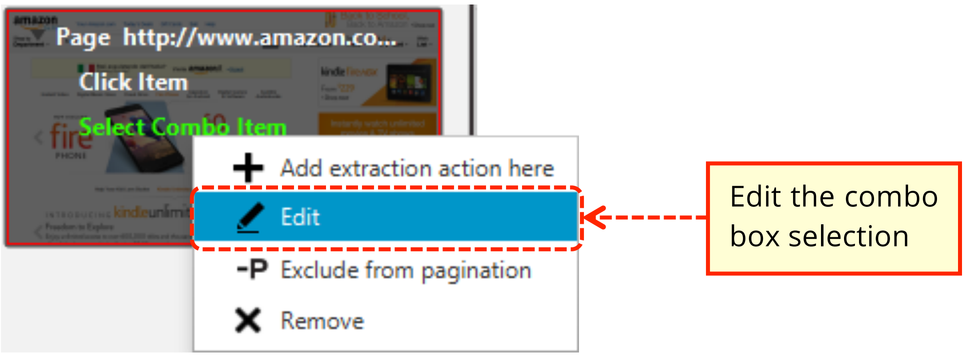
6. Click *Yes* if you want to delete the next actions, *No* to keep them, Cancel to abort the update.

7. To make the changes persistent, save the task from the File menu.

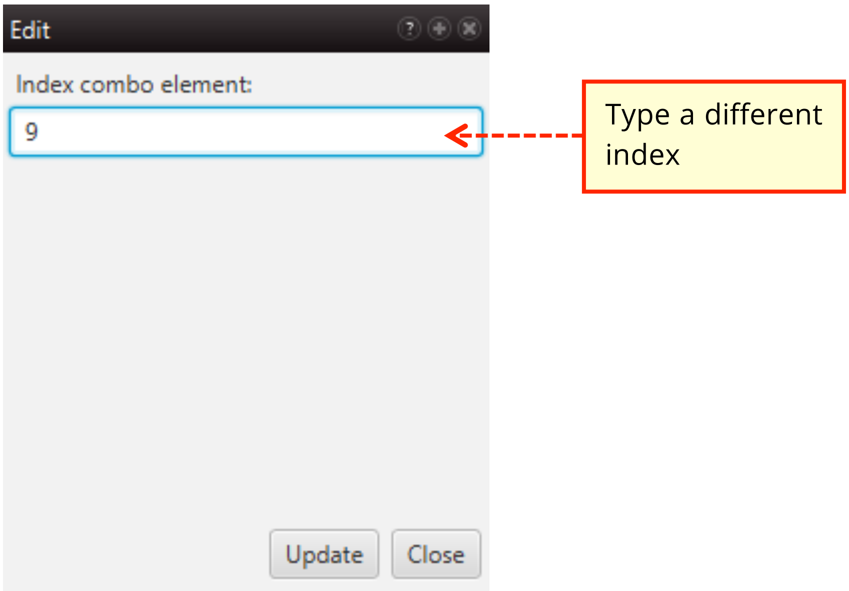
Editing a combo box selection

In order to edit the element selected from a combo box:

1. Right click on the click action and select the command *Edit*


Page http://www.amazon.co...
Click Item
Select Combo Item
Add extraction action here
Edit
-P Exclude from pagination
Remove
Edit the combo box selection

2. In the edit panel change the index of combo element


Edit
Update
Close
Type a different index

Index combo element: 9

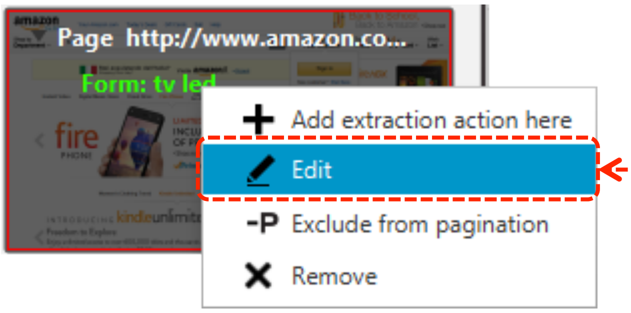
8. Click *Update* and *Close*.

9. To make the changes persistent, save the task from the File menu.

Editing a form filling action

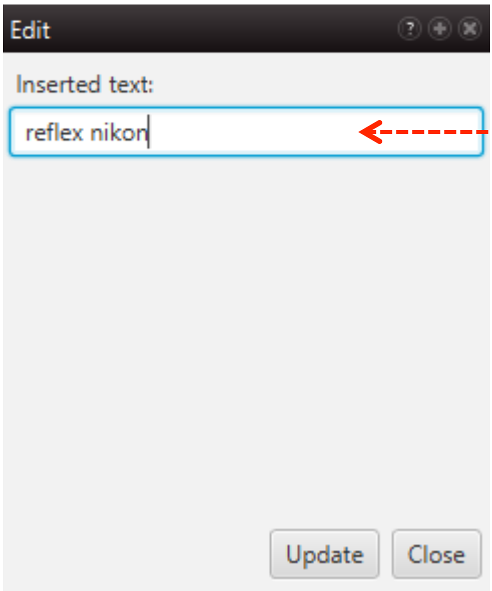
If you want to edit the text inserted at form filling recording time,:

1. Right click on the click action and select the command *Edit*


Page http://www.amazon.co...
Form: tv led
Add extraction action here
Edit
Exclude from pagination
Remove

Edit the combo box selection

2. change the inserted text In the edit panel


Edit
Update
Close

Inserted text: reflex nikon

Type a different text

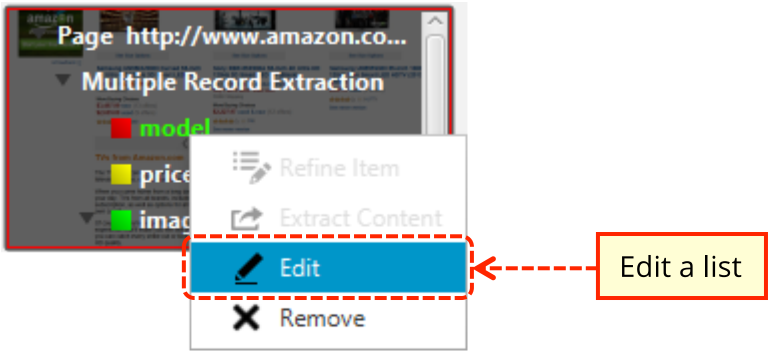
3. Click *Update* and *Close*.

4. , save the task from the File menu To make the changes persistent.

Editing a list

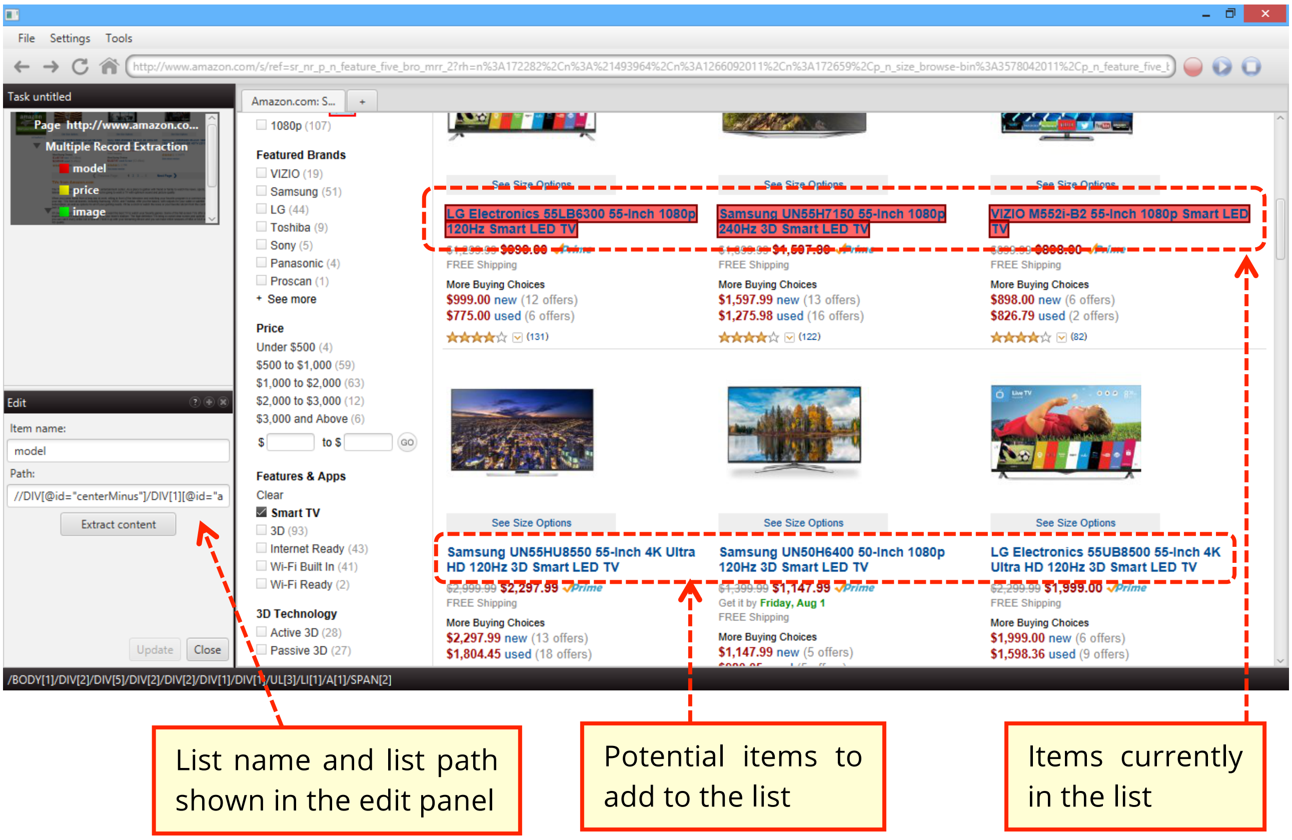
You can refine a list by adding or removing an element at any time. To do this:

1. Right click on the list and select the command *Edit*


Page http://www.amazon.co...
Multiple Record Extraction
model
price
imag
Refine Item
Extract Content
Edit
Remove

Edit a list

2. the elements currently in the list are highlighted in the web page and the path and the name of the list are shown in the edit panel


File Settings Tools
Task untitled
Page http://www.amazon.co...
Multiple Record Extraction
model
price
image
Amazon.com: S...
Edit
Extract content
Update
Close

Item name: model

Path: //DIV[@id="centerMinus"]/DIV[1][@id="a

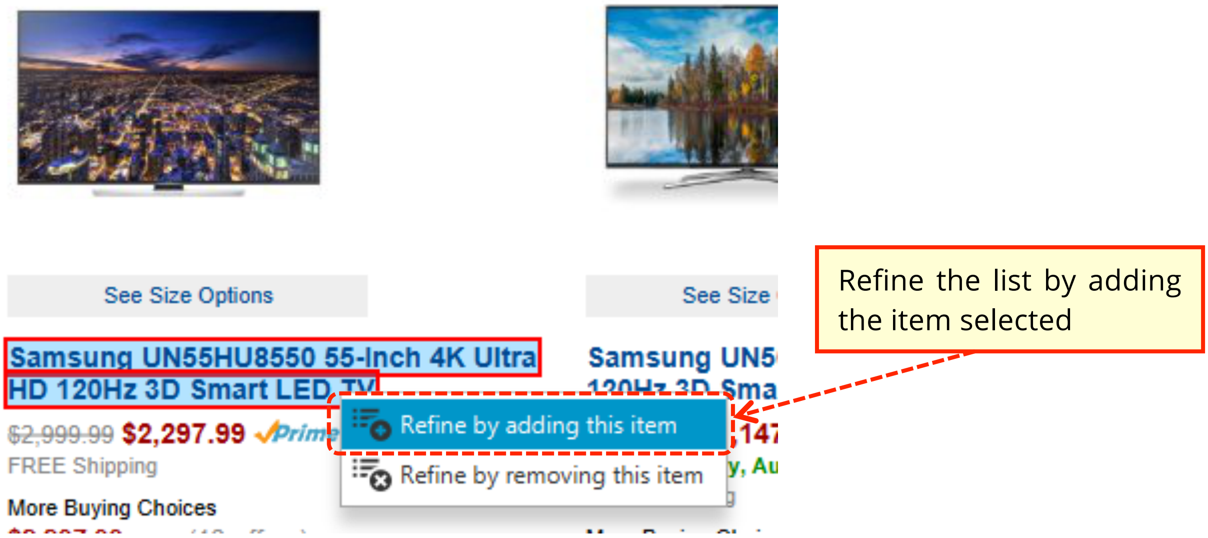
/BODY[1]/DIV[2]/DIV[5]/DIV[2]/DIV[2]/DIV[1]/DIV[1]/UL[3]/LI[1]/A[1]/SPAN[2]

List name and list path shown in the edit panel

Potential items to add to the list

Items currently in the list

   2.1. To change the name type a different name
   2.2. To refine the list by adding (or removing) an item right click on it and select the command *Refine by adding (or removing) this item*.


See Size Options
See Size
Samsung UN55HU8550 55-Inch 4K Ultra HD 120Hz 3D Smart LED TV
$2,999.99 $2,297.99 Prime
FREE Shipping
More Buying Choices
Refine by adding this item
Refine by removing this item
Refine the list by adding the item selected

2.3. To extract a different content select the command *Extract content* from the edit pane

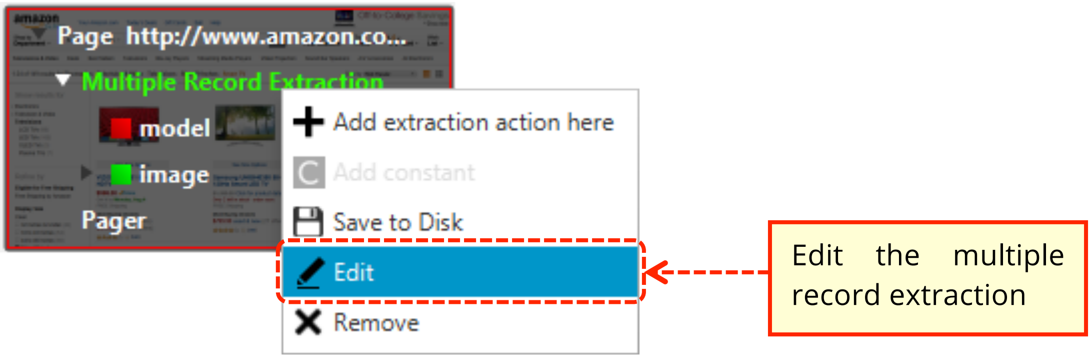
3. Click *Update* and *Close*.

4. To make the changes persistent, save the task from the File menu.

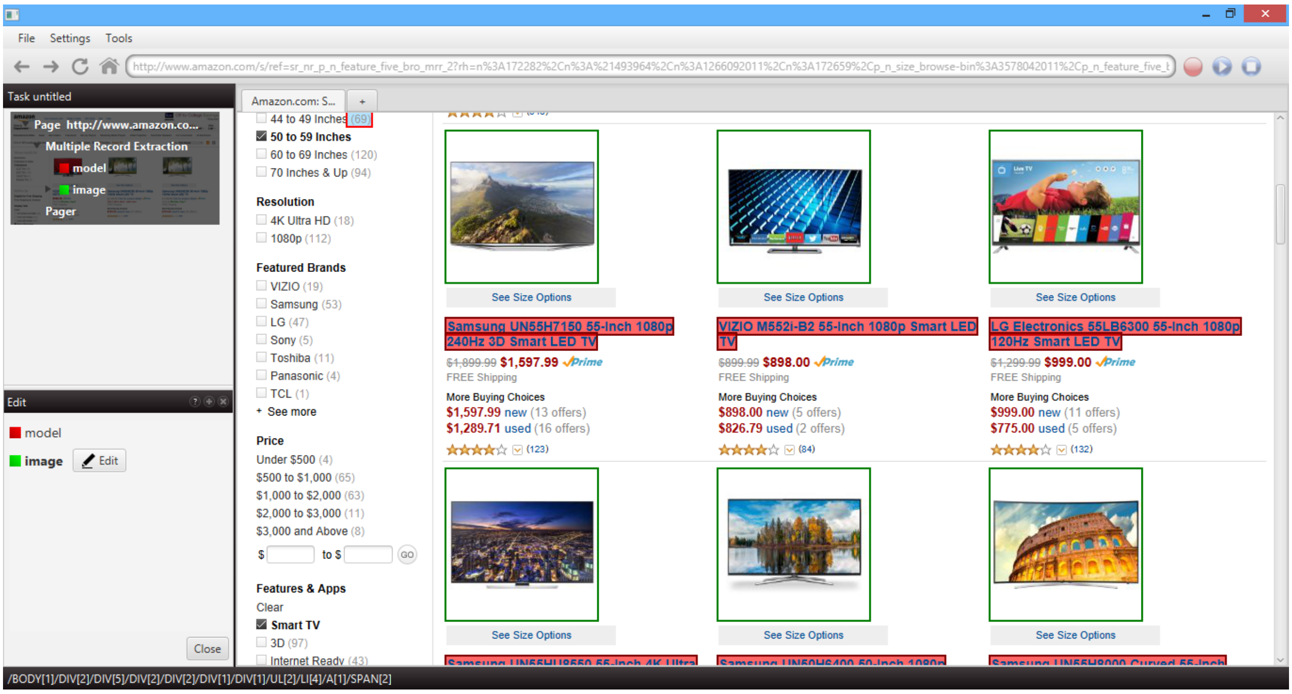
Adding a new list

You can use the edit function to add a new list to an existing multiple record extraction. To do this:

1. Right click on the multiple record extraction and select the command *Edit*

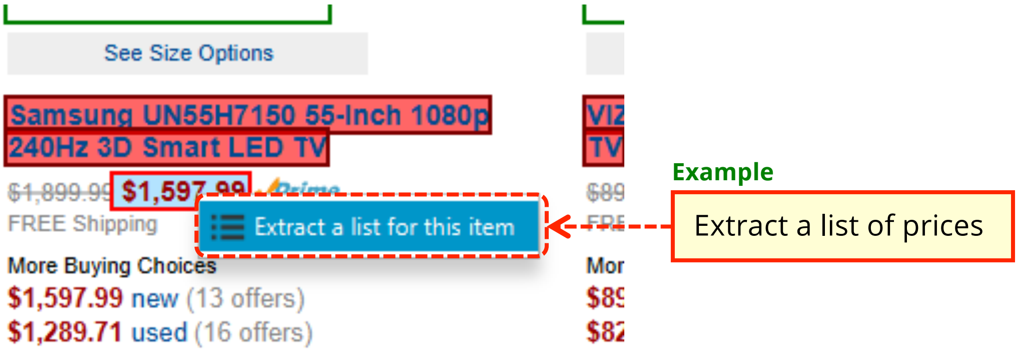

Page http://www.amazon.co...
Multiple Record Extraction
model
image
Pager
Add extraction action here
Add constant
Save to Disk
Edit
Remove
Edit the multiple record extraction

The elements of the multiple record extraction are shown in the edit panel, while the currently selected item is highlighted in the web page.

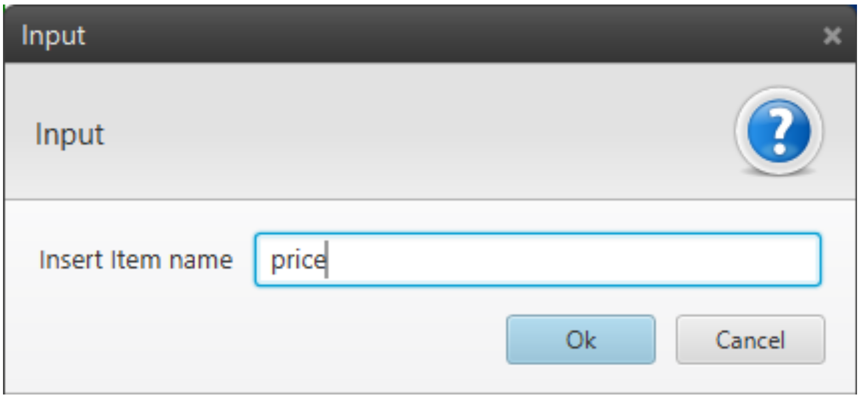

File
Settings
Tools
Task untitled
Page http://www.amazon.co...
Multiple Record Extraction
model
image
Pager
Edit
model
image
Edit
Close
Amazon.com: S...
Samsung UN55H7150 55-Inch 1080p 240Hz 3D Smart LED TV
VIZIO M552i-B2 55-Inch 1080p Smart LED TV
LG Electronics 55LB6300 55-Inch 1080p 120Hz Smart LED TV
See Size Options

```
/BODY[1]/DIV[2]/DIV[5]/DIV[2]/DIV[2]/DIV[1]/DIV[1]/UL[2]/LI[4]/A[1]/SPAN[2]
```

2. To extract a new list, right click on an item in the web page and select the command *Extract a list for this item*

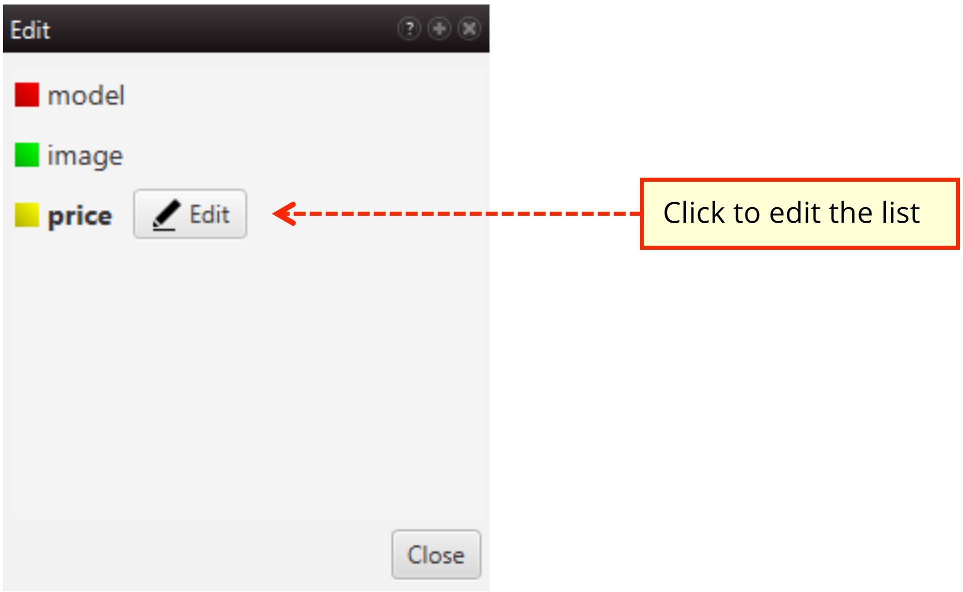

See Size Options
Samsung UN55H7150 55-Inch 1080p 240Hz 3D Smart LED TV
$1,899.99 $1,597.99
FREE Shipping
More Buying Choices
$1,597.99 new (13 offers)
$1,289.71 used (16 offers)
Extract a list for this item
Example
Extract a list of prices

3. Type the name of the new list


Input
Input
Ok
Cancel

Insert Item name: price

4. The new list is added in the multiple record extraction and in the edit panel. If you want to refine it or specify contents to be extracted, point the mouse over the list and choose the edit button


Edit
model
image
price
Edit
Close
Click to edit the list

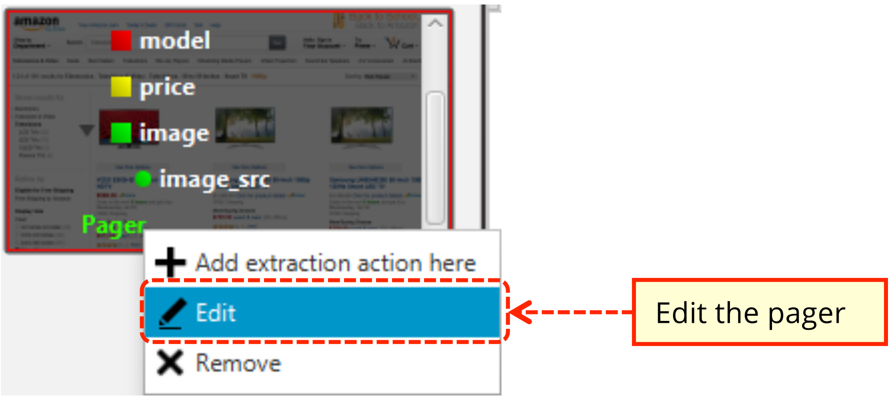
5. Click *Close*.

6. To make the changes persistent, save the task from the File menu.

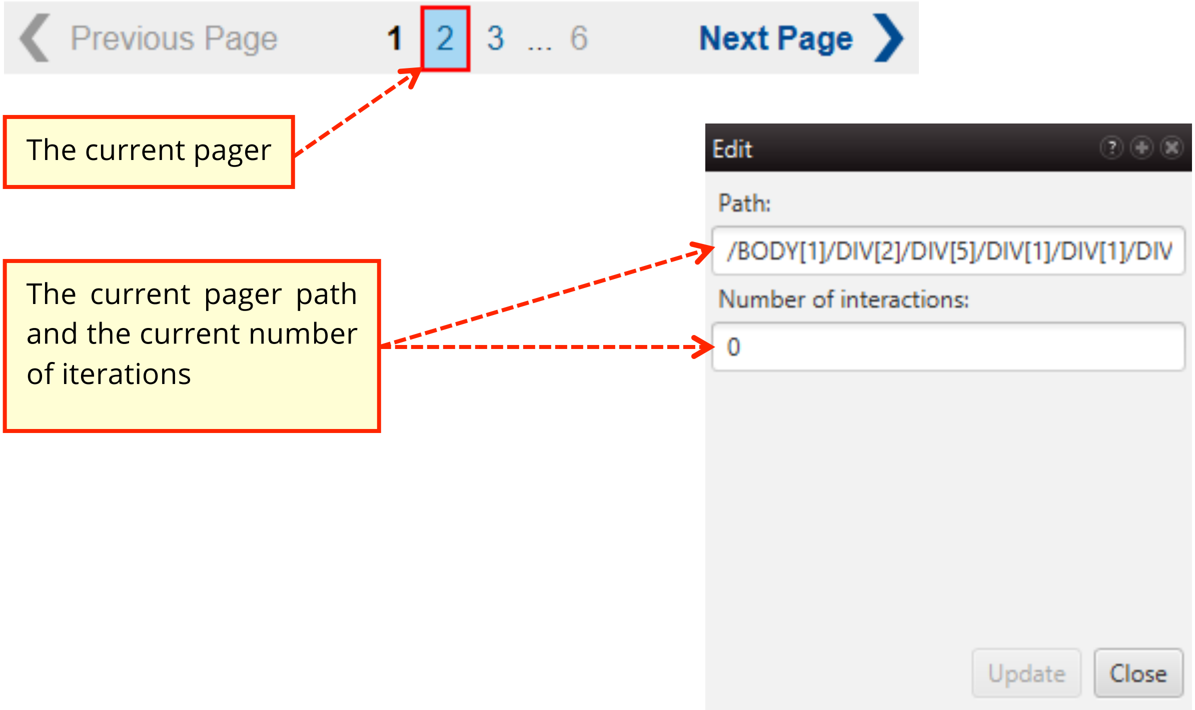
Edit a pager

You can use this function to change both the position and the number of iterations of a pager.

1. Right click on the pager action and select the command *Edit*


model
price
image
image_src
Pager
Add extraction action here
Edit
Remove
Edit the pager

2. The current pager is highlighted in the web page, while in the edit panel the current pager path and the current number of iterations are shown:


Previous Page
1 2 3 … 6
Next Page
Edit
Update
Close

The current pager

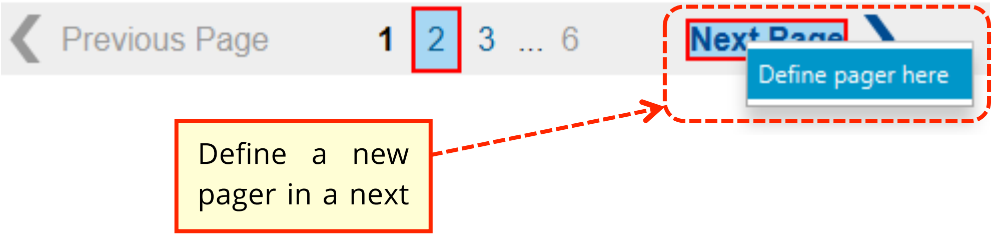
The current pager path and the current number of iterations

Path: /BODY[1]/DIV[2]/DIV[5]/DIV[1]/DIV[1]/DIV

Number of interactions: 0

2.1. Type a different number in the field To change the number of iterations,.

2.2. Right click on a different pager and select *Define a pager here* to change the pager position.


Previous Page
1 2 3 … 6
Next Page
Define pager here

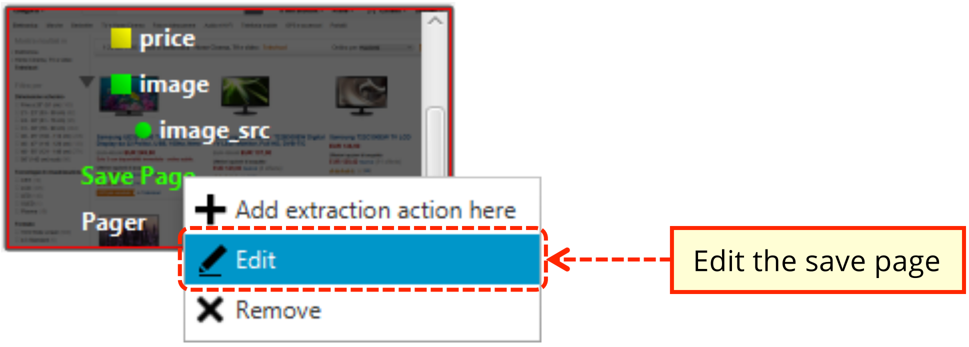
Define a new pager in a next

3. Click *Update* and *Close*.
4. Save the task from the File menu to make the changes persistent.

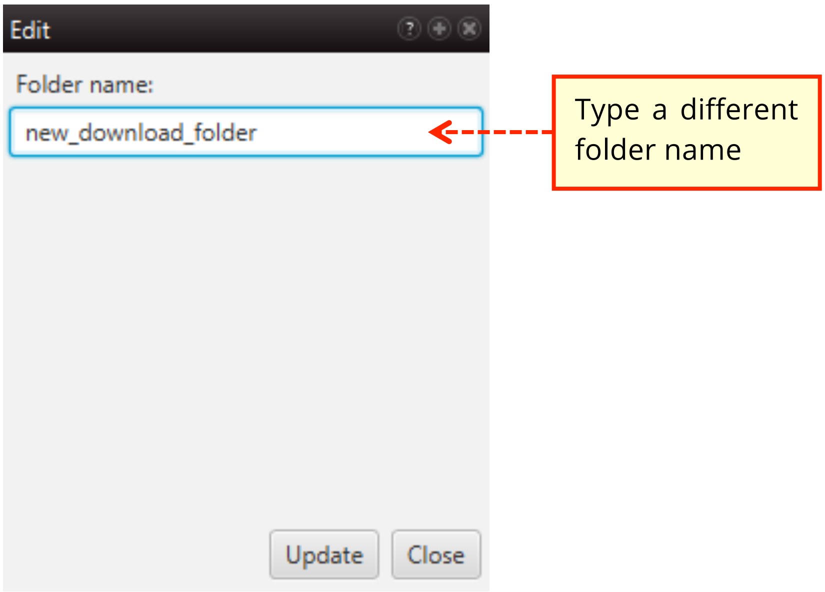
Editing a save page

During a save page action recording, you set the folder in which the files will be downloaded. To change this folder:

1. Right click on the save page action and select the command *Edit.*


price
image
image_src
Save Page
Pager
Add extraction action here
Edit
Remove
Edit the save page

2. Change the folder name In the edit panel.


Edit
Update
Close
Type a different folder name

Folder name: new_download_folder

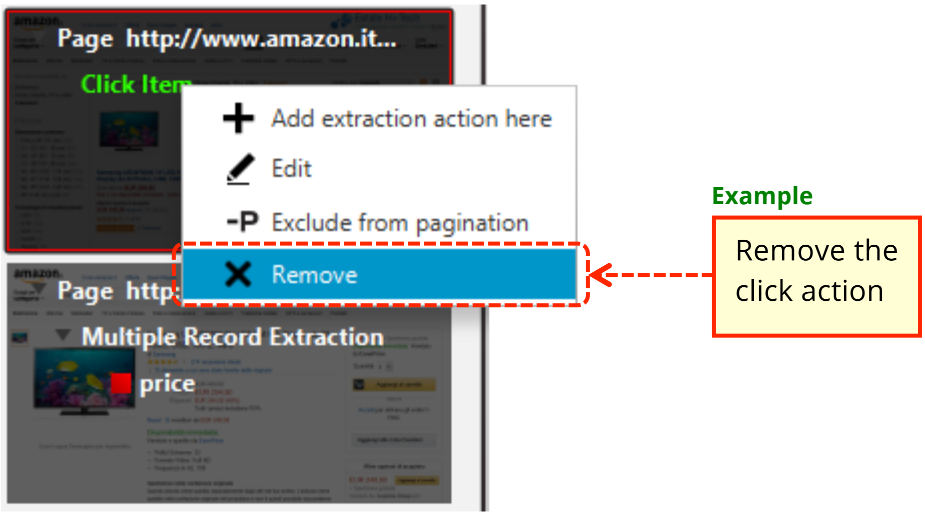
3. Click *Update* and *Close*.

4. Save the task from the File menu to make the changes persistent.

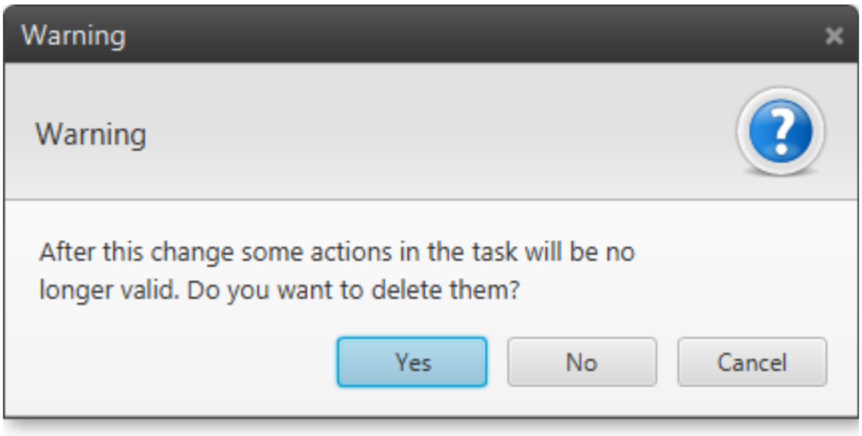
Remove actions

You can remove a recorded action at any time with a right click on it and selecting the Remove command from the context menu. Some actions may be critical because their removal can invalidate other actions. Suppose for example to remove a click action that leaves to another page where other actions were recorded. In this case, these actions will no longer be valid and a warning dialog will notify you. In the following the procedure is shown step by step:

1. Right click on the action that you want to remove and select the command *Remove*


Page http://www.amazon.it...
Click Item
Add extraction action here
Edit
-P Exclude from pagination
Remove
Page http
Multiple Record Extraction
price
Example
Remove the click action

2. A warning message will appear to ask you what to do with the next actions that may no longer be valid.


Warning
Warning

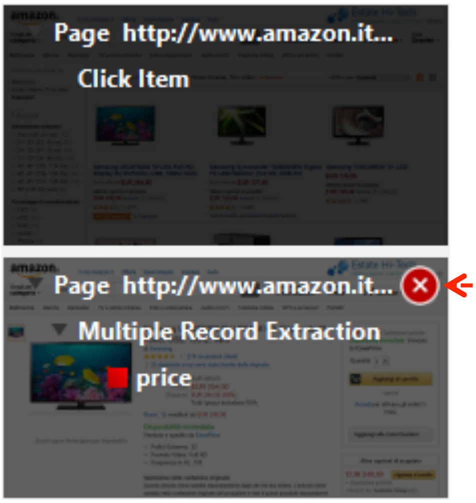
After this change some actions in the task will be no longer valid. Do you want to delete them?

Yes
No
Cancel

3. Click *Yes* if you want to delete the invalid actions, *No* to keep them, Cancel to abort the remove.

4. Save the task from the File menu To make the changes persistent,.

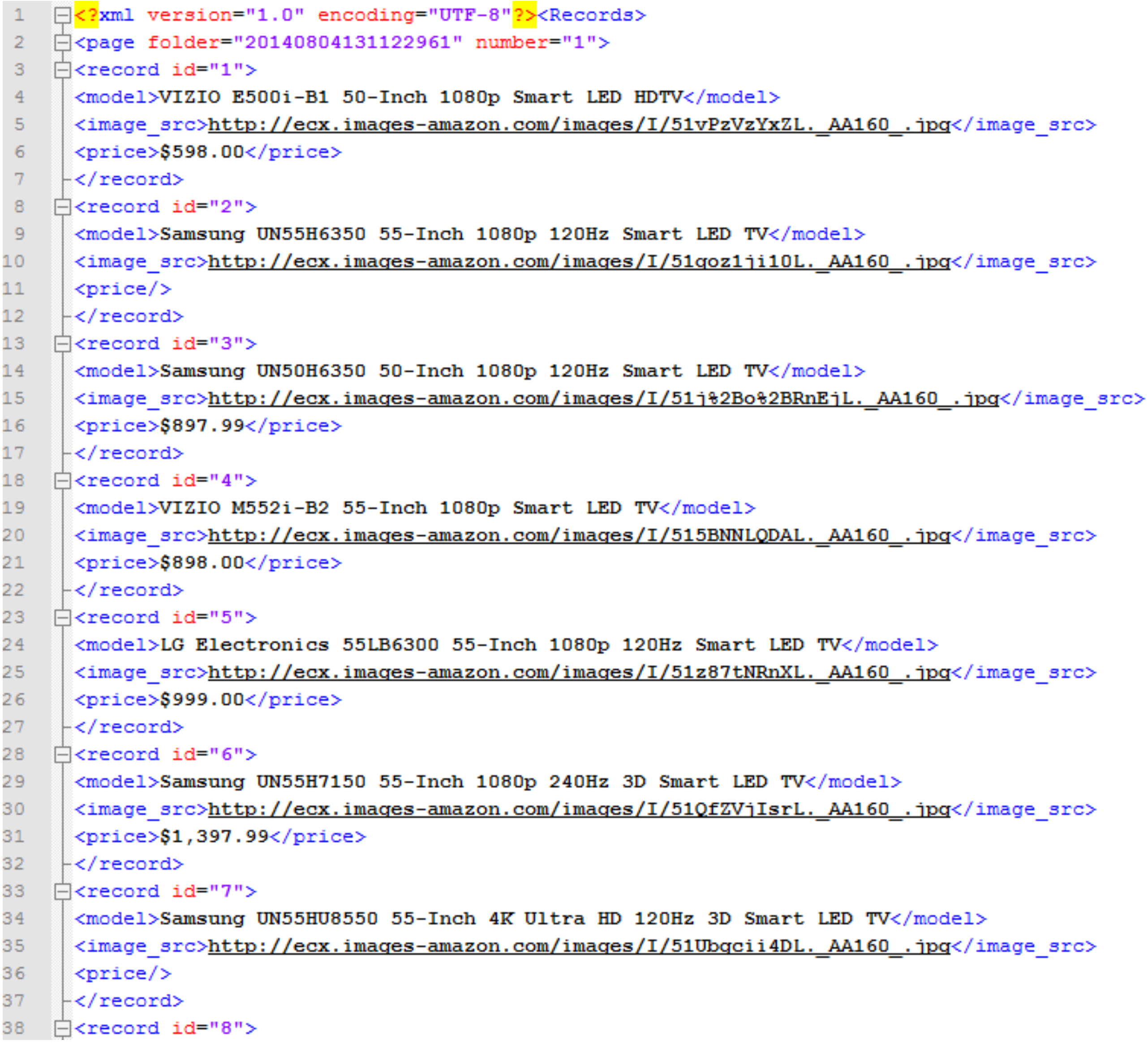
If you want to remove a complete page action, point the mouse at the top-right corner and press the red button that will appear.


Page http://www.amazon.it...
Click Item
Page http://www.amazon.it...
Multiple Record Extraction
price

Click to remove the page action

CHAPTER 5 – RUNNING A TASK

After recording a task you can run it by pressing the button Play in the toolbar. During the task running, the actions previously recorded will be automatically executed. The extracted data are recorded in XML file with the same name of the task (untitled by default). In the example below the extraction result of the amazon TV task is shown.



```
<?xml version="1.0" encoding="UTF-8"?><Records>
<page folder="20140804131122961" number="1">
<record id="1">
<model>VIZIO E500i-B1 50-Inch 1080p Smart LED HDTV</model>
<image_src>http://ecx.images-amazon.com/images/I/51vPzVzYxZL._AA160_.jpg</image_src>
<price>$598.00</price>
</record>
<record id="2">
<model>Samsung UN55H6350 55-Inch 1080p 120Hz Smart LED TV</model>
<image_src>http://ecx.images-amazon.com/images/I/51goz1ji10L._AA160_.jpg</image_src>
<price/>
</record>
<record id="3">
<model>Samsung UN50H6350 50-Inch 1080p 120Hz Smart LED TV</model>
<image_src>http://ecx.images-amazon.com/images/I/51j%2Bo%2BRnEjL._AA160_.jpg</image_src>
<price>$897.99</price>
</record>
<record id="4">
<model>VIZIO M552i-B2 55-Inch 1080p Smart LED TV</model>
<image_src>http://ecx.images-amazon.com/images/I/515BNNLQDAL._AA160_.jpg</image_src>
<price>$898.00</price>
</record>
<record id="5">
<model>LG Electronics 55LB6300 55-Inch 1080p 120Hz Smart LED TV</model>
<image_src>http://ecx.images-amazon.com/images/I/51z87tNRnXL._AA160_.jpg</image_src>
<price>$999.00</price>
</record>
<record id="6">
<model>Samsung UN55H7150 55-Inch 1080p 240Hz 3D Smart LED TV</model>
<image_src>http://ecx.images-amazon.com/images/I/51QfZVjIsrL._AA160_.jpg</image_src>
<price>$1,397.99</price>
</record>
<record id="7">
<model>Samsung UN55HU8550 55-Inch 4K Ultra HD 120Hz 3D Smart LED TV</model>
<image_src>http://ecx.images-amazon.com/images/I/51Ubgcii4DL._AA160_.jpg</image_src>
<price/>
</record>
<record id="8">
```

Figure 13: XML file result of the amazon TV task

CHAPTER 6 – KEYBOARD SHORTCUT

Key combination	Result
Ctrl + O	Open a task
Ctrl + S	Save a task on disk
Ctrl + Shift + S	Save a task on disk as a specific name
Ctrl + R	Start a new task recording
Ctrl + P	Run the task
Ctrl + Q	Stop the task recording/running
Ctrl + ↑	Select the parent element of the current selected element
Ctrl + →	Select the right sibling of the current selected element
Ctrl + ←	Select the left sibling of the current selected element
Ctrl + ↓	Select a random child of the current selected element
Ctrl + X	Close the application

CHAPTER 7 - TROUBLESHOOTING

In this section you can find solutions to the major problems that you might have with the MWE. If you cannot solve your problem, please contact us.

An error occurred during the login

Error type	Cause	Solution
Account not valid	The username/password combination is incorrect	Check that the login credentials entered are correct and try again.
Unable to verify your account	Probably the account file in the MWE folder is no longer available or it is corrupted	Insert again your login credentials
Error during the communication with the server	The MWE is unable to contact the server to validate your account due to your internet connection or firewall	Check that your internet connection is available or add an exception rule in your firewall.
License expired	Your license is expired	Contact us to renew it

An error occurred during the registration

Error type	Cause	Solution
First Name: field cannot be empty	You have not entered your first name	Enter your first name
Last Name: field cannot be empty	You have not entered your last name	Enter your last name
Company: field cannot be empty	You have not entered your company name	Enter your company name
Position: field cannot be empty	You have not entered the position in your company	Enter the position in your company
Email: field cannot be empty	You have not entered your email	Enter your email

Email: invalid email	You have entered a wrong email	Check it and verify that corresponds to your email
Email: email already exists	An account associated to the email entered already exists.	Check that the email entered is correct. If yes, use the forgot password functionality to recover it.
Error occurred sending the registration confirmation email.	Probably something goes wrong during the email sending. The causes may be different.	Check the email field and try again. If the error persist contact us.
Error occurred during the account creation	An error occurred server side during the account creation.	Try again or contact us.