

Extracción con Scraper y breve introducción a XPath

Por: Silvia Gutiérrez [CC BY 2.0]

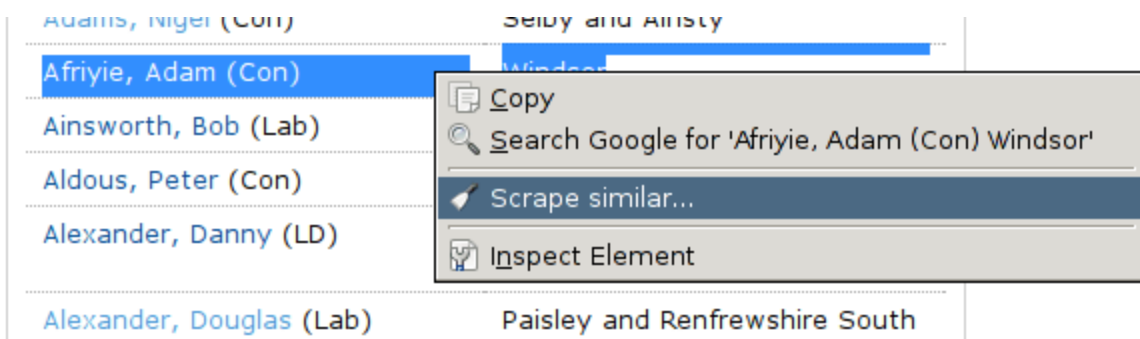
Traducido y adaptado de: <https://schoolofdata.org/handbook/recipes/scraper-extension-for-chrome/>

Ejemplo 1

1. Busca la extensión de Scraper en la tienda de Google Chrome:
<https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaaffohmbkdlecaccepngjd>
2. Añádela a tus extensiones
3. Regresa a la lista de la sesión anterior: <http://www.parliament.uk/mps-lords-and-offices/mps/>
4. Selecciona una de las entradas

Adams, Nigel (Con)	Selby and Ainsty
Afriyie, Adam (Con)	Windsor
Ainsworth, Bob (Lab)	Coventry North East
Aldous, Peter (Con)	Waveney

5. Da clic derecho y selecciona "scrape similar..."



6. Aparecerá una nueva ventana, la consola del "scraper", en la que verás el contenido extraído

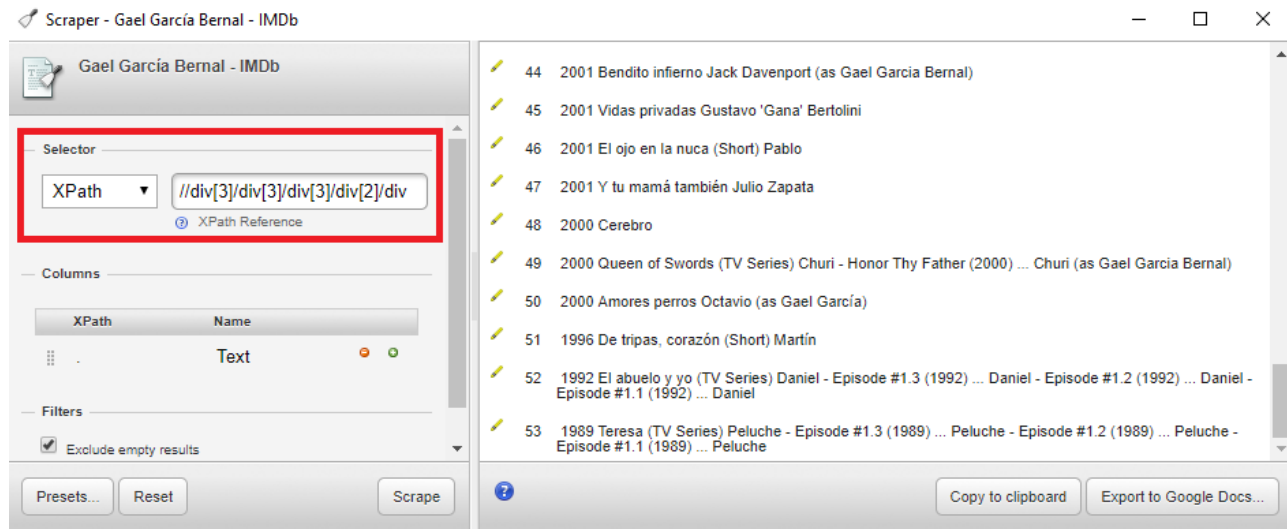
	Surname, First name	Constituency
1	A	
2	Abbott, Diane (Lab)	Hackney North and Stoke Newington
3	Abrahams, Debbie (Lab)	Oldham East and Saddleworth
4	Adams, Nigel (Con)	Selby and Ainsty
5	Afryie, Adam (Con)	Windsor
6	Ainsworth, Bob (Lab)	Coventry North East
7	Aldous, Peter (Con)	Waveney
8	Alexander, Danny (LD)	Inverness, Nairn, Badenoch and Strathspey
9	Alexander, Douglas (Lab)	Paisley and Renfrewshire South

7. Selecciona "Copy to clipboard". Abre una hoja de cálculo y pega el contenido
8. Repite los pasos 5 y 6 pero ahora selecciona "Save to Google Docs..." y otorga los permisos a Scraper para conectarse con tu cuenta.

Ejemplo 2

El ejemplo anterior es un caso ideal, pero la información no siempre está estructurada de forma tan sencilla. Veremos un ejemplo ahora con los datos de IMDB (para información como esta también se puede consultar [DBpedia](#) o [Freebase](#))

1. Digamos que quieres encontrar todas las películas en las que ha participado Gael García y vas a la lista (bastante completa) de IMDB:
<http://www.imdb.com/name/nm0305558/>
2. Intenta extraer la información de las películas en que ha sido actor con el método anterior (ya sea copiando y pegando la información o guardándola en GDocs)
3. Verás que esta vez, el procedimiento no es tan *smooth*. Esto es porque la estructura de esta tabla no es tan sencilla, y para seleccionar de mejor manera la información es que usaremos la sección de "Selector" donde aparece "Xpath" como opción de default



👉 **XPath** es un **lenguaje de consulta** para HTML y XML. XPath utiliza un lenguaje de “rutas” (path expressions) para seleccionar los nodos o grupo de nodos de un documento con lenguaje de marcado. Para aprender más:

https://www.w3schools.com/xml/xpath_intro.asp

Terminología básica

Dado el siguiente documento de HTML (que usamos en la introducción), vamos a aprender aspectos básicos de XPath

```
<!DOCTYPE html>
<html>
  <body>
    <h1>My First Heading</h1>
    <div>
      <p style="color:red;">My red paragraph.</p>
      <p style="font-size:160%;">Big text </p>
      <p style="text-align:center;">Centered text</p>
    </div>
  </body>
</html>
```

Abre <https://codebeautify.org/Xpath-Tester>, copia y pega el documento de HTML y prueba las diferentes sintaxis como se indica después de (👉)

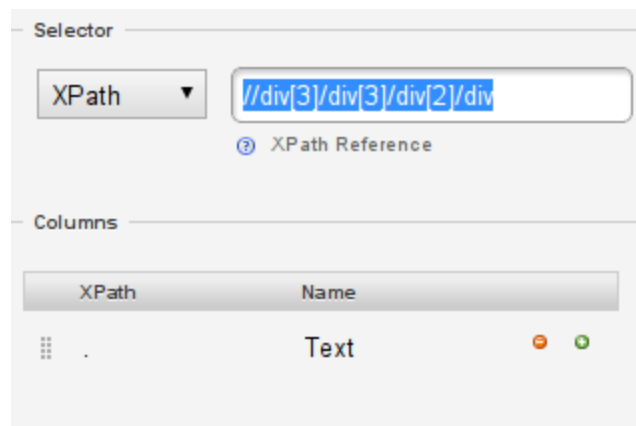
Tres tipos de nodos:

- root element node / elemento-nodo-raíz <body>
- element node / elemento-nodo <h1> <div> <p>
- attribute node / nodo-atributo style="color:blue; style="font-size:160%; style="text-align:center;">

Relaciones entre nodos y sintaxis de XPath

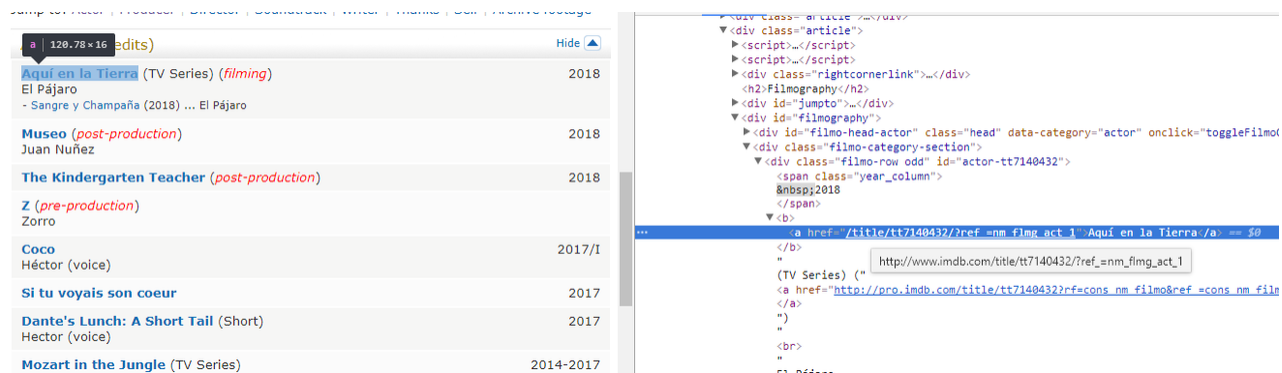
- **parent / progenitor** [*ancestors / ancestros*]: cada elemento-nodo tiene un progenitor, <div> es el progenitor de los tres <p>
👉 //p/..
- **child / hijx** [*descendants / descendientes*]: cada elemento-nodo puede tener cero o más hijxs, <h1> <div> son los hijxs de <body>
👉 //div/p (descendientes)
👉 //div/p[1] (primer hijo)
- **siblings / hermanxs**: elementos-nodo que tienen el mismo progenitor, los tres <p> del ejemplo son hermanxs

4. Regresemos al XPath de nuestro ejercicio con Gael García; el cual es:
“//div[3]/div[3]/div[2]/div”.



¿Cómo se traduce esta ruta (en términos de relaciones entre nodos y sintaxis como vimos arriba)?

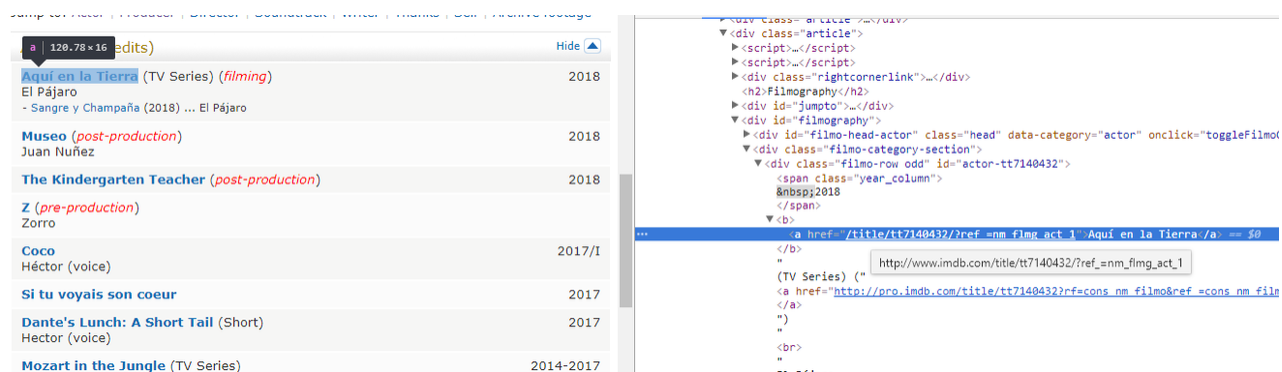
5. Ahora que sabes cómo seleccionó los datos, podremos cambiar algunas cosas de la ruta para que no aparezcan en una sola columna como aparecen hasta ahora. Primero, inspeccionemos la fila de la información con el método visto en el paso 2 del tutorial anterior:



The screenshot shows the IMDb page for the TV series 'Aquí en la Tierra'. The table lists the following entries:

Year	Role	Category
2018	El Pájaro	(TV Series) (filming)
2018	Sangre y Champaña	(TV Series) (post-production)
2018	Museo	(TV Series) (post-production)
2018	Juan Nuñez	(TV Series) (post-production)
2018	The Kindergarten Teacher	(TV Series) (post-production)
2018	Z	(TV Series) (pre-production)
2018	Zorro	(TV Series) (pre-production)
2017/I	Coco	(TV Series) (voice)
2017	Héctor	(TV Series) (voice)
2017	Si tu voyais son cœur	(TV Series) (voice)
2017	Dante's Lunch: A Short Tail	(TV Series) (voice)
2017	Hector	(TV Series) (voice)
2014-2017	Mozart in the Jungle	(TV Series)

The HTML structure on the right shows the corresponding DOM tree, highlighting the 'year_column' and 'actor' elements.

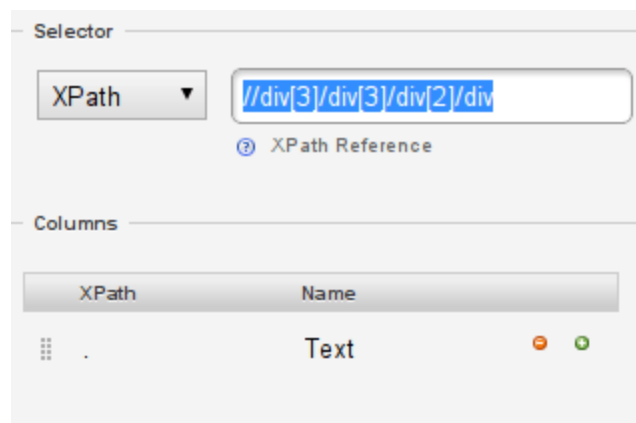


This screenshot is identical to the one above, showing the IMDb page for 'Aquí en la Tierra' and its HTML structure. The table lists the following entries:

Year	Role	Category
2018	El Pájaro	(TV Series) (filming)
2018	Sangre y Champaña	(TV Series) (post-production)
2018	Museo	(TV Series) (post-production)
2018	Juan Nuñez	(TV Series) (post-production)
2018	The Kindergarten Teacher	(TV Series) (post-production)
2018	Z	(TV Series) (pre-production)
2018	Zorro	(TV Series) (pre-production)
2017/I	Coco	(TV Series) (voice)
2017	Héctor	(TV Series) (voice)
2017	Si tu voyais son cœur	(TV Series) (voice)
2017	Dante's Lunch: A Short Tail	(TV Series) (voice)
2017	Hector	(TV Series) (voice)
2014-2017	Mozart in the Jungle	(TV Series)

The HTML structure on the right shows the corresponding DOM tree, highlighting the 'year_column' and 'actor' elements.

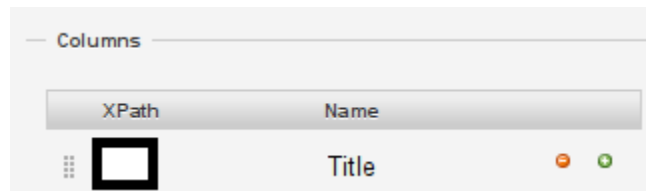
6. ¿Entre qué etiquetas está el título?
7. Para seleccionar esa etiqueta tenemos que usar otra expresión que no conocíamos: ".", indica la dirección actual; la dirección actual es la ruta que tenemos aquí:



The screenshot shows the XPath Selector tool. The 'XPath' field contains the path `//div[3]/div[3]/div[2]/div`. The 'Columns' section shows a table with the following columns:

XPath	Name
.	Text

En la ventana de columna, tienes que decirle que de la ruta actual, quieres seleccionar todo el contenido de las etiquetas b's ¿cómo lo harías?



8. En la sección de “Columns” cambia el nombre de la primera columna a “title” como en la imagen de arriba.
9. Ahora, ¿entre qué etiquetas está la información del año?
10. Repite los pasos 7 y 8 ahora para integrar la información del año de la película



11. Da clic en el botón de “scrape” y observa tus datos ahora:

