

Webscraping, three main scenarios

Akitaka Matsuo

Essex IADS

Scenario 1: Data in table format

Name ◆	Scope ◆	Years active ◆	Subject matter ◆
African Court of Justice	Africa	2009–present	Interpretation of AU treaties
African Court on Human and Peoples' Rights	Africa	2006–present	Human rights
Appellate Body of the World Trade Organization	Global	1995–present	Trade disputes within the WTO
Benelux Court of Justice	Benelux	1975–present	Trade disputes within Benelux
Caribbean Court of Justice	Caribbean	2005–present	General disputes
COMESA Court of Justice	Africa	1998–present	Trade disputes within COMESA
Common Court of Justice and Arbitration of the	Africa	1998–present	Interpretation of OHADA treaties and

Scenario 2: Data in unstructured format

Find MPs

Name, postcode or location

Party

All

[Show more options](#)

Search

Total results 650 (page 1 of 33)

1

2

3











4

...


11


>


>>


 <div>Ms Diane Abbott Labour Hackney North and Stoke Newington</div>	 <div>Debbie Abrahams Labour Oldham East and Saddleworth</div>
 <div>Nigel Adams Conservative Selby and Ainsty</div>	 <div>Bim Afolami Conservative Hitchin and Harpenden</div>
 <div>Adam Afriyie Conservative Windsor</div>	 <div>Imran Ahmad Khan Conservative Wakefield</div>
 <div>Nickie Aiken Conservative Cities of London and Westminster</div>	 <div>Peter Aldous Conservative Waveney</div>
 <div>Rushanara Ali Labour Bethnal Green and Bow</div>	 <div>Tahir Ali Labour Birmingham, Hall Green</div>


Scinario 3: hidden behind web forms


 MONITOR
LEGISLATIVO


 INICIO


 PERFIL IDEAL


 NOTICIAS

 CANDIDATOS

 ASAMBLEA NACIONAL

 ABUSOS

 CONTÁCTENOS



Seleccione ▾

Partido ▾

BUSCAR

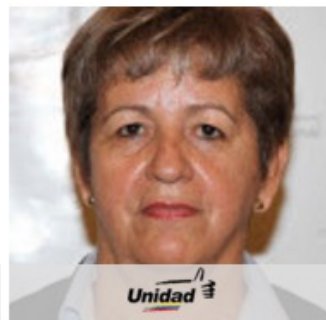
DIPUTADOS ENCONTRADOS



Julio Ygarza
Estado: Amazonas



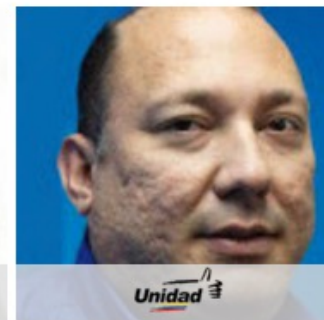
Mauligmer Baloa
Estado: Amazonas



Nirma Guarulla
Estado: Amazonas



José Brito
Estado: Anzoátegui



Chaim Bucarán
Estado: Anzoátegui



Richard Arteaga
Estado: Anzoátegui

Three main scenarios

1.Data in **table** format

- Automatic extraction with `pandas.read_html()`

2.Data in **unstructured** format

- Element identification
 - `selectorGadget`
 - **Inspect** in browser
- Identify the target with **CSS** (or **xpath**) selector
- Automatic extraction with `BeautifulSoup`

3.Data hidden behind web forms

- Automation of web browser behavior with `selenium`

The rules of the game

1. Respect the hosting site's wishes:

- Check if an API exists or if data are available for download
- Keep in mind where data comes from and give credit (and respect copyright if you want to republish the data!)
- Some websites *disallow* scrapers on `robots.txt` file

2. Limit your bandwidth use:

- Wait one or two seconds after each hit
- Scrape only what you need, and just once

3. When using APIs, read documentation

- Is there a batch download option?
- Are there any rate limits?
- Can you share the data?