# Tipologia i cicle de vida de les dades

# Pràctica 1

Web Scraping de les escales de vaixells al Port de Tarragona durant l'any 2019

Arnau Giribet March
Oscar Buisan Viniegra

#### 1. Títol

Web Scraping de les escales de vaixells al Port de Tarragona durant l'any 2019.

#### 2. Context

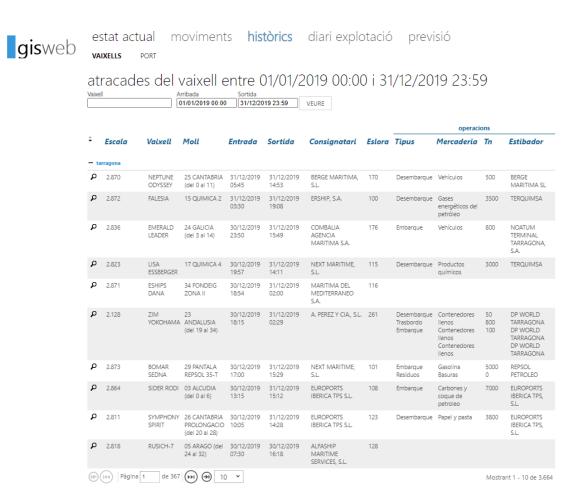
Les dades recollides pel scraper eren necessàries per l'estudi previ a la creació de l'impost d'emissions dels grans vaixells que actualment s'està desenvolupant des de la conselleria d'Acció Climàtica i la conselleria d'Economia de la Generalitat de Catalunya.

### 3. Descripció del dataset

El dataset conté un registre per a cada escala de cada vaixell realitzada al Port de Tarragona.

Cada escala està caracteritzada pel vaixell en qüestió i les seves propietats, el temps que va durar, les empreses implicades, el moll on es va realitzar i el tipus d'operació o operacions que es van dur a terme.

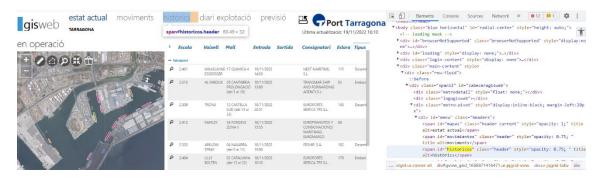
#### 4. Representació Gràfica



Aquesta és una captura de pantalla de l'aplicació amb la qual s'interactua.

Com es pot observar, les dades que hem d'obtenir estan distribuïdes en 367 pàgines i taules de 10 files que el scraper haurà de navegar.

Com que podem augmentar fins a 30 el número de registres per pàgina, el nostre programa només haurà d'iterar 123 pàgines.



Per identificar cada lloc on el scraper ha de clicar o obtenir-ne dades s'utilitza l'opció Inspecciona del cercador.

I aquí es pot veure la instrucció que li donem al scraper per a que detecti, en aquest cas, la pestanya històrics. Més endavant analitzarem el codi amb detall.

#### 5. Contingut

S'han recollit totes les escales realitzades durant l'any 2019 (al Port de Tarragona).

De cada escala se n'enregistra el **número d'escala, nom del vaixell**, a quin **moll** s'ha realitzat, **data i hora d'entrada i de sortida**, el **consignatari**, **l'eslora** del vaixell, el **tipus d'operació**, les **mercaderies** transportades, el **tonatge i l'estibador**. Cal dir que aquestes últimes quatre dades – operació, mercaderia, tonatge i estibador - poden ser múltiples dins de cada escala (diverses operacions per una sola escala), de manera que en un posterior tractament de les dades caldrà tenir-ho en compte.

#### 6. Propietari

Les dades són propietat del Port de Tarragona.

Per poder fer-les accessibles van crear l'eina GISWEB, que és la plataforma amb la qual nosaltres interactuem amb el scraper. La seva intenció és proporcionar-les obertament a la ciutadania però no acaben de facilitar-ne la descàrrega directa i massiva.

Es pot trobar més informació en el següent enllaç:

https://www.porttarragona.cat/ca/situacio-de-vaixells-link

#### · Accessibilitat a través de scraper

L'eina GISWEB (amb enllaç <a href="https://tarragona.posidoniaport.com/">https://tarragona.posidoniaport.com/</a>) no disposa de fitxer robots.txt. Es pot constatar seguint l'adreça <a href="https://tarragona.posidoniaport.com/robots.txt">https://tarragona.posidoniaport.com/robots.txt</a>, que no condueix enlloc. Per tant, assumim que navegar-la amb un scraper no serà problemàtic.

La pàgina web del Port de Tarragona sí que té fitxer robots.txt i les úniques adreces que inhabilita són les següents, sense fer esment de GISWEB:

```
User-agent: *
Allow: /*.js*
Allow: /*.css*
Allow: /*.png*
Allow: /*.jpg*
Allow: /*.gif*
Disallow: /administrator/
Disallow: /bin/
Disallow: /cache/
Disallow: /cli/
Disallow: /includes/
Disallow: /installation/
Disallow: /language/
Disallow: /layouts/
Disallow: /libraries/
Disallow: /logs/
Disallow: /tmp/
```

## 7. Inspiració

Aquestes dades són interessants ja que permeten observar amb tota transparència els moviments que es duen a terme al Port de Tarragona.

A més a més, estan resultant pràctiques en l'elaboració de l'impost a les emissions dels grans vaixells que s'està desenvolupament conjuntament a les conselleries d'Acció Climàtica i la d'Economia. A partir del temps al port, tonatge i tipologia de vaixell es pot fer un estudi de les emissions contaminants habituals al llarg de l'any i a partir d'aquí gravar tan justament com es pugui els grans vaixells.

#### 8. Llicència

La Ilicència escollida és Open Database License (ODbL-1.0).\*

El motiu és que, en tractar-se de dades obertes que facilita el Port de Tarragona, sembla coherent mantenir el dataset resultant també obert.

\*Tot i creure que aquesta és la llicència més adient per la base de dades, en la publicació a Zenodo aquesta no era una de les opcions possibles, i s'ha acabat aplicant-hi la llicència Creative Commons Attribution 4.0 International.

#### 9. Codi

El codi íntegre del scraper es pot trobar a:

https://github.com/arnaugiribet/scrapings/blob/main/source/Escales%20Vaixells.py

Una explicació detallada del funcionament del codi es pot trobar a:

https://github.com/arnaugiribet/scrapings/blob/main/Explicaci%C3%B3%20codi.pdf

#### 10. Dataset

El dataset s'ha publicat a Zenodo amb DOI 10.5281/zenodo.7337235.

## https://doi.org/10.5281/zenodo.7337235

## 11. Vídeo

El vídeo de presentació de la pràctica es pot trobar a:

<u>Enllaç</u>