Exercise

Advanced Web Scraping

Summary

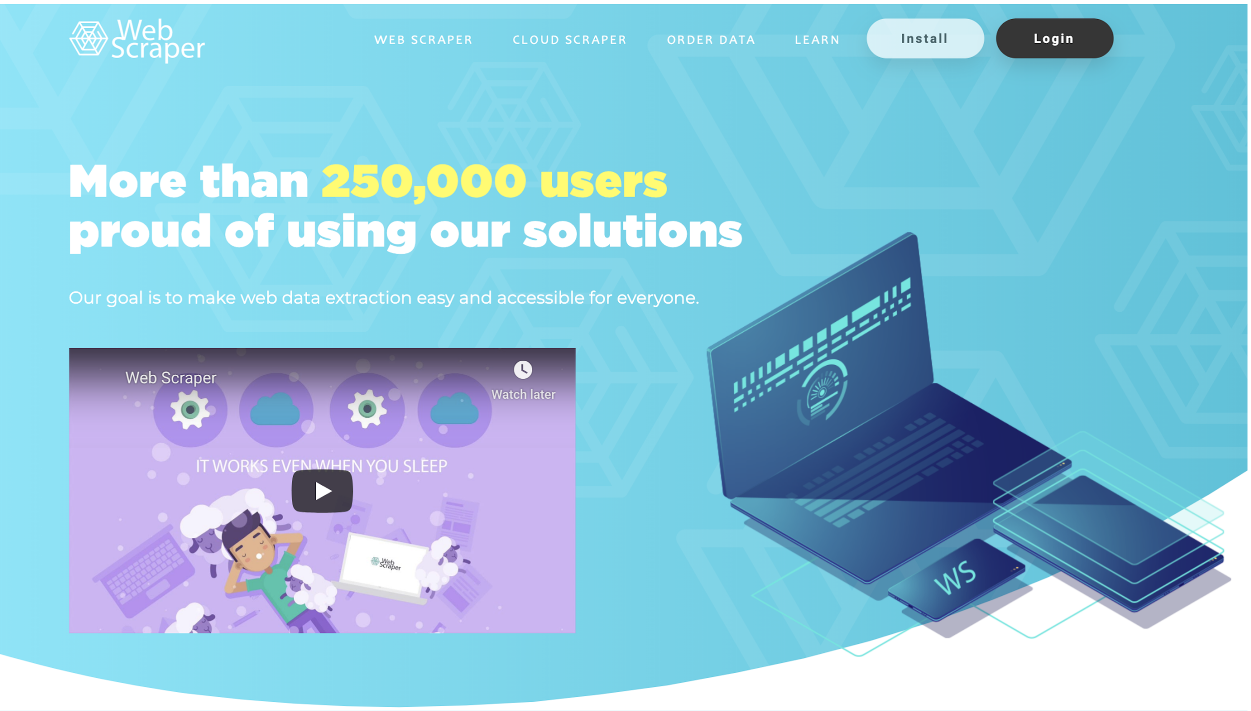
1. Module 2: Finding Data, Data Collection & Data Formats[1]
2. Lesson: Finding Data Online
3. Sub-topic: Data Scraping
4. Objective: Web scraping with webscraper.io
5. Time Allotment: 90 minutes

Steps

First, the webscraping tool that we will be using uses the Chrome web browser, so you have to use Chrome if you already have it installed on your computer, or else you will have to first download and install it.

After you have Chrome running on your computer, go to webscraper.io to install the Web Scraper browser extension.

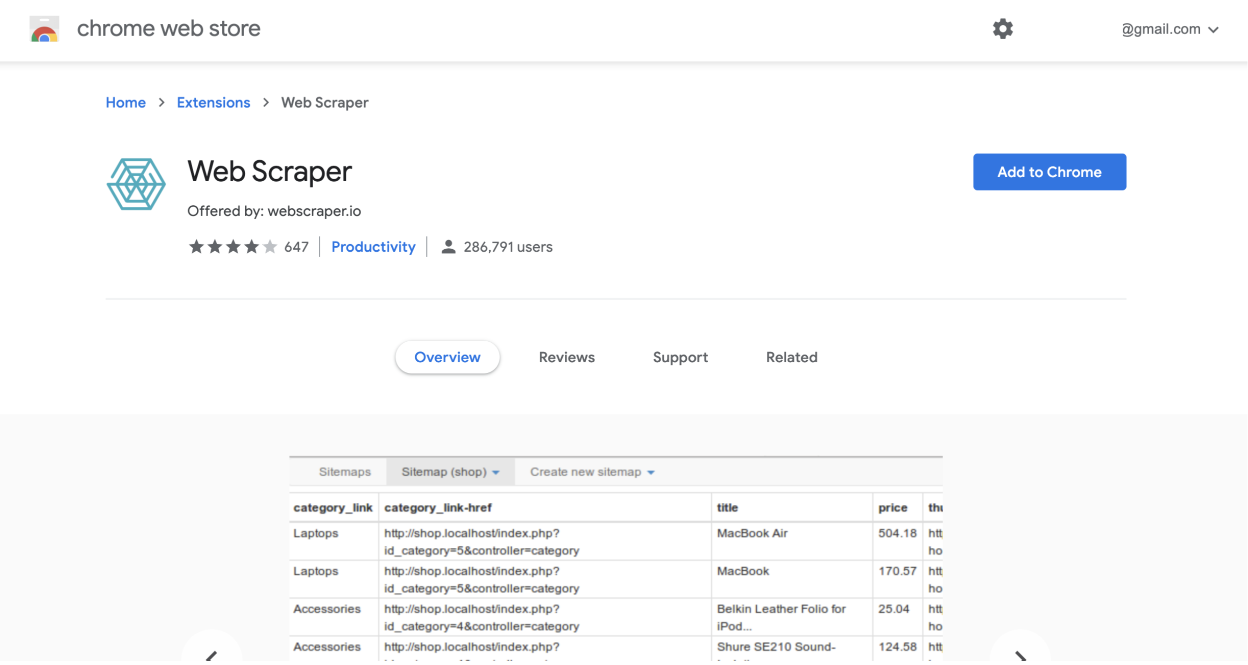
[1] This lesson was adapted from the World Bank's Introduction to Data Literacy training manual by Eva Constantaras, and adapted by Yan Naung Oak, Open Development Cambodia and Open Development Initiative, and is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. For full terms of use, see here.



More than 250,000 users proud of using our solutions

Our goal is to make web data extraction easy and accessible for everyone.

If you click on the “Install” button at the top, you will be taken to the Chrome web store where you will be able to add the Web Scraper extension to your browser.



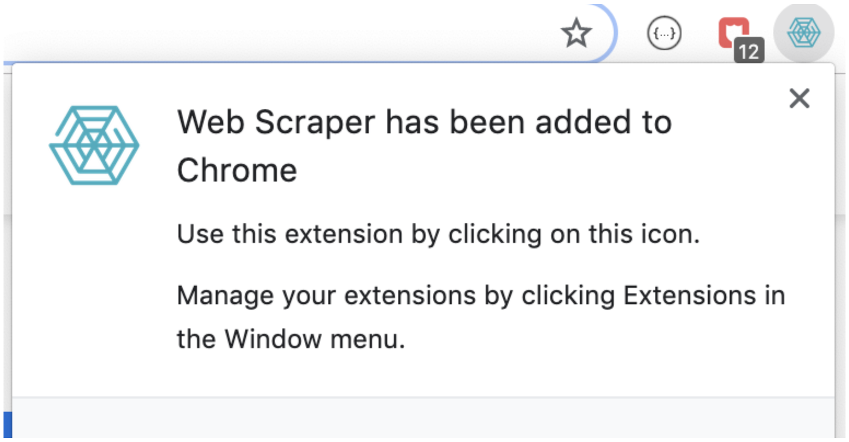
Web Scraper

Offered by: webscraper.io

★★★★☆ 647 | Productivity | 286,791 users

category_link	category_link-href	title	price	thu
Laptops	http://shop.localhost/index.php?id_category=5&controller=category	MacBook Air	504.18	htt ho
Laptops	http://shop.localhost/index.php?id_category=5&controller=category	MacBook	170.57	htt ho
Accessories	http://shop.localhost/index.php?id_category=4&controller=category	Belkin Leather Folio for iPod...	25.04	htt ho
Accessories	http://shop.localhost/index.php?	Shure SE210 Sound-	124.58	htt

Click on “Add to Chrome”, and you will see a message that says it has been added to your browser.

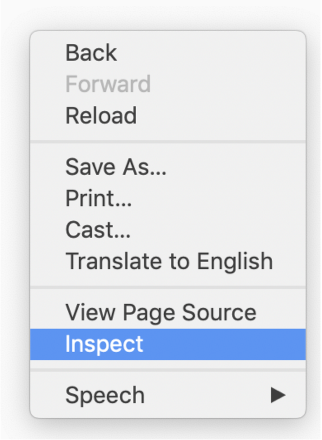


Web Scraper has been added to Chrome

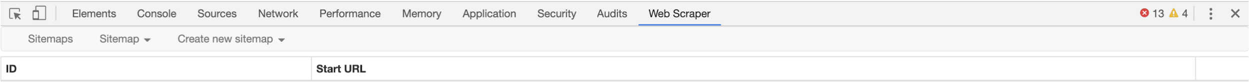
Use this extension by clicking on this icon.

Manage your extensions by clicking Extensions in the Window menu.

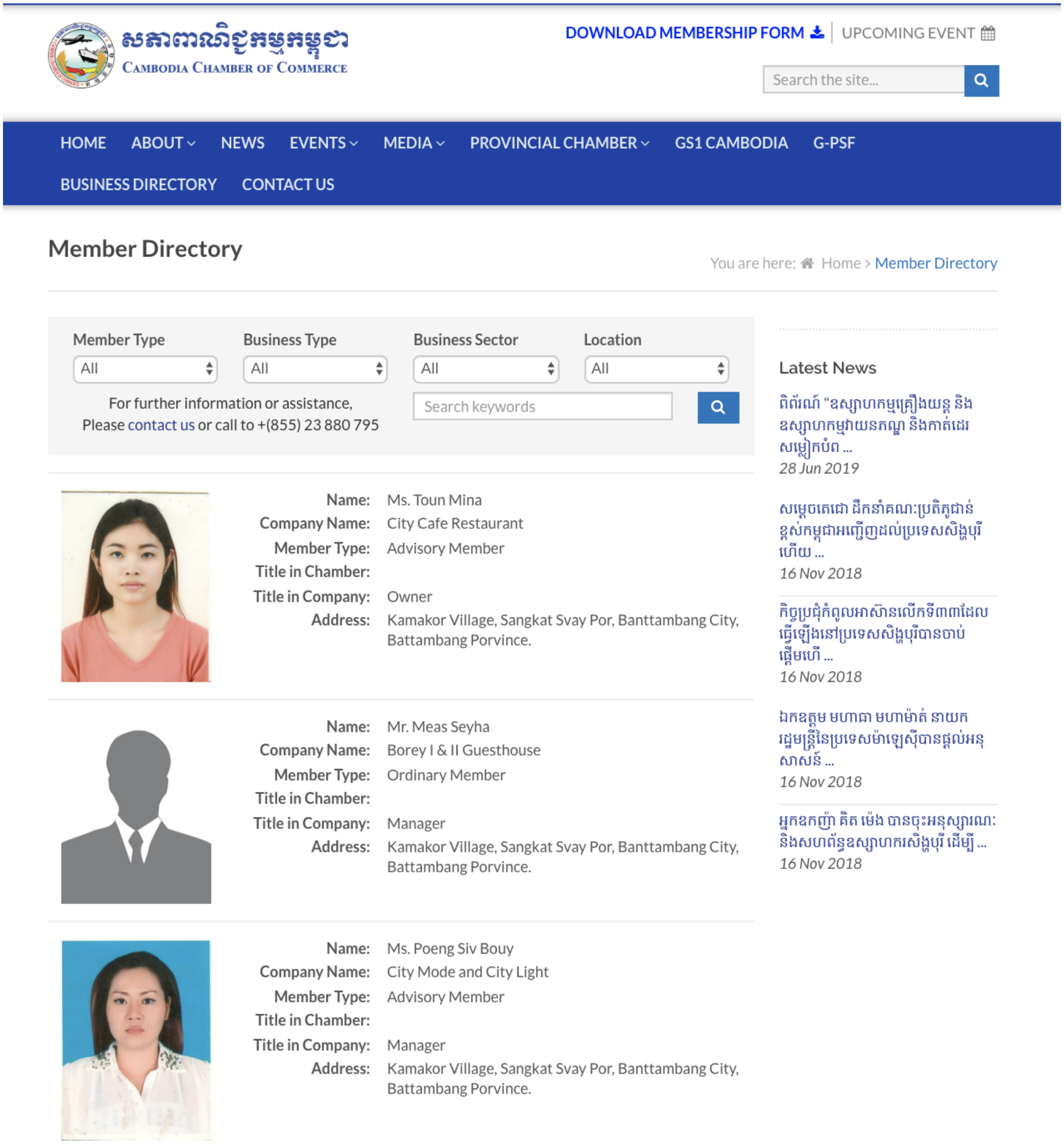
If you right click anywhere on the browser window to bring up the context menu, you can click on “inspect” to bring up the developer tools.



After that you will see a tab called “Web Scraper” appear on the developer tools.



Now let's look at the data that we are going to scrape. We will be scraping the profiles of the members of the Cambodian Chamber of Commerce. As seen below:



សភាពាណិជ្ជកម្មកម្ពុជា
CAMBODIA CHAMBER OF COMMERCE

DOWNLOAD MEMBERSHIP FORM | UPCOMING EVENT

Search the site...

Member Directory

You are here: Home > Member Directory

Member Type	Business Type	Business Sector	Location
All	All	All	All

For further information or assistance,
Please contact us or call to +(855) 23 880 795

Search keywords

Name:	Ms. Toun Mina
Company Name:	City Cafe Restaurant
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Owner
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name:	Mr. Meas Seyha
Company Name:	Borey I & II Guesthouse
Member Type:	Ordinary Member
Title in Chamber:	
Title in Company:	Manager
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name:	Ms. Poeng Siv Bouy
Company Name:	City Mode and City Light
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Manager
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Latest News

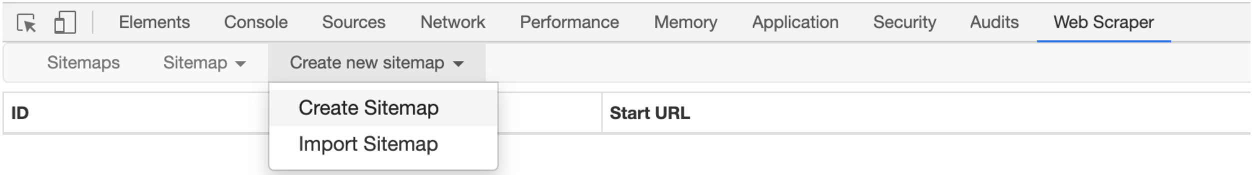
ពិព័រណ៍ "ឧស្សាហកម្មគ្រឿងយន្ត និង ឧស្សាហកម្មវាយនភណ្ឌ និងកាត់ដេរសម្លៀកបំព ...
28 Jun 2019

សម្តេចតេជោ ដឹកនាំគណៈប្រតិភូជាន់ខ្ពស់កម្ពុជាអញ្ជើញដល់ប្រទេសសិង្ហបុរីហើយ ...
16 Nov 2018

កិច្ចប្រជុំកំពូលអាស៊ានលើកទី៣៣ដែលធ្វើឡើងនៅប្រទេសសិង្ហបុរីបានចាប់ផ្តើមហើ ...
16 Nov 2018

ឯកឧត្តម មហាធា មហាម៉ាត់ នាយករដ្ឋមន្ត្រីនៃប្រទេសម៉ាឡេស៊ីបានផ្តល់អនុសាសន៍ ...
16 Nov 2018

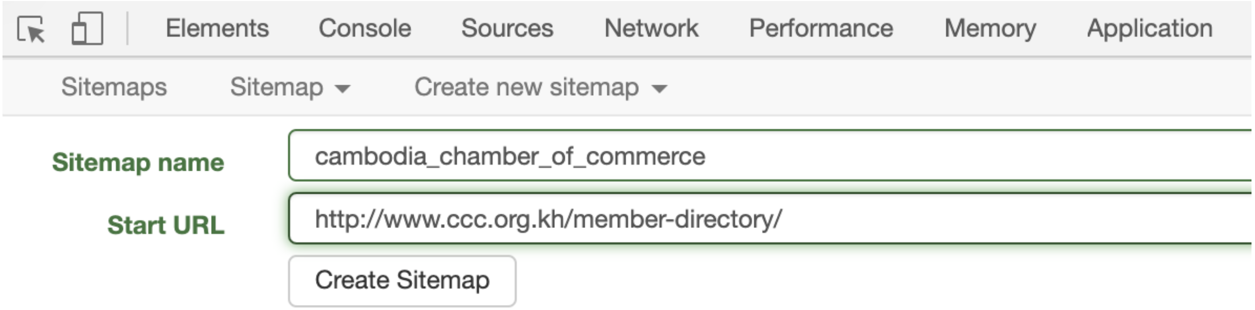
អ្នកឧកញ៉ា គិត ម៉េង បានចុះអនុស្សារណៈ និងសហព័ន្ធឧស្សាហករសិង្ហបុរី ដើម្បី ...
16 Nov 2018



Next, we put in the Sitemap name and Start URL as shown below:

Sitemap name: cambodia_chamber_of_commerce

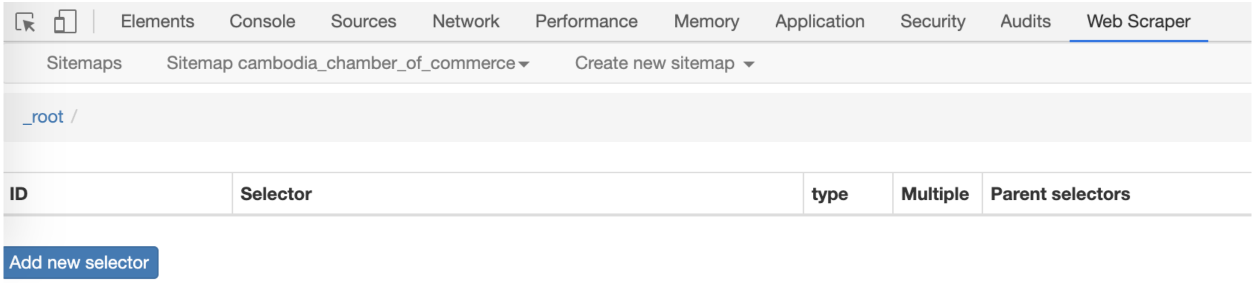
Start URL: http://www.ccc.org.kh/member-directory/



Sitemap name: cambodia_chamber_of_commerce

Start URL: http://www.ccc.org.kh/member-directory/

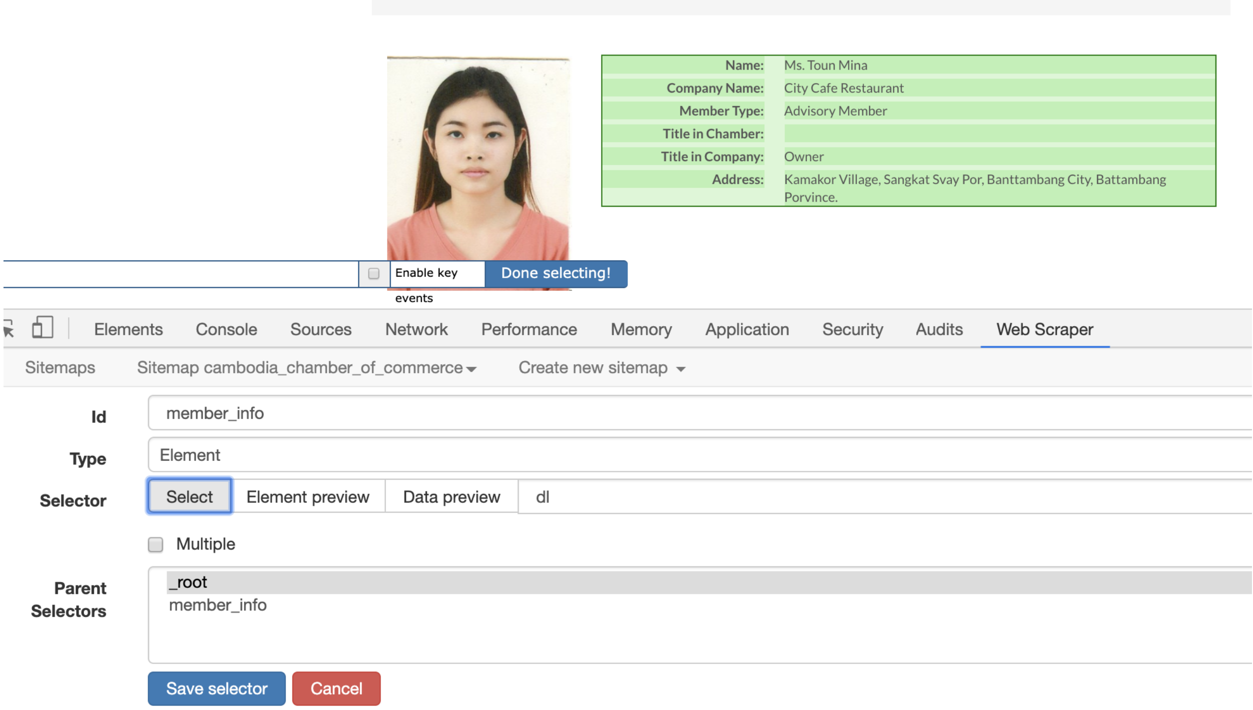
When you click on “Create Sitemap”, a new display appears that lets you add new selectors. A selector is one element on the web page that Web Scraper will scrape, such as a label, or paragraph of text, or the link to an image.



Click on “Add new selector”. This will let you start clicking on elements on the web page to select them.

On the new selector screen, put in “member_info” in the Id box, and select “Element” for Type.

Next, we actually select the element. Click on “Select” and move your mouse over until it is selecting the box that is shown in the screenshot below. This box is the container for the actual information that we want, which is the Name, Company Name, Member Type, etc.



Name:	Ms. Toun Mina
Company Name:	City Cafe Restaurant
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Owner
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Id: member_info

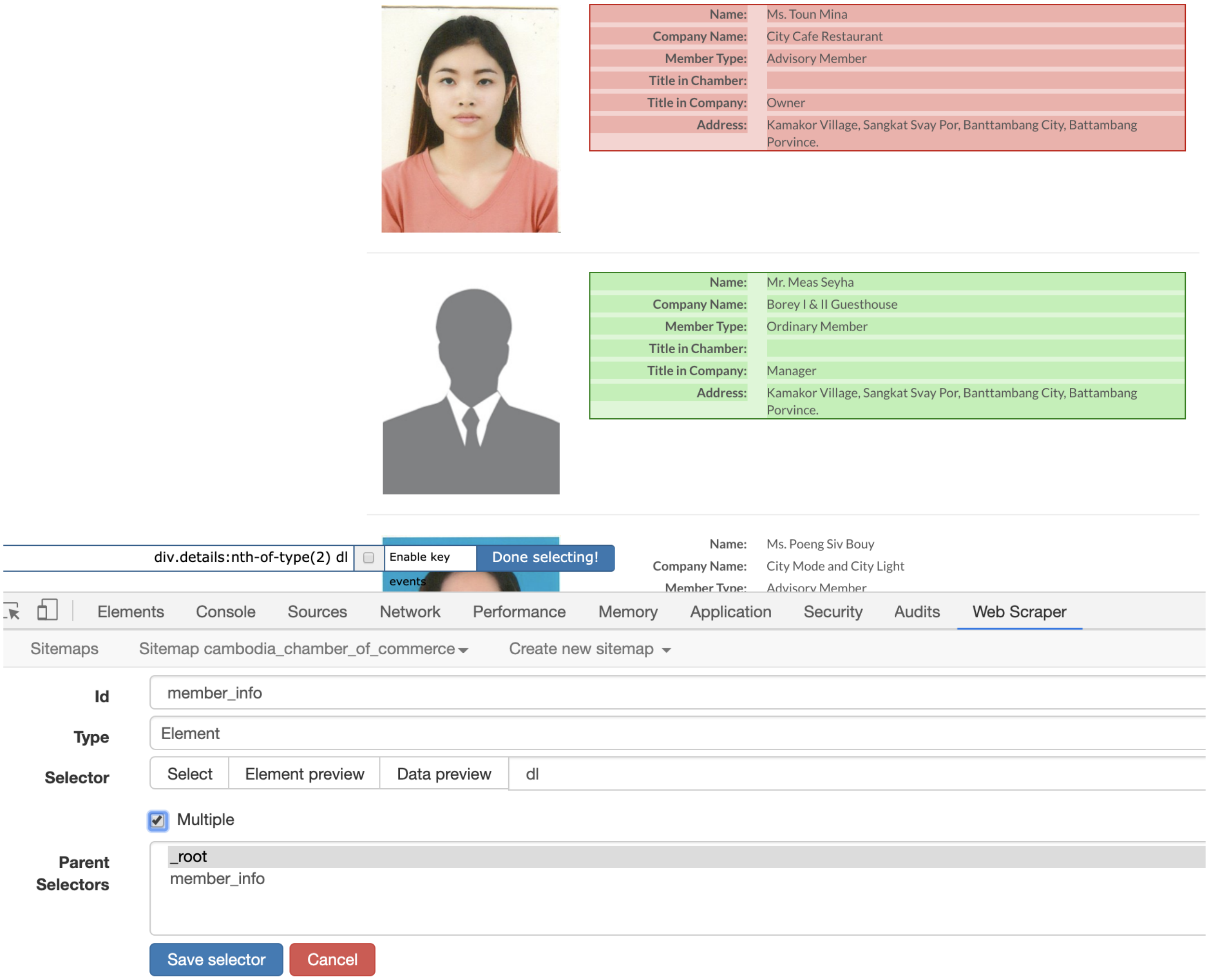
Type: Element

Selector: dl

□ Multiple

Parent Selectors: _root, member_info

Now, since we want to let the Web Scraper know that there are multiple boxes like this that contains the member’s information on this page, we want to click on the “Multiple” check box and after that move your move to select the same box for the next person, as shown below:



Name:	Ms. Toun Mina
Company Name:	City Cafe Restaurant
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Owner
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

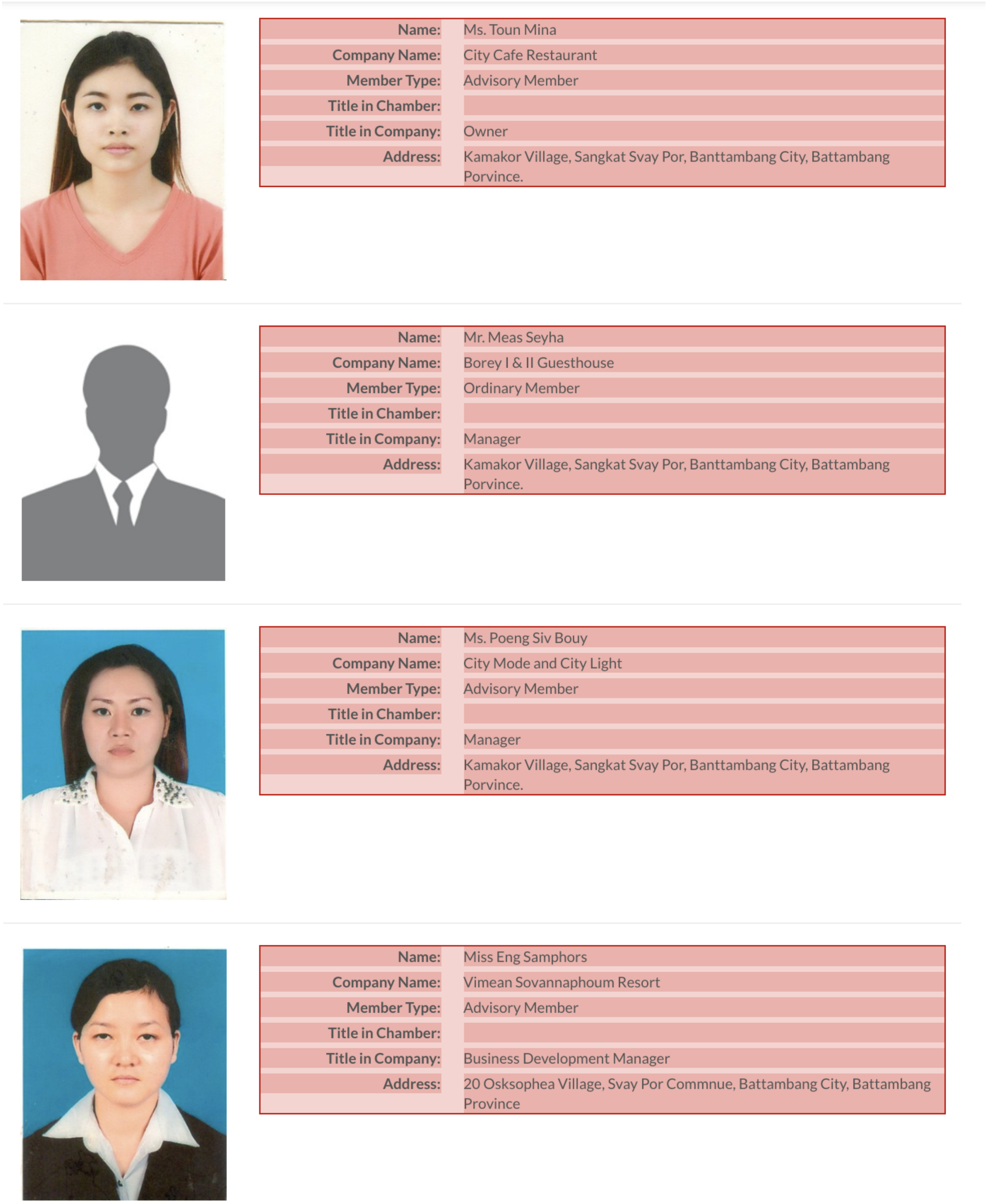
Name:	Mr. Meas Seyha
Company Name:	Borey I & II Guesthouse
Member Type:	Ordinary Member
Title in Chamber:	
Title in Company:	Manager
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name:	Ms. Poeng Siv Bouy
Company Name:	City Mode and City Light
Member Type:	Advisory Member

div.details:nth-of-type(2) dl

Id	member_info
Type	Element
Selector	Select / Element preview / Data preview / dl
	☑ Multiple
Parent Selectors	_root member_info

Now, you will see that all the containers for the information we want on the chamber members will have been selected, as seen below. After that, click on “Done selecting!”.



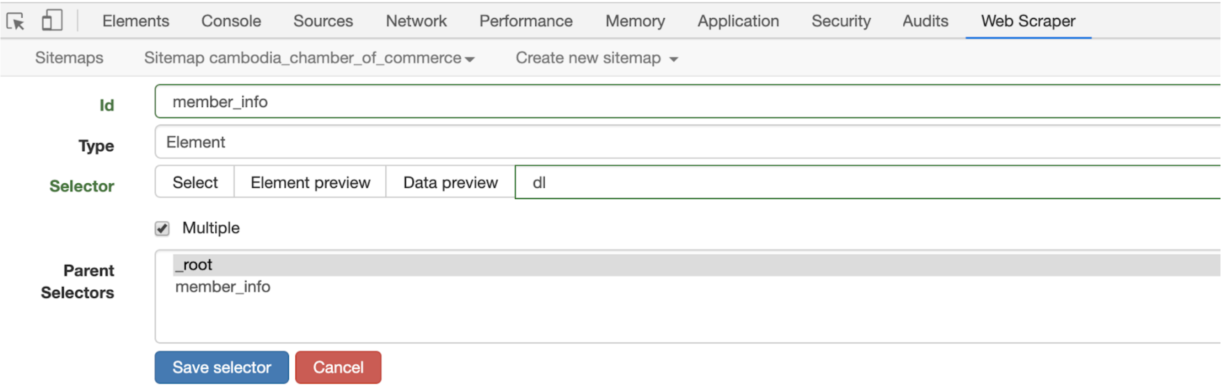
Name:	Ms. Toun Mina
Company Name:	City Cafe Restaurant
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Owner
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name:	Mr. Meas Seyha
Company Name:	Borey I & II Guesthouse
Member Type:	Ordinary Member
Title in Chamber:	
Title in Company:	Manager
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name:	Ms. Poeng Siv Bouy
Company Name:	City Mode and City Light
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Manager
Address:	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name:	Miss Eng Samphors
Company Name:	Vimean Sovannaphoum Resort
Member Type:	Advisory Member
Title in Chamber:	
Title in Company:	Business Development Manager
Address:	20 Osksophea Village, Svay Por Commnue, Battambang City, Battambang Province

Your selector screen should look like the screenshot below. Now click on “Save selector”.



Id: member_info

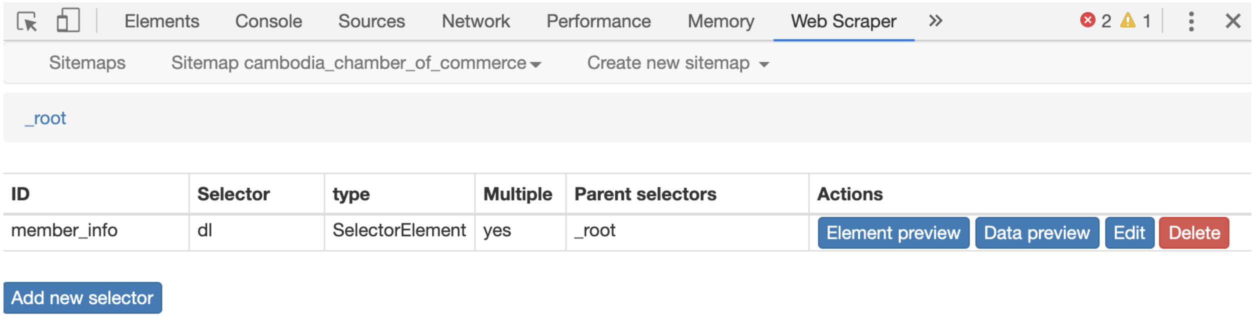
Type: Element

Selector: dl

☑ Multiple

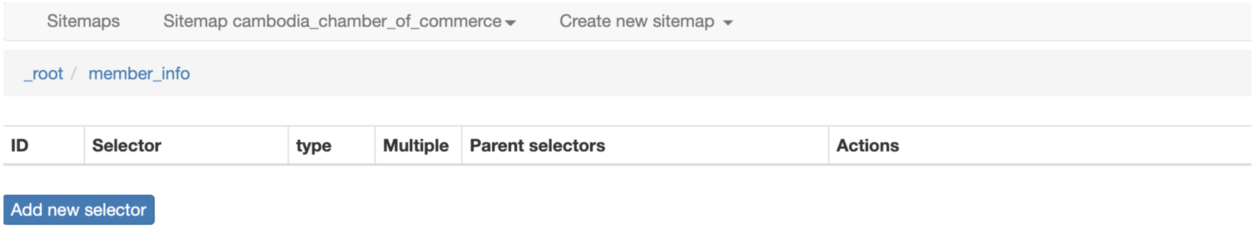
Parent Selectors: _root
member_info

After that, you will be brought back to the list of selectors, which will now have the “member_info” selector that you just created.



ID	Selector	type	Multiple	Parent selectors	Actions
member_info	dl	SelectorElement	yes	_root	Element preview Data preview Edit Delete

Click on the “member_info” row to go into the member_info selector and define items that will be child selectors.



ID	Selector	type	Multiple	Parent selectors	Actions

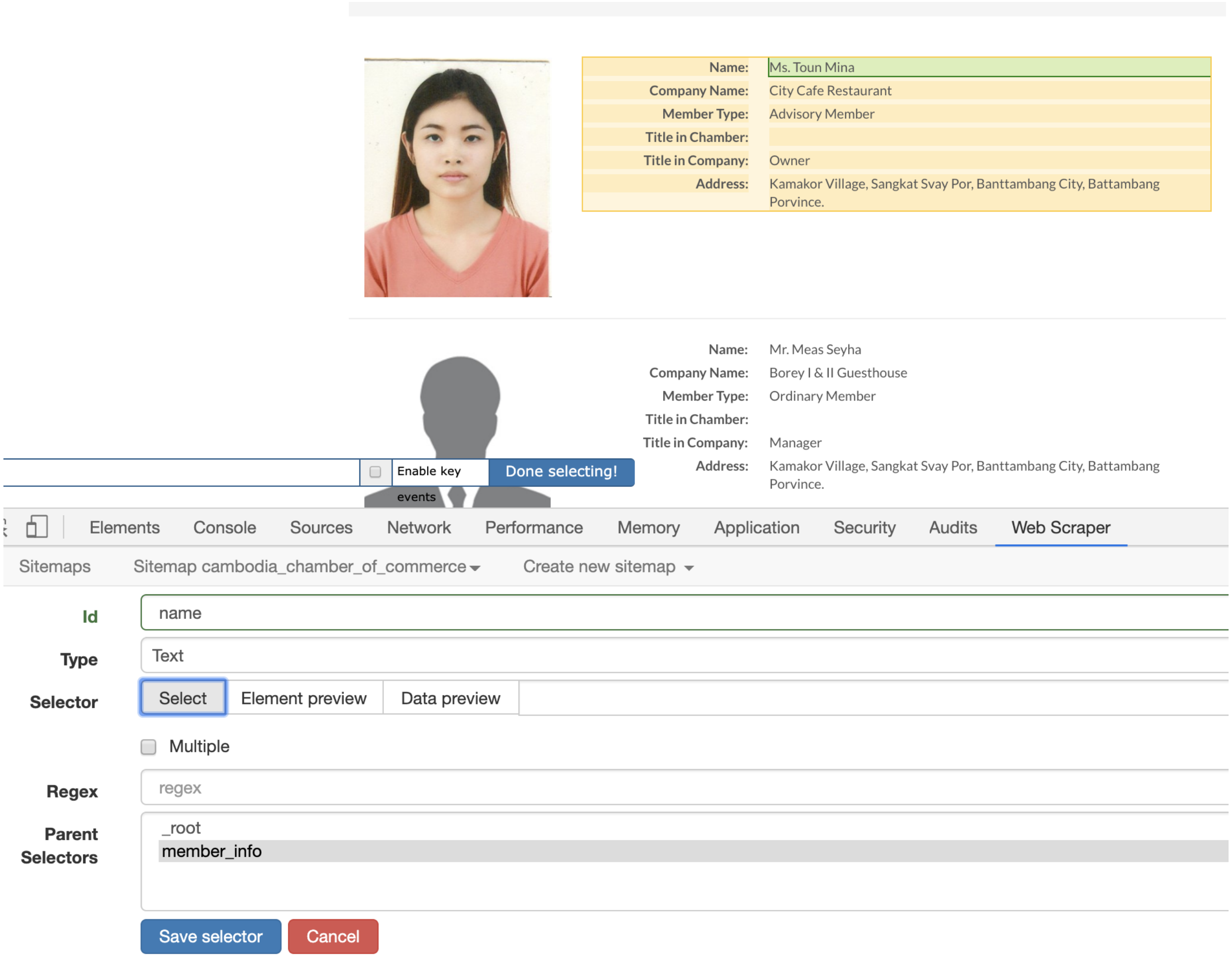
You should see “_root / member_info” at the top bar of the selector list view now. Click on “Add new selector”.

Now we will add a few selectors for each of the items in that container. Notice that since you are selecting “children” of the member_info selector, once you click on “Select”, a yellow box already highlights the places that you can select from.

We will now create a few selectors, starting with Name.

Select the label on the page that shows the chamber member’s name, and type in “name” for the Id, the Type should be selected as “Text”.

Click on “Done selecting!”, and “Save selector” after that.



Name: Ms. Toun Mina
Company Name: City Cafe Restaurant
Member Type: Advisory Member
Title in Chamber:
Title in Company: Owner
Address: Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

Name: Mr. Meas Seyha
Company Name: Borey I & II Guesthouse
Member Type: Ordinary Member
Title in Chamber:
Title in Company: Manager
Address: Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

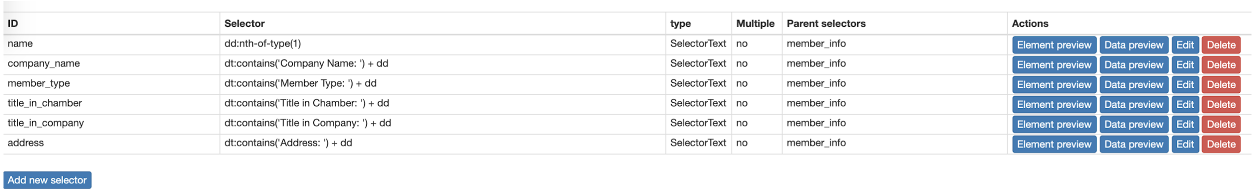
Id: name
Type: Text
Selector
□ Multiple
Regex: regex
Parent Selectors: _root
member_info

Create a few more selectors, as shown below:

- name
- company_name
- member_type
- title_in_chamber
- title_in_company

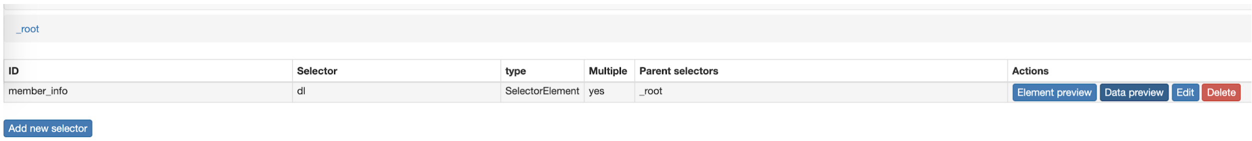
- address

After you have saved all the selectors, you can see the following list under the “member_info” parent selector:



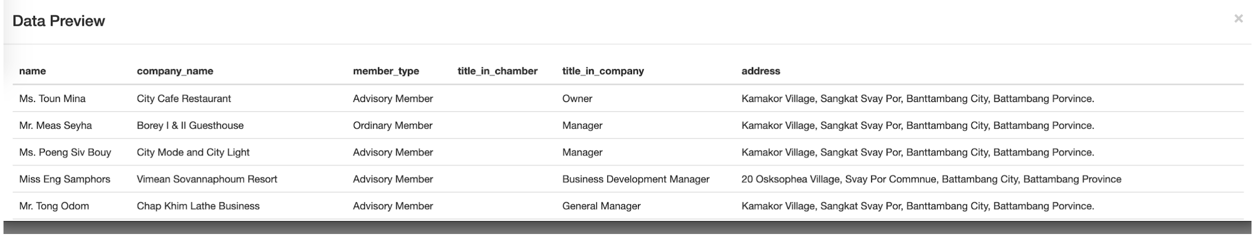
ID	Selector	type	Multiple	Parent selectors	Actions
name	dd:nth-of-type(1)	SelectorText	no	member_info	Element preview Data preview Edit Delete
company_name	dt:contains('Company Name: ') + dd	SelectorText	no	member_info	Element preview Data preview Edit Delete
member_type	dt:contains('Member Type: ') + dd	SelectorText	no	member_info	Element preview Data preview Edit Delete
title_in_chamber	dt:contains('Title in Chamber: ') + dd	SelectorText	no	member_info	Element preview Data preview Edit Delete
title_in_company	dt:contains('Title in Company: ') + dd	SelectorText	no	member_info	Element preview Data preview Edit Delete
address	dt:contains('Address: ') + dd	SelectorText	no	member_info	Element preview Data preview Edit Delete

Click on “_root” in the top bar to get back to the original list of selectors.



ID	Selector	type	Multiple	Parent selectors	Actions
member_info	dl	SelectorElement	yes	_root	Element preview Data preview Edit Delete

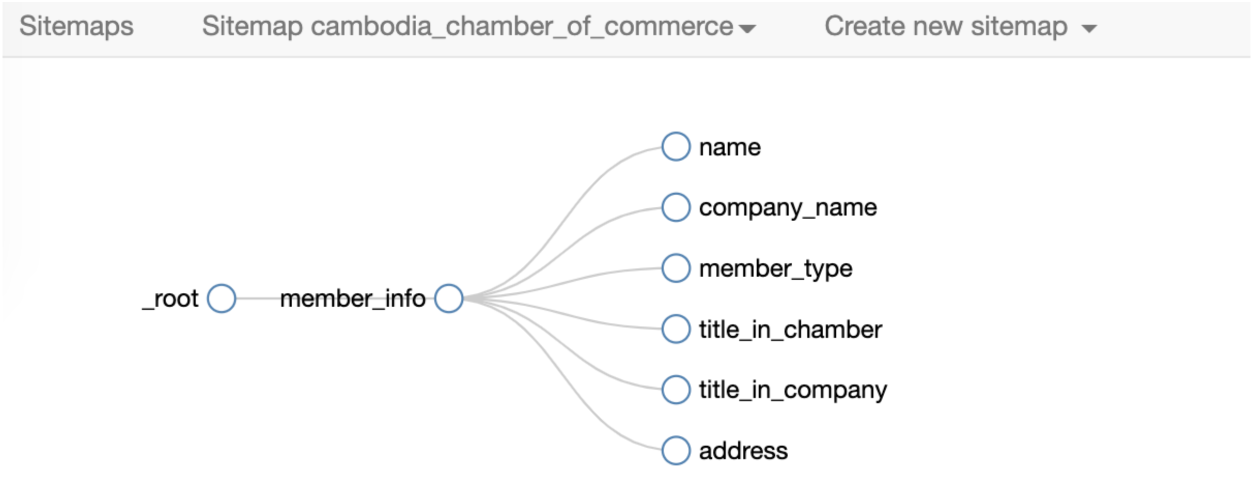
Now, click on “Data preview” in the “member_info” row, and you will see a preview of the data that will be scraped.



Data Preview

name	company_name	member_type	title_in_chamber	title_in_company	address
Ms. Toun Mina	City Cafe Restaurant	Advisory Member		Owner	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.
Mr. Meas Seyha	Borey I & II Guesthouse	Ordinary Member		Manager	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.
Ms. Poeng Siv Bouy	City Mode and City Light	Advisory Member		Manager	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.
Miss Eng Samphors	Vimean Sovannaphoum Resort	Advisory Member		Business Development Manager	20 Osksophea Village, Svay Por Commnue, Battambang City, Battambang Province
Mr. Tong Odom	Chap Khim Lathe Business	Advisory Member		General Manager	Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

You can look at the sitemap that we have created in the form of a graph by choosing Selector Graph under the sitemap menu. Here you can see the parent-child relationships that you have created for your selectors very clearly.

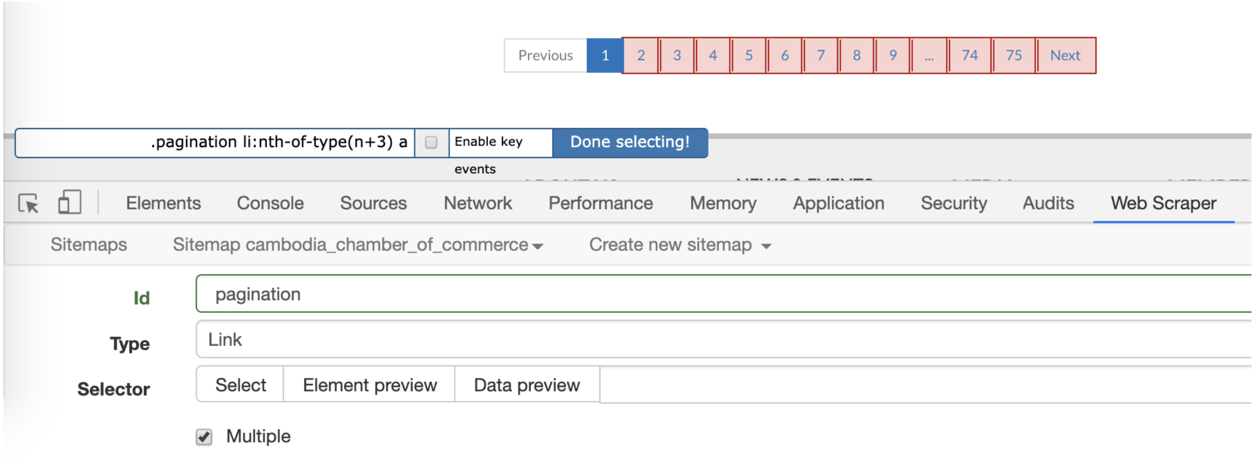


So far, we've successfully scraped the data from one webpage, but notice that this website has multiple pages, as shown by the page selector at the bottom of the page. How will we scrape data from these other pages as well?

We are using the Web Scraper plugin for this exercise precisely because it allows us to easily scrape data that is spread out over multiple pages.

First, go back to the "_root" page of the Web Scraper view, and click on "Add new selector".

For this selector, put in "pagination" as the Id, and choose "Link" for type. You must also click on the "Multiple" checkbox. Next, click on the "Select" button, and choose all the page links that are at the bottom of the screen, as shown in the screenshot below.



.pagination li:nth-of-type(n+3) a

□ Enable key events

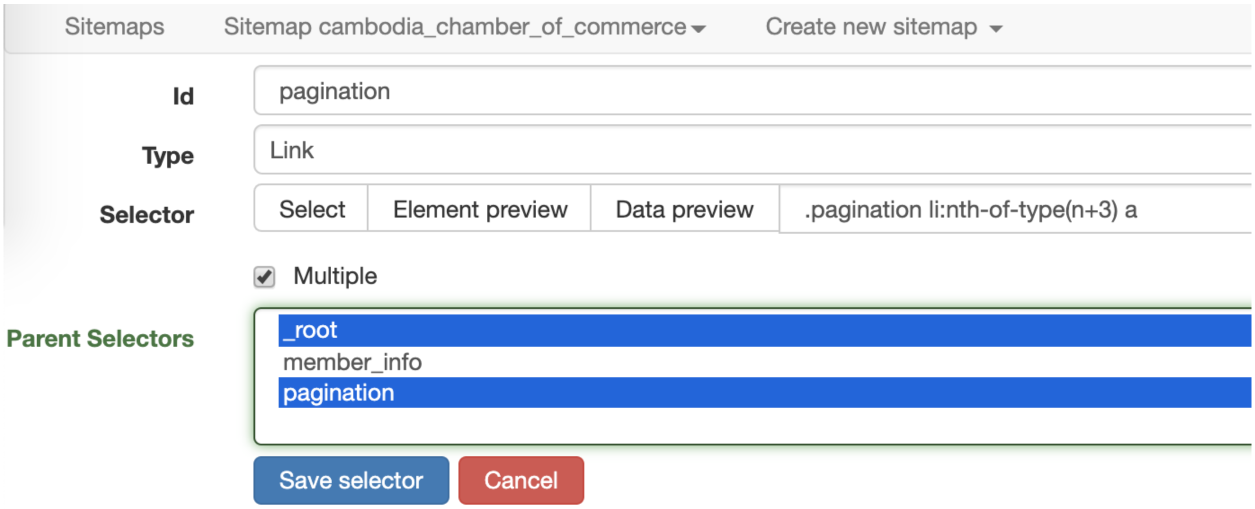
Id: pagination

Type: Link

Selector

☑ Multiple

Another very important step is to choose both “_root” and “pagination” as Parent Selectors. You can do this by holding down the “Ctrl” key (if you’re on a Windows computer) or the “Command” key (if you’re on a Mac) while you click.



Id: pagination

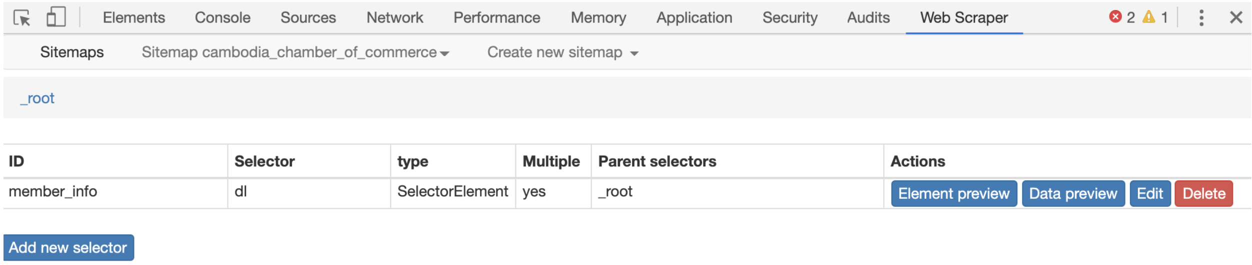
Type: Link

Selector: .pagination li:nth-of-type(n+3) a

☑ Multiple

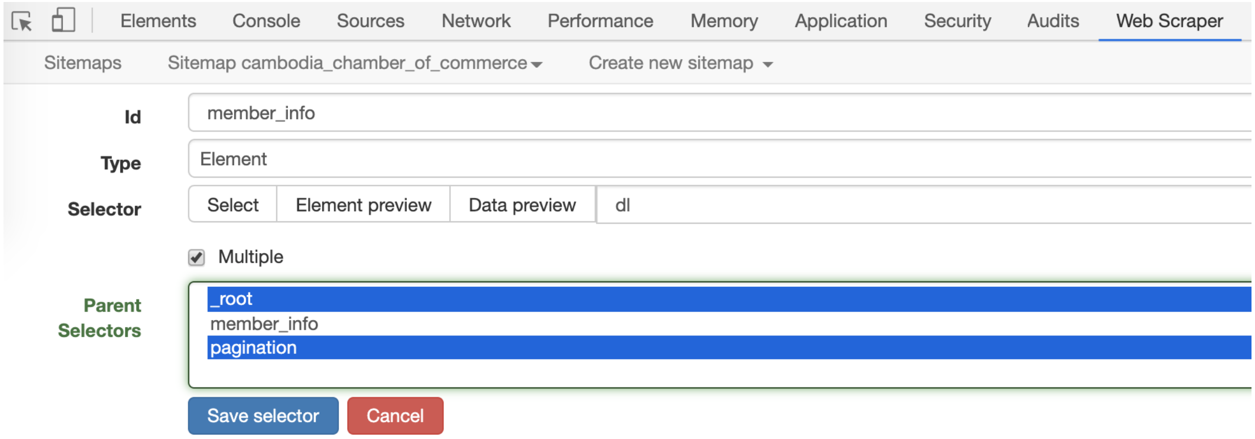
Parent Selectors: _root (selected), member_info, pagination (selected)

You also need to go back to your “member_info” selector and choose both “pagination” and “_root” as its parent selectors. To do this, first go back to the “_root” page, and then click on “Edit” on the “member_info” row.



ID	Selector	type	Multiple	Parent selectors	Actions
member_info	dl	SelectorElement	yes	_root	Element preview Data preview Edit Delete

Now select both “pagination” and “_root” as parent selectors, and click on “Save selector”.



Id: member_info

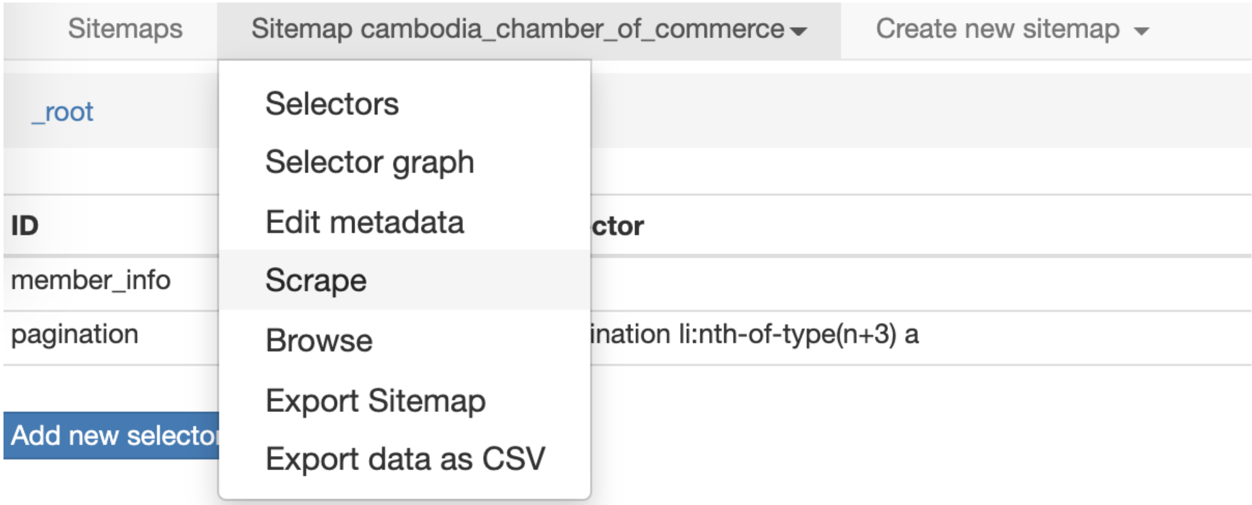
Type: Element

Selector: dl

☑ Multiple

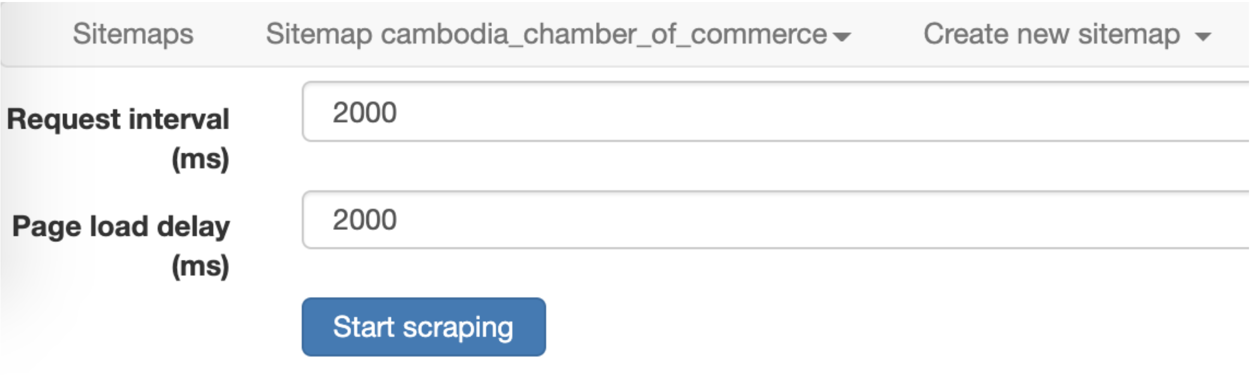
Parent Selectors: _root, member_info, pagination

Now it's time to start your scraper. Click on the "Sitemap cambodia_chamber_of_commerce" menu and select "Scrape".



ID	ctor
member_info	
pagination	ination li:nth-of-type(n+3) a

A new screen that pops up will ask you to fill in the "Request Interval" and "Page load delay". This sets how much the time scraper should wait before requesting additional pages. You can just leave both at the default values of 2000ms. Now click on "Start scraping".

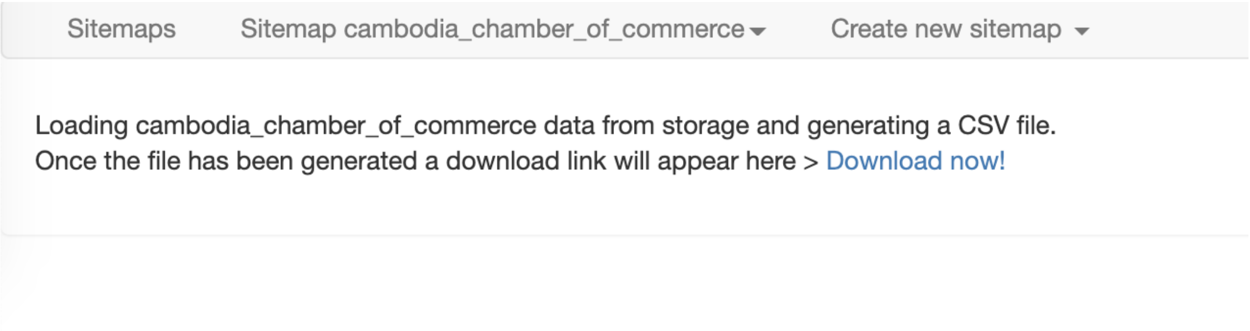


Request interval (ms): 2000

Page load delay (ms): 2000

Once you start scraping, a new pop-up window will start loading all the pages on the site that you have requested and start collecting the data according to what you configured in the selectors. This will take a few minutes since the scraper will have to go through all the pages on the site.

After the scraping is finished, you will see a message as shown in the screen shot below with a link to the Download of the data in CSV format.



Loading cambodia_chamber_of_commerce data from storage and generating a CSV file.
Once the file has been generated a download link will appear here > Download now!

Further Practice

Try to scrape another multi-page website of your choosing.