

doc-pract-1

April 7, 2022

1 Tipología y ciclo de vida de los datos: PRÁCTICA I

Componentes de la práctica:

Jonás Medina Brito (jmedinabrit@uoc.edu)

Miguel Rafael Esteban Martín (mestebanmart@uoc.edu)

1.1 Índice de contenidos

- Contexto
- Título
- Descripción del dataset
- Representación gráfica
- Contenido
- Agradecimientos
- Inspiración
- Licencia
- Código
- Dataset

1.2 Contexto

1.2.1 Recogida de datos

Aunque [Eurostat](#) tiene una API pública (1) que permite obtener los diferentes dataset, en esta práctica hemos decido hacer web scraping de los datos que se ofrecen en su “*Data browser*” accesible desde un navegador.

- (1) Eurostat web services: <https://ec.europa.eu/eurostat/en/web/main/data/web-services> (Última vista abril del 2020)

Para explicar el estudio que se ha hecho para la obtención de datos se tomará como ejemplo el dataset de “*Componentes del precio de la electricidad para consumidores domésticos*” ([Electricity prices components for household consumers - annual data \(from 2007 onwards\)](#))

The screenshot shows the Eurostat Data Browser interface. The title is "Electricity prices components for household consumers - annual data (from 2007 onwards)". The data code is "NRG_PC_204_C". The interface includes a selection panel with "Geographical entity (reporting)" set to "EU-28 (EU-27 + EU-28)" and "Time" set to "Year [Yr]". The table displays electricity price components for various countries, including EU-28, EU-27, and individual member states like Austria, Belgium, Bulgaria, Czechia, Denmark, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, and the UK.

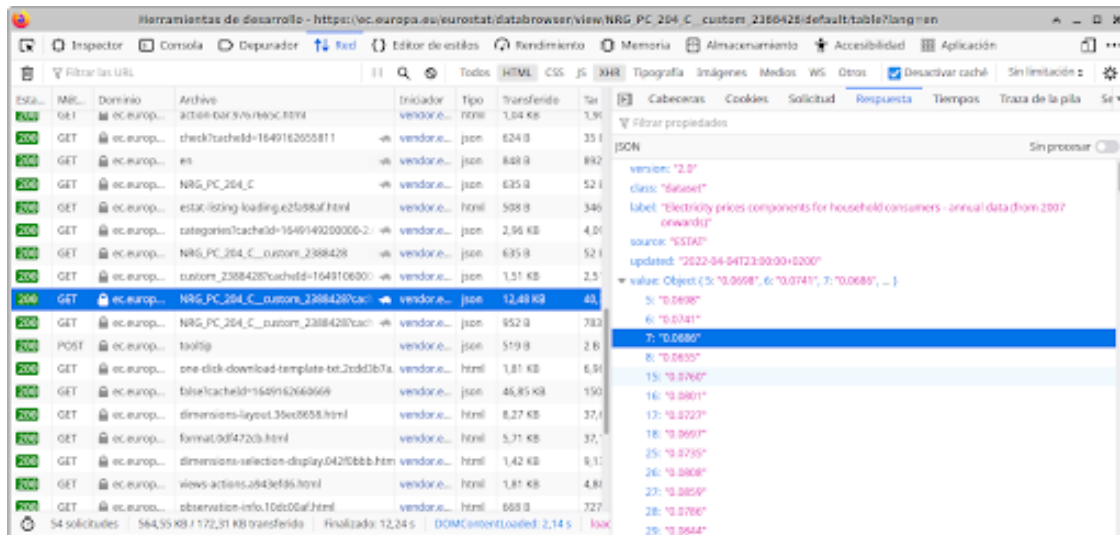
Geographical entity (reporting)	2007	2008	2009	2010	2011
EU-28 (EU-27 + EU-28)	0.0706	0.0697	0.0693	0.0693	0.0694
Austria	0.0706	0.0697	0.0693	0.0693	0.0694
Belgium	0.0691	0.0691	0.0694	0.0693	0.0694
Bulgaria	0.0691	0.0691	0.0694	0.0693	0.0694
Czechia	0.0691	0.0691	0.0694	0.0693	0.0694
Denmark	0.0691	0.0691	0.0694	0.0693	0.0694
Germany (incl. HH service territory of the FRG)	0.0691	0.0691	0.0694	0.0693	0.0694
Greece	0.0691	0.0691	0.0694	0.0693	0.0694
Hungary	0.0691	0.0691	0.0694	0.0693	0.0694
Ireland	0.0691	0.0691	0.0694	0.0693	0.0694
Italy	0.0691	0.0691	0.0694	0.0693	0.0694
Latvia	0.0691	0.0691	0.0694	0.0693	0.0694
Lithuania	0.0691	0.0691	0.0694	0.0693	0.0694
Luxembourg	0.0691	0.0691	0.0694	0.0693	0.0694
Malta	0.0691	0.0691	0.0694	0.0693	0.0694
Netherlands	0.0691	0.0691	0.0694	0.0693	0.0694
Poland	0.0691	0.0691	0.0694	0.0693	0.0694
Portugal	0.0691	0.0691	0.0694	0.0693	0.0694
Romania	0.0691	0.0691	0.0694	0.0693	0.0694
Slovakia	0.0691	0.0691	0.0694	0.0693	0.0694
Slovenia	0.0691	0.0691	0.0694	0.0693	0.0694
Spain	0.0691	0.0691	0.0694	0.0693	0.0694
Sweden	0.0691	0.0691	0.0694	0.0693	0.0694
Switzerland	0.0691	0.0691	0.0694	0.0693	0.0694
UK	0.0691	0.0691	0.0694	0.0693	0.0694

Captura de pantalla de la interfaz “Data Browser” del informe de los componentes del precio de la electricidad para consumidores domésticos

El sitio web de Eurostat utiliza HTML dinámico para generación de la vista, esto quiere decir que el navegador procesa la información y la presenta generando código HTML, y no sólo se limita a representar el código HTML que responde el servidor. Este componente de generación de código por parte del navegador hace complicado el web scraping tradicional.

Utilizando ingeniería inversa y las herramientas de desarrollo del navegador (Firefox), se puede ver

cual es la llamada que devuelve los datos que se representan en la tabla dinámica. Un ejemplo es la petición que es utilizada en la construcción de la tabla con los datos que se muestra en la página web.



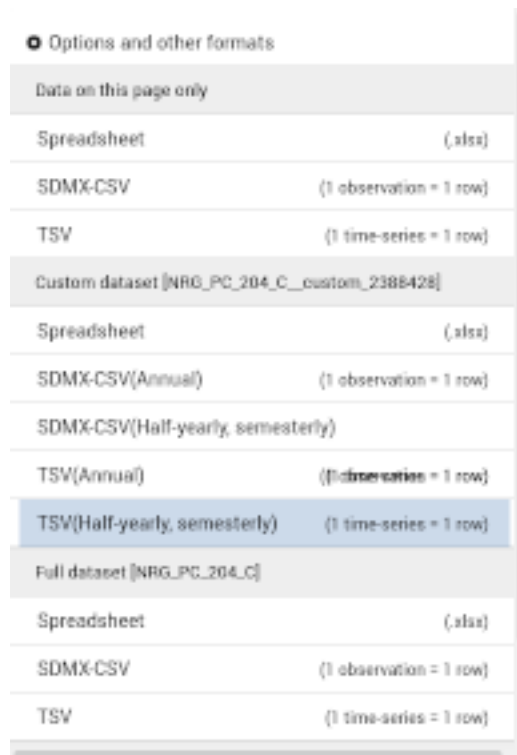
Captura de pantalla de la consola del navegador con la petición de datos de dataset y respuesta JSON

Si se exporta la petición esta es la llamada que se realiza (comando CURL):

```
$ curl 'https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/LIVE/false/json/en/NRG_PC_204_C_custom_2388428?cacheId=1649106000000-2.6.3%2520-%25202022-03-17%252005%253A43' ...
```

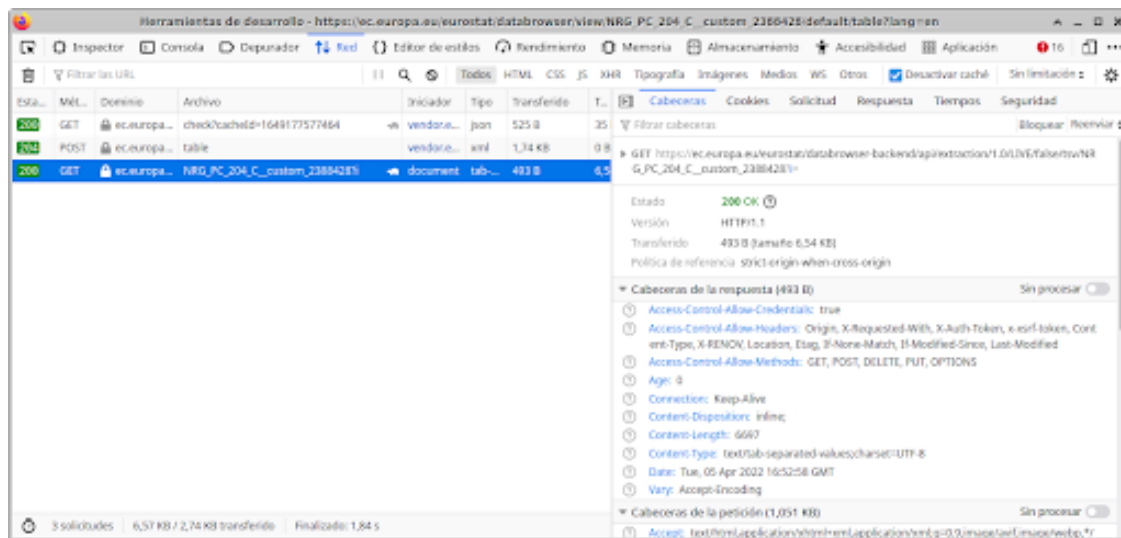
La respuesta de esta llamada es en formato json con una estructura específica, difícil de procesar y aprovechar para el propósito de acceder a los datos.

Sin embargo la página tiene un formulario de descarga que permite descargar los dataset, en formato TSV mediante el navegador.



Captura de pantalla del enlace de descarga del dataset

Si se hace click en ese enlace la petición que hace el navegador es la siguiente:



Captura de pantalla de la interfaz “Data Browser” de la descarga del informe de los componentes del precio de la electricidad para consumidores domésticos en formato TSV

- 1.3 Título
- 1.4 Descripción del dataset
- 1.5 Representación gráfica
- 1.6 Contenido
- 1.7 Agradecimientos
- 1.8 Inspiración
- 1.9 Licencia
- 1.10 Código
- 1.11 Dataset