

Exercise

การดึงข้อมูลจากเว็บมาใช้งาน

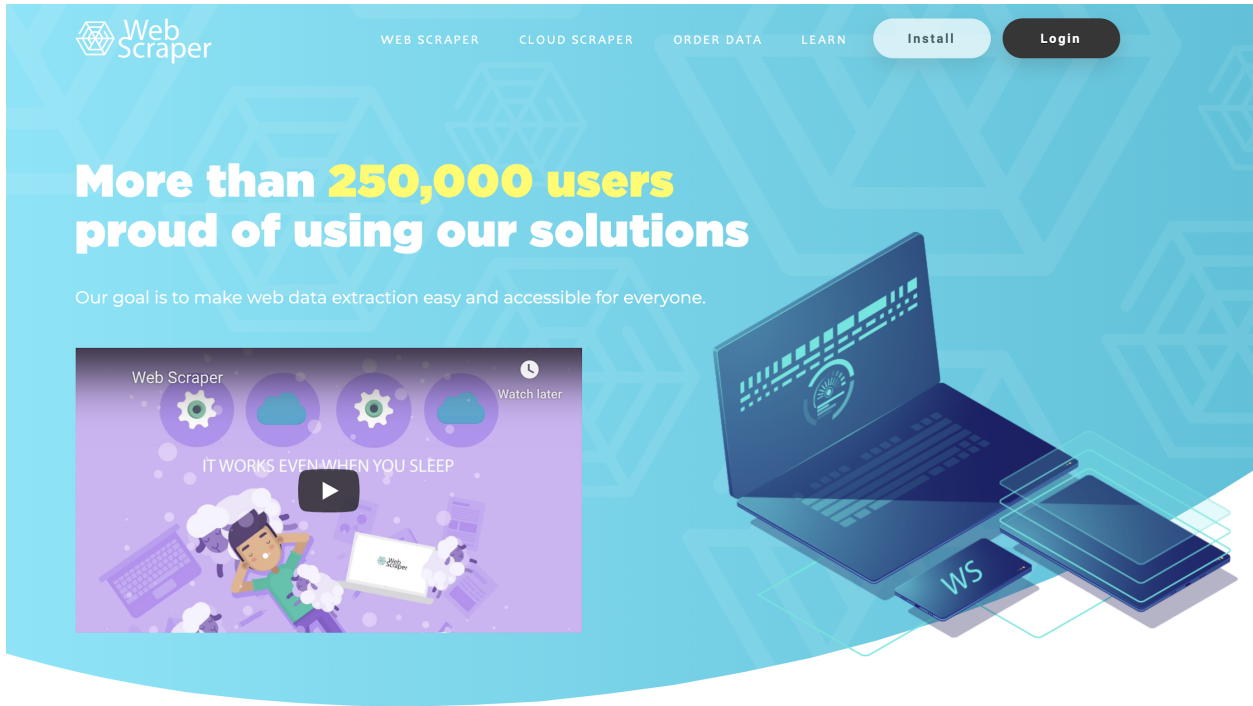
สาระสำคัญ

1. โมดูล: ค้นหา รวบรวม และจัดทำรูปแบบข้อมูลที่เหมาะสม¹
2. บทเรียน: หาข้อมูลออนไลน์
3. หัวข้อย่อย: การดึงข้อมูลจากเว็บมาใช้งาน
4. วัตถุประสงค์: ดึงข้อมูลจากเว็บด้วย webscraper.io
5. เวลา: 45-60 นาที

ขั้นตอน

1. ขั้นแรก เครื่องมือ webscraping ที่เราจะใช้นั้นจะใช้เว็บเบราว์เซอร์ [Chrome](#) เป็นหลัก ดังนั้นคุณต้องใช้เบราว์เซอร์ Chrome หากยังไม่มีคุณจะต้องดาวน์โหลดและติดตั้งเสียก่อน
2. หลังจากคุณเปิดใช้ Chrome ในคอมพิวเตอร์ของคุณแล้ว ให้ไปที่ [webscraper.io](#) เพื่อติดตั้งส่วนขยายเบราว์เซอร์ Web Scraper

¹ เอกสารนี้ดัดแปลงมาจากคู่มือการอบรม Introduction to Data Literacy ของธนาคารโลก โดย Eva Constantaras และปรับปรุงโดย Yan Naung Oak, Open Development Cambodia และ Open Development Initiative ซึ่งได้รับอนุญาตภายใต้ [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International](#)



3. หากคุณคลิกที่ปุ่ม "[ติดตั้ง](#)" ที่ด้านบนคุณจะถูกนำไปที่ Chrome เว็บสโตร์ซึ่งคุณสามารถเพิ่ม Web Scraper extension ลงในเบราว์เซอร์ของคุณ
4. คุณอาจจะต้องปิด - เปิดเพื่อเริ่ม Google Chrome ใหม่ (ขั้นตอนเพิ่มเติม)

chrome web store @gmail.com

Home > Extensions > Web Scraper

Web Scraper

Offered by: webscraper.io

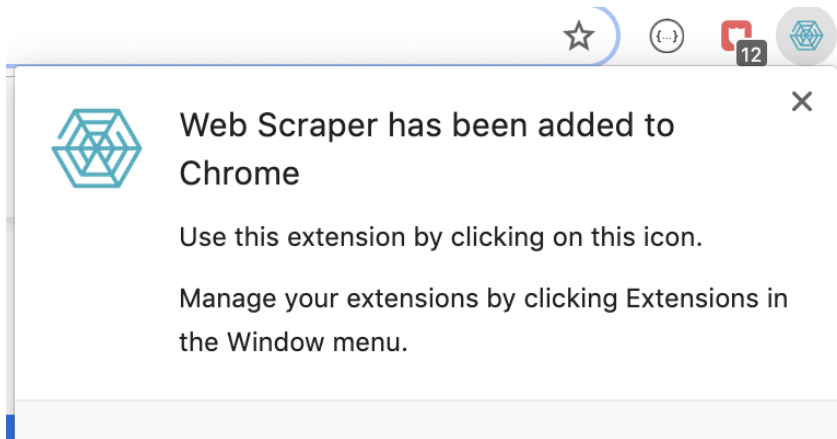
★★★★☆ 647 | Productivity | 286,791 users

Add to Chrome

Overview
Reviews
Support
Related

category_link	category_link-href	title	price	thi
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


5. คลิกที่ "เพิ่มลงใน Chrome" (Add to Chrome) แล้วคุณ将会เห็นข้อความแจ้งว่า Web Scraper ใต้รับการเพิ่มลงในเบราว์เซอร์ของคุณแล้ว



6. ทีนี้มาดูข้อมูลที่เรากำลังจะดึงมาใช้กัน เราจะดึงข้อมูล [members of the Cambodian Chamber of Commerce](#) ([ทดลองขั้นสูง](#)) ตามตัวอย่างด้านล่าง:

Member Directory

You are here: Home > Member Directory

Member Type All	Business Type All	Business Sector All	Location All
For further information or assistance, Please contact us or call to +(855) 23 880 795		Search keywords	

Latest News

ពិធីប្រជុំ "ឧស្សាហកម្មគ្រឿងយន្ត និង ឧស្សាហកម្មវាយតម្លៃ និងកាត់ដេរ សម្លៀកបំពាក់ ..."
28 Jun 2019

សម្តេចតេជោ ដឹកនាំគណៈប្រតិភូជាន់ខ្ពស់កម្ពុជាអញ្ជើញដល់ប្រទេសសិង្ហបុរី ហើយ ...
16 Nov 2018

កិច្ចប្រជុំកំពូលអាស៊ានលើកទី៣៣ដែលធ្វើឡើងនៅប្រទេសសិង្ហបុរីបានចាប់ផ្តើមហើយ ...
16 Nov 2018

ឯកឧត្តម មហាធា មហាម៉ាត់ នាយករដ្ឋមន្ត្រីនៃប្រទេសម៉ាឡេស៊ីបានផ្តល់អនុសាសន៍ ...
16 Nov 2018

អ្នកឧកញ៉ា គិត ម៉េង បានចុះអនុស្សាវរណៈ និងសហព័ន្ធនៃឧស្សាហកម្មសិង្ហបុរី ដើម្បី ...
16 Nov 2018



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.

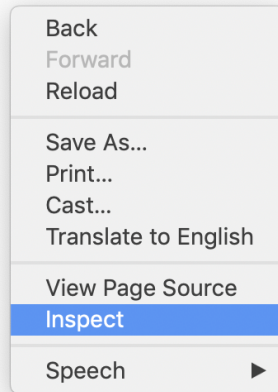


Name: Mr. Meas Seyha
Company Name: Borey I & II Guesthouse
Member Type: Ordinary Member
Title in Chamber:
Title in Company: Manager
Address: Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

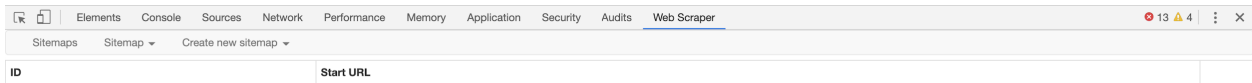


Name: Ms. Poeng Siv Bouy
Company Name: City Mode and City Light
Member Type: Advisory Member
Title in Chamber:
Title in Company: Manager
Address: Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince.

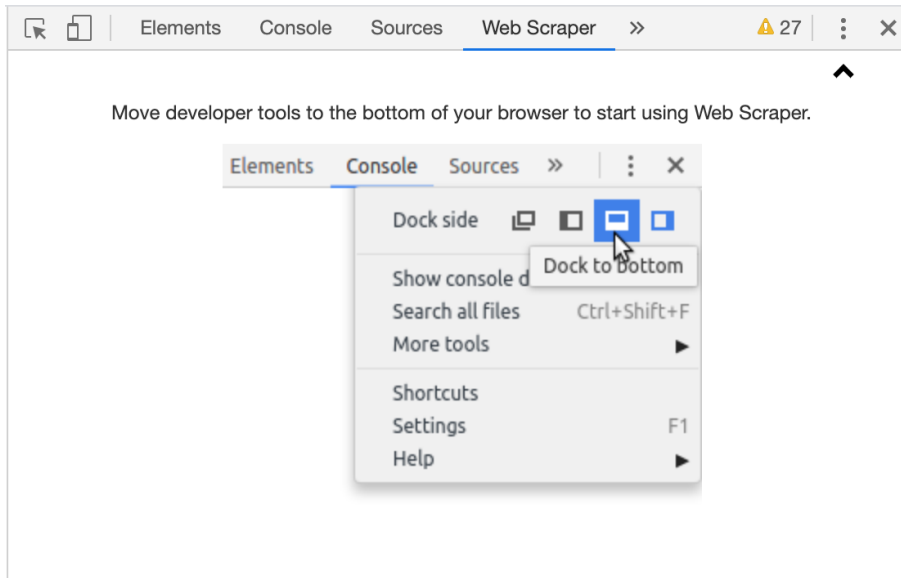
1. កុំចុច F12 ដើម្បីបើក devtool ហើយជ្រើសរើស 'Web scraper'
2. បើកពិនិត្យលើកុំប្រើប្រាស់កូដដែលបានផ្តល់ឱ្យ ដើម្បីបើកប្រព័ន្ធគ្រប់គ្រងគេហទំព័រ ហើយបើកពិនិត្យលើកុំប្រើប្រាស់កូដដែលបានផ្តល់ឱ្យ (inspect) ដើម្បីបើកប្រព័ន្ធគ្រប់គ្រងគេហទំព័រឡើងវិញ



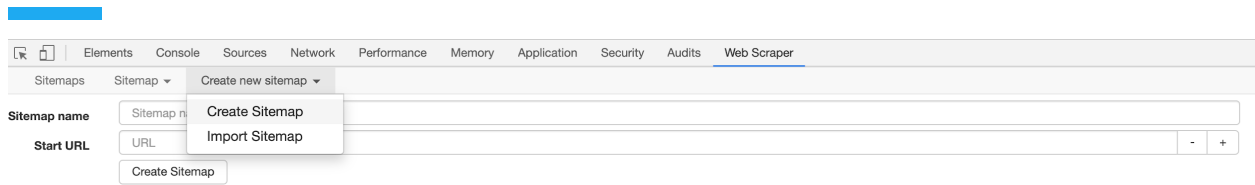
3. หลังจากนั้น คุณจะเห็นแท็บ“ Web Scaper” ปรากฏอยู่ในเครื่องมือสำหรับนักพัฒนา



4. คลิกไอคอน hamburger (...) เพื่อเลือกการแสดงผลหน้าต่าง "Dock side" ไปที่ด้านล่างของหน้าต่างเบราว์เซอร์



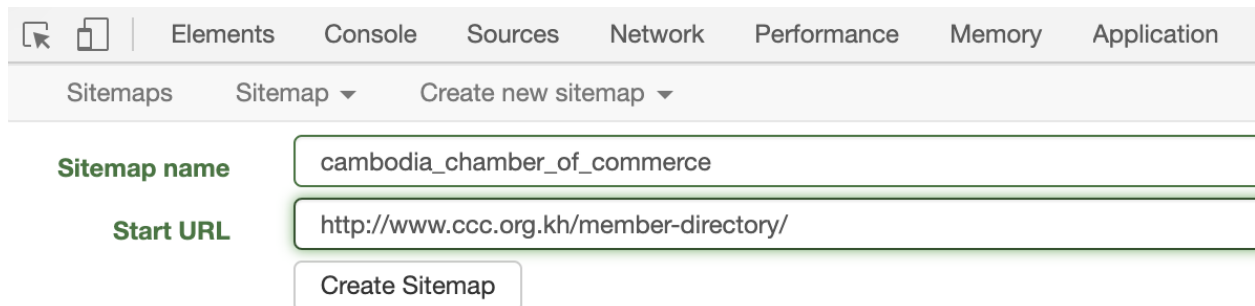
5. ไปที่ “สร้างแผนผังไซต์ใหม่” > “สร้างแผนผังไซต์”



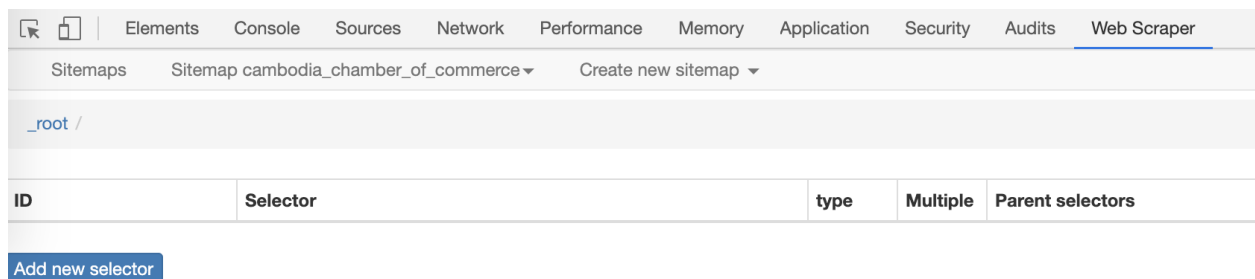
หลังจากเลือกสร้างไซต์แมปใหม่ “Create Sitemap” ให้กรอกชื่อดังนี้:

ให้กรอกในช่อง Sitemap name: **cambodia_chamber_of_commerce**

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



และกดปุ่ม “Create Sitemap” ระบบจะแสดงหน้าจอใหม่ให้คุณเลือกกดปุ่ม “Add new selector” เพื่อเลือกข้อมูลบนเว็บให้ระบบสคริปข้อมูล ไม่ว่าจะป็นป้ายชื่อ ข้อความทั้งย่อหน้า หรือลิงก์ของรูป

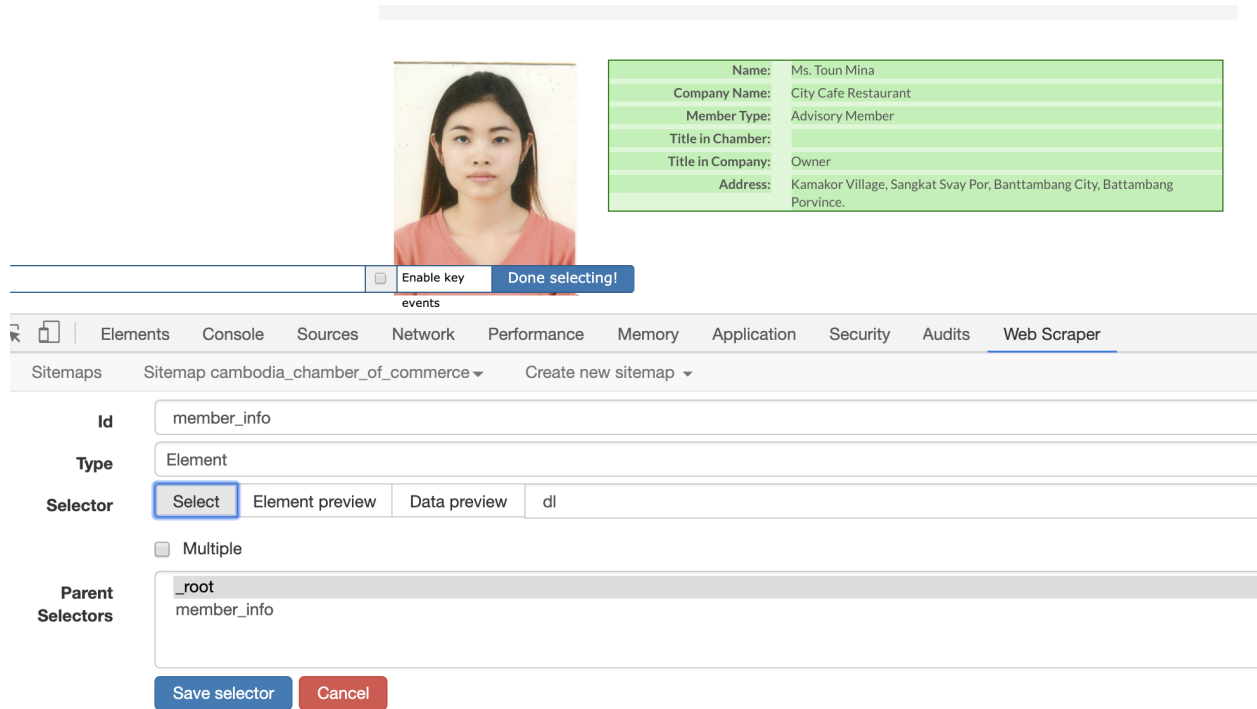


ให้กดปุ่ม “Add new selector”

ระบบจะแสดงหน้าจอให้ป้อนข้อมูลการเลือกองค์ประกอบที่จะดึงจากเว็บ

ป้อนในช่อง Id box เป็น “member_info” และเลือกในช่อง Type เป็น “Element”

จากนั้น กดปุ่ม “Select” และเลื่อนเมาส์ของคุณมาบริเวณองค์ประกอบของเว็บที่คุณต้องการดึงข้อมูล คุณจะเห็นแถบกล่องสีปรากฏ และเลื่อนเมาส์ให้ครอบคลุมข้อมูลที่ต้องการทั้งหมด ดังภาพ



The screenshot displays a web scraper interface. At the top, a member profile is shown with a photo and a table of details:

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, Bantambang City, Battambang Porvince.

Below the profile, there are buttons for "Enable key" and "Done selecting!". The main interface is the "Web Scraper" tool, which shows the following configuration:

- Id:** member_info
- Type:** Element
- Selector:** Select (highlighted), Element preview, Data preview, dl
- Multiple
- Parent Selectors:** _root, member_info
- Buttons: Save selector, Cancel

ตอนนี้ เราต้องการให้โปรแกรม Web Scraper รู้ว่ามีบริเวณองค์ประกอบที่ต้องการดึงข้อมูล ซึ่งเป็นข้อมูลสมาชิกในหน้านี้อยู่หลายกล่อง เราจำเป็นต้องเลือก “Multiple” และเลื่อนให้เมาส์ครอบคลุมกับกล่องอื่น ๆ ของสมาชิกทุกคนที่แสดงในหน้านั้น



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.

div.details:nth-of-type(2) dl Enable key Done selecting!

events

Elements Console Sources Network Performance Memory Application Security Audits Web Scraper

Sitemaps Sitemap cambodia_chamber_of_commerce Create new sitemap

Id member_info

Type Element

Selector Select Element preview Data preview dl

Multiple

Parent Selectors _root member_info

Save selector Cancel

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

ตอนนี้ เราจะเห็นข้อมูลสมาชิกของทุกคนมีแถบสีครอบคลุมดังภาพด้านล่าง ให้เราคลิกปุ่ม “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 Oskospeha Village, Svay Por Commnue, Battambang City, Battambang Province

หน้าจอข้อมูลของข้อมูลที่ต้องการดึงควรจะแสดงผลเช่นเดียวกับหน้าจอด้านล่างนี้ จากนั้นให้กดปุ่ม “Save selector”.

The screenshot shows the 'Web Scraper' configuration panel. The 'Id' field is set to 'member_info', the 'Type' is 'Element', and the 'Selector' is 'dl'. The 'Multiple' checkbox is checked. The 'Parent Selectors' list contains '_root' and 'member_info'. At the bottom, there are 'Save selector' and 'Cancel' buttons.

หลังจากนั้น โปรแกรมจะแสดงหน้าจอตามด้านล่าง ซึ่งแสดงข้อมูลของสมาชิก “member_info” selector ที่ได้สร้างไว้

The screenshot shows the 'Web Scraper' interface with a table of selectors. The table has columns for ID, Selector, type, Multiple, Parent selectors, and Actions. One selector is listed: 'member_info' with selector 'dl', type 'SelectorElement', 'Multiple' set to 'yes', and parent selectors '_root'. The Actions column contains buttons for 'Element preview', 'Data preview', 'Edit', and 'Delete'. Below the table is an 'Add new selector' button.

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

กดที่ชื่อ “member_info” เพื่อเข้าไปกำหนดรายละเอียดย่อยของสมาชิกที่ต้องการให้โปรแกรมดึงข้อมูล

The screenshot shows the 'Web Scraper' interface with the path '_root / member_info' selected. Below the path is a table with columns for ID, Selector, type, Multiple, Parent selectors, and Actions. The table is currently empty. Below the table is an 'Add new selector' button.

ID	Selector	type	Multiple	Parent selectors	Actions
----	----------	------	----------	------------------	---------

คุณควรจะเห็นเส้นทางของข้อมูล Path “_root / member_info” บนมุมมองตัวเลือกที่เราเลือก ให้กด “Add new selector” เพื่อเพิ่มข้อมูลที่ต้องการเลือกภายใต้ข้อมูลสมาชิก

เราจำเป็นต้องเลือกข้อมูลในรายละเอียดของแต่ละกล่องภายใต้ข้อมูลสมาชิก (“children” of the member_info selector) เมื่อคุณกดปุ่ม “Select” แถบกล่องสีเหลืองจะปรากฏเหนือบริเวณที่คุณได้เลือกไว้

เราจะสร้างตัวเลือก เริ่มจากชื่อสมาชิก (Name) โดยเลื่อนเมาส์ไปเลือกชื่อสมาชิก

พิมพ์ในช่อง Id ว่า “name” และเลือกประเภทในช่อง Type เป็น “Text”

กดปุ่ม “Done selecting!” และเลือกกด “Save selector” เพื่อบันทึกข้อมูล

The screenshot displays a web scraper interface with two member profiles and a selector configuration panel. The first profile shows a photo of a woman and her details: Name: Ms. Toun Mina, Company Name: City Cafe Restaurant, Member Type: Advisory Member, Title in Chamber: (blank), Title in Company: Owner, Address: Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince. The second profile shows a silhouette and details: Name: Mr. Meas Seyha, Company Name: Borey I & II Guesthouse, Member Type: Ordinary Member, Title in Chamber: (blank), Title in Company: Manager, Address: Kamakor Village, Sangkat Svay Por, Banttambang City, Battambang Porvince. The selector configuration panel at the bottom shows: Id: name, Type: Text, Selector: Select (highlighted), Multiple: (unchecked), Regex: regex, Parent Selectors: _root, member_info (highlighted). Buttons for 'Save selector' and 'Cancel' are visible.

สร้างข้อมูลตัวเลือกในลักษณะเดียวกันเพื่อเพิ่มข้อมูลของสมาชิกตาม Id ต่อไปนี้

- name
- company_name
- member_type
- title_in_chamber
- title_in_company
- address

หลังจากที่บันทึกตัวเลือกข้อมูลทั้งหมดแล้ว คุณจะมีลิสต์ภายใต้ “member_info” parent selector ดังนี้

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

[Add new selector](#)

กด “_root” เพื่อย้อนกลับไปตัวเลือกที่สร้างไว้ในตอนแรก

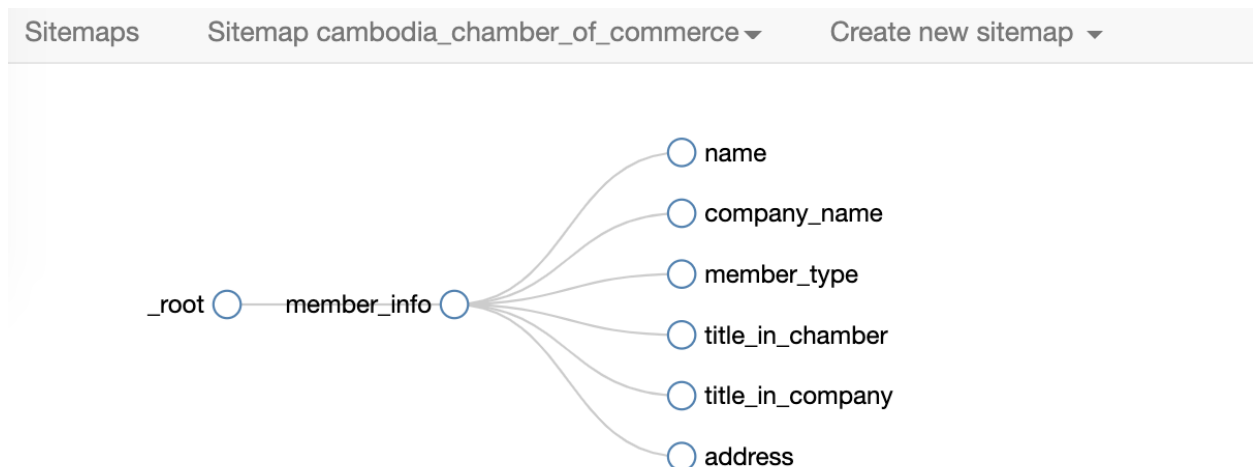
ID	Selector	type	Multiple	Parent selectors	Actions
member_info	dl	SelectorElement	yes	_root	Element preview Data preview Edit Delete

[Add new selector](#)

ตอนนี้ คุณคลิกดูข้อมูลตัวอย่าง “Data preview” ของ “member_info” ได้ และคุณ将会เห็นข้อมูลตัวอย่างที่จะทำการดึง ดังภาพด้านล่าง

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, Bantambang City, Battambang Porvince.
Mr. Meas Seyha	Borey I & II Guesthouse	Ordinary Member		Manager	Kamakor Village, Sangkat Svay Por, Bantambang City, Battambang Porvince.
Ms. Poeng Siv Bouy	City Mode and City Light	Advisory Member		Manager	Kamakor Village, Sangkat Svay Por, Bantambang City, Battambang Porvince.
Miss Eng Samphors	Vimean Sovannaphoum Resort	Advisory Member		Business Development Manager	20 Oskophea Village, Svay Por Commnue, Battambang City, Battambang Province
Mr. Tong Odom	Chap Khim Lathe Business	Advisory Member		General Manager	Kamakor Village, Sangkat Svay Por, Bantambang City, Battambang Porvince.

คุณสามารถดูไซต์แม็บที่สร้างในรูปแบบของกราฟเชื่อมต่อได้ภายใต้เมนู Sitemap ซึ่งจะแสดงความสัมพันธ์ของข้อมูลที่คุณได้เลือกไว้อย่างชัดเจน

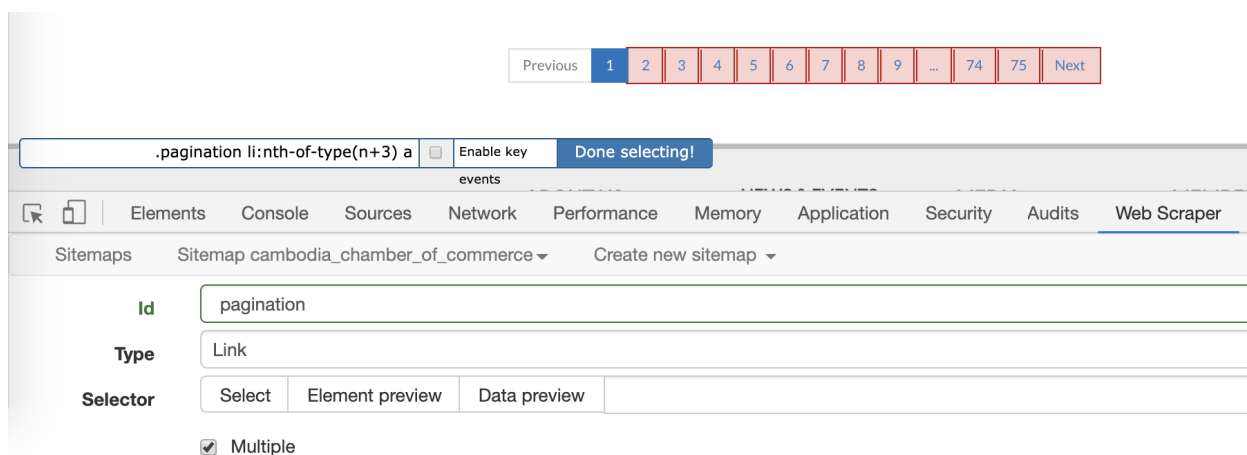


ถึงขั้นนี้ เราจะมีข้อมูลที่จะดึงจากหน้าเว็บเพียง 1 หน้าเท่านั้น แต่เว็บไซต์นี้มีสมาชิกหลายคนแสดงให้เห็น เลขหน้าที่ด้านล่างของเว็บไซต์ เราจะดึงข้อมูลทั้งหมดนี้ได้อย่างไร?

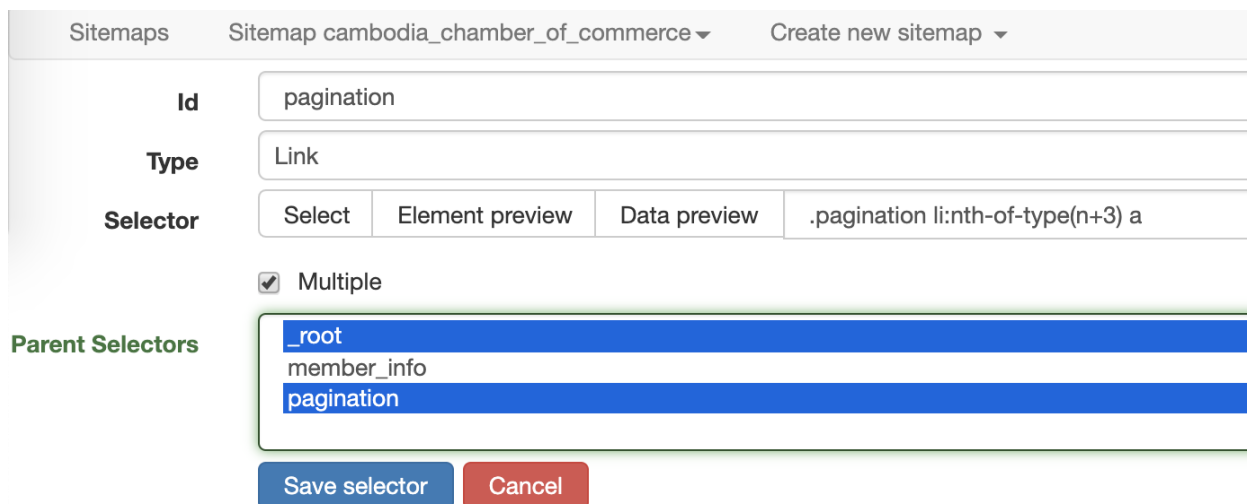
เราจะใช้เครื่องมือ Web Scrape ตัวนี้ดึงข้อมูลจากทุกหน้าได้อย่างง่ายดาย

ขั้นแรก กลับไปยังหน้าแรก “_root” page และกดปุ่ม “Add new selector”

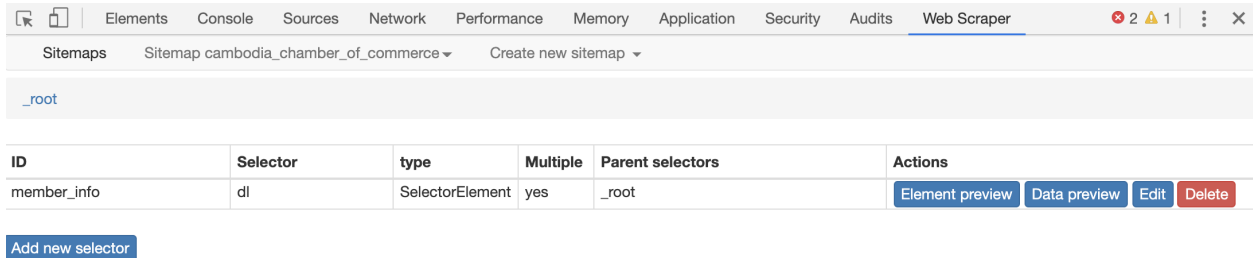
ตอนนี้ เราต้องสร้างตัวเลือกหน้า ดังนั้นกำหนดชื่อ Id ให้เป็น “pagination” และเลือกประเภทของข้อมูลเป็น “Link” คุณต้องกดเลือก “Multiple” ด้วย จากนั้น กดปุ่ม “Select” และเลื่อนเมาส์ไปเลือกเลขหน้าทั้งหมดที่อยู่ด้านล่างของเว็บไซต์ ดังภาพ



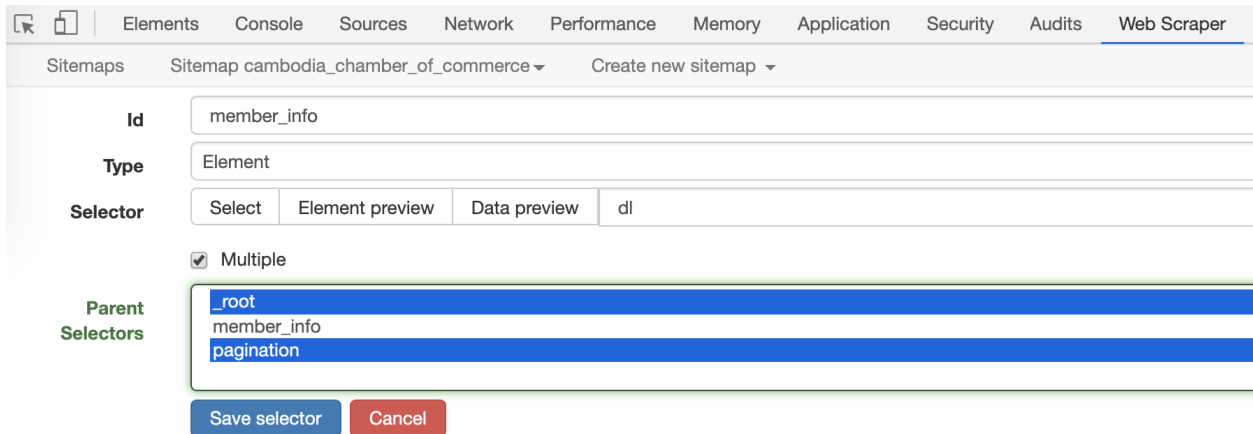
สิ่งสำคัญในขั้นตอนนี้คือ เลือกทั้ง “_root” และ “pagination” ให้เป็นตัวเลือกหลัก คุณสามารถทำได้โดยกดปุ่ม “Ctrl” (สำหรับ Windows) หรือปุ่ม “Command” (สำหรับ Mac) ในขณะที่กดปุ่ม



คุณต้องกลับไปทีตัวเลือกข้อมูลสมาชิก เพื่อเลือกให้ทั้ง “pagination” และ “_root” เป็นตัวเลือกหลักด้วยเช่นกัน โดยกลับไปทีหน้า “_root” และกดปุ่ม “Edit” ของแถว “member_info” เพื่อแก้ไขข้อมูล



ตอนนี้ คุณก็สามารถเลือก “pagination” และ “_root” เป็นตัวเลือกหลักได้ตามรูป และกดปุ่ม “Save selector”.



ตอนนี้ ถึงเวลาเริ่มสคริป ดึงข้อมูล ให้กดไปที่เมนู “Sitemap cambodia_chamber_of_commerce” และเลือก “Scrape”

Sitemaps	Sitemap cambodia_chamber_of_commerce ▼	Create new sitemap ▼
_root	Selectors	
	Selector graph	
ID	Edit metadata	ctor
member_info	Scrape	
pagination	Browse	ination li:nth-of-type(n+3) a
Add new selector	Export Sitemap	
	Export data as CSV	

หน้าจอใหม่จะปรากฏ ดังภาพด้านล่าง คุณสามารถปล่อยข้อมูลไว้ตามนั้นไม่ต้องแก้ไข และคลิกให้เริ่มทำงาน “Start scraping”.

Sitemaps	Sitemap cambodia_chamber_of_commerce ▼	Create new sitemap ▼
Request interval (ms)	<input type="text" value="2000"/>	
Page load delay (ms)	<input type="text" value="2000"/>	
	Start scraping	

หลังจากนั้น จะมีหน้าจอ pop-up แสดงหน้าข้อมูลที่คุณให้โปรแกรมดึงข้อมูลตามที่กำหนดตัวเลือกไว้ โปรแกรมจะใช้เวลาเล็กน้อยในการดึงข้อมูล

เมื่อทำงานจบแล้ว คุณจะเห็นข้อความดังด้านล่างแสดงให้เห็นพร้อมลิงก์ให้ดาวน์โหลดไฟล์ในรูปแบบ CSV

Sitemaps	Sitemap cambodia_chamber_of_commerce ▼	Create new sitemap ▼
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!		



การประเมิน

ทดสอบการแปลงข้อมูลจำนวนมากจากเว็บ เพื่อนำไปใช้งานและการวิเคราะห์ต่อได้

