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 Indice de contenidos

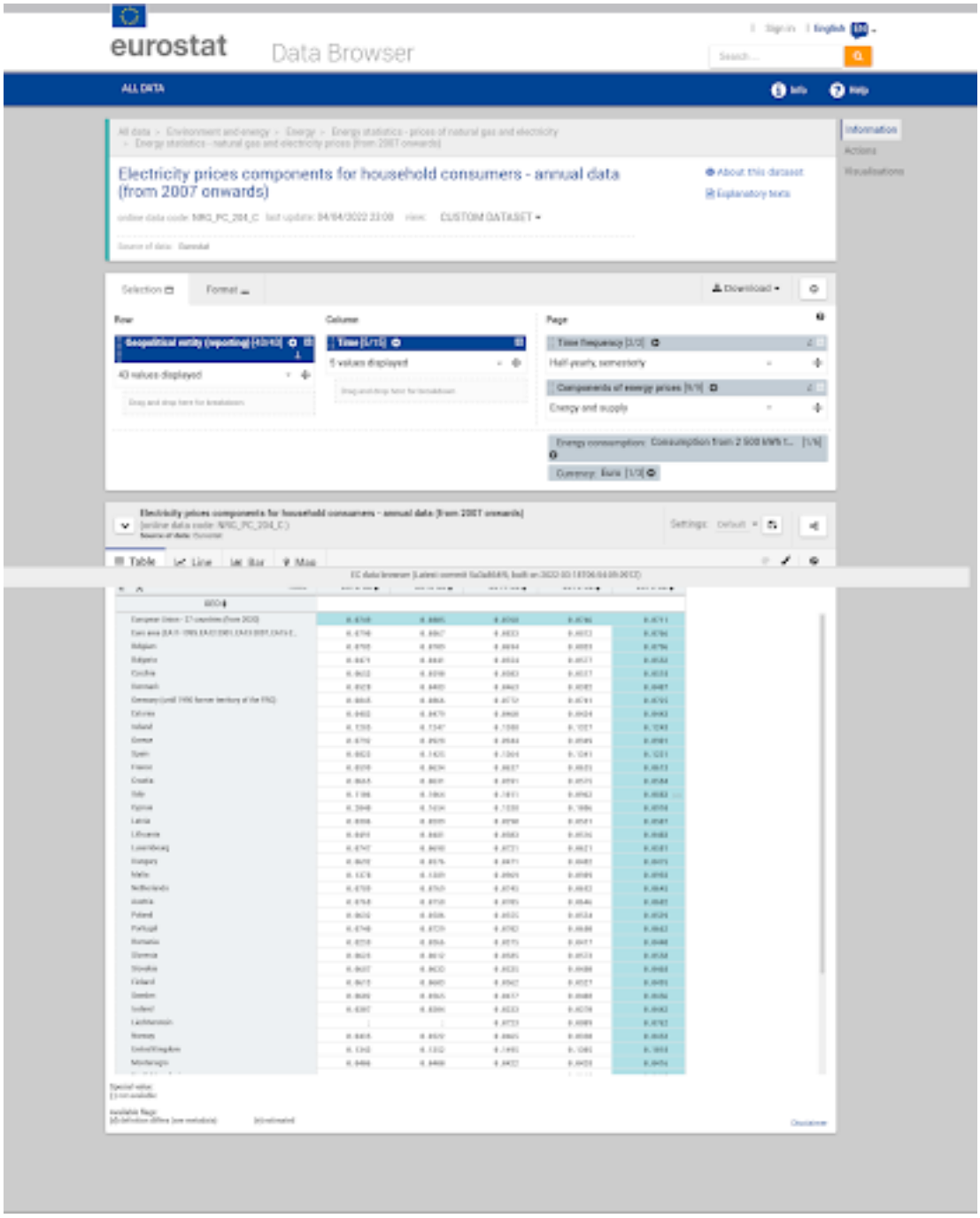
1.2 Contexto

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)

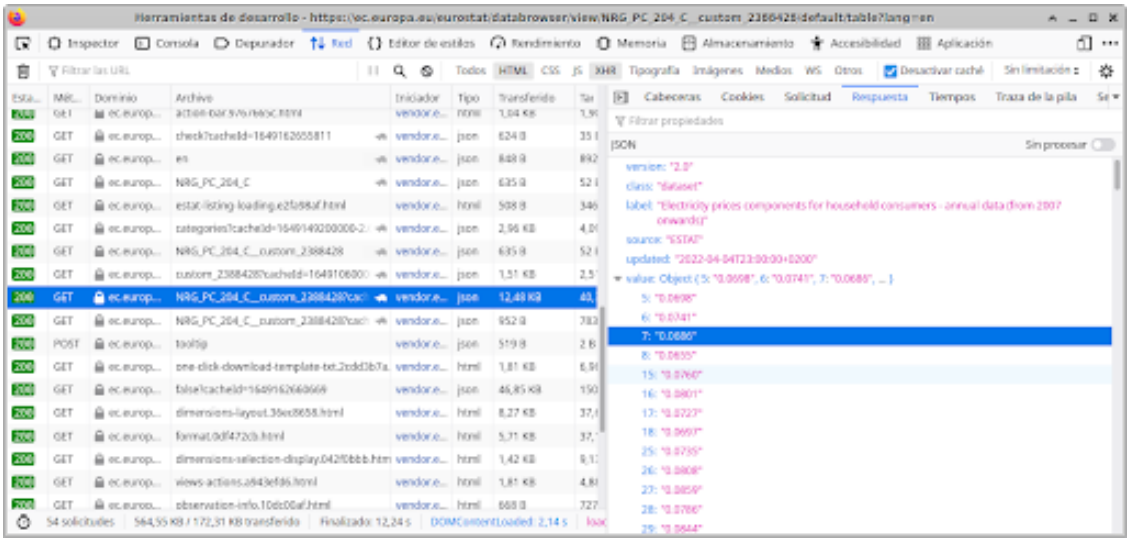


Captura de pantalla de la interfaz "Data Browser" del informe de los componentes del precio de la electricidad para consumidores domésticos

El sitio web de Eurostat utiliza HTML dinámico para generación de la vista, esto quiere decir qie 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.



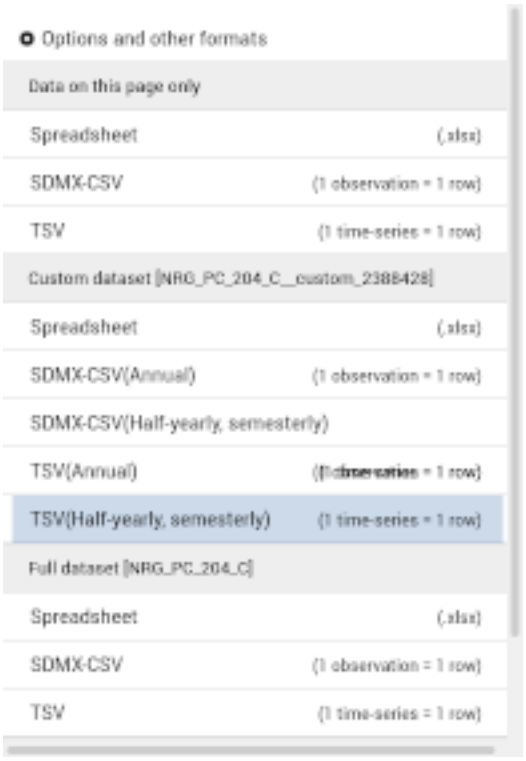
Captura de pantalla de la consola del navegador con la petición de datos de dataset y respuesta JSON

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' …
```

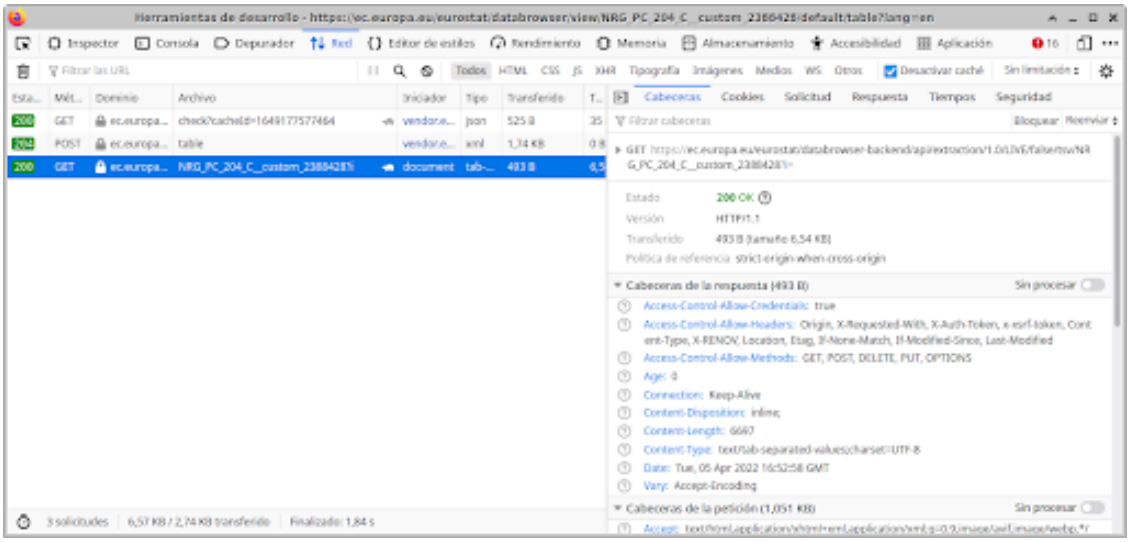
La respuesta de esta llamada es en formato JSON con una estructura específica, difícil de procesar y aprovechar para neustro propósito de acceder a los datos.

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:



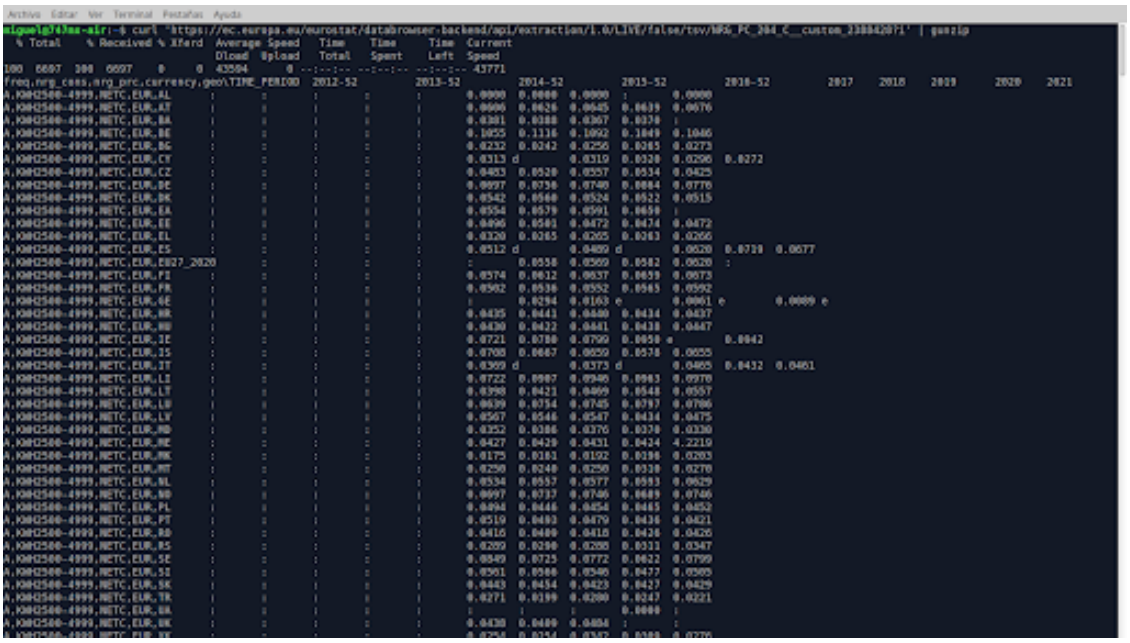
Captura de pantalla de la interfaz "Data Browser" de la descarga del informe de los componentes del precio de la electricidad para consumidores domésticos en formato TSV

Qué si se exporta la petición en formato CURL es:

```
$ curl 'https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/LIVE/
```

`false/tsv/NRG_PC_204_C__custom_2388428?i' …`

Y que si se ejecuta desde una consola se obtiene los datos en el formato TSV



Captura de pantalla de la respuesta utilizando el comando CURL

1.2.2 Llamada HTTP para la obtención del los datos de los dataset

Analizando la petición anterior la llamada que se hará para la obtención de los diferentes dataset que componen el dataset principal será una petición GET a la URL

`https://ec.europa.eu/eurostat/databrowser-backend/api/extraction/1.0/LIVE/false/tsv/<key_dataset>?i`

Donde `<key_dataset>` indica el identificador del dataset del sistema de eurostat, identificado como "*online data code*" en cada uno de los conjuntos de datos que son accesibles desde el "*Data Browser*" de Eurostat. Esa llamada es fácilmente procesable desde las utilidades que ofrece la librería PANDAs para la carga de dataset.