# PRA1: WEB SCRAPING

## 1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Des del mes de març d'enguany estem vivint un context de pandèmia global causat pel coronavirus *Covid19-Sars2* que està tenint un fort impacte tant des del punt de vista econòmic com en les pautes de conducta social. Degut a la ràpida transmissió del virus i la incidència en la població, les administracions públiques reporten diàriament les xifres de què disposen pel que fa a casos i morts en compliment a criteris de transparència. Si bé les administració, o els instituts sanitaris en què deleguen, proporcionen les dades de què disposen en els repositoris que controla, obtenir una representació a escala global que recapituli la incidència de la Covid-Sars2 de forma actualitzada a escala global no és una tasca immediata.

En aquesta situació, la pàgina web www.worldometers.info s'ha convertit en un referent per a aquest tipus de consulta. *Worldometers* recull dades globals en temps real de múltiples paràmetres demogràfics, despesa pública, de consum energètic, ús de recursos naturals o relacionades amb la salut. Entre ells, podem accedir a una taula en què podem informar-nos sobre la incidència de Covid19 a nivell global, per continent o per país.

## 2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

El títol que hem escollit per al dataset és: *Comparativa de l'evolució de la incidència de la Covid19-Sars2 a l'Estat Espanyol respecte del conjunt d'Europa entre el 20 d'octubre i el 4 de novembre de 2020*.

## 3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

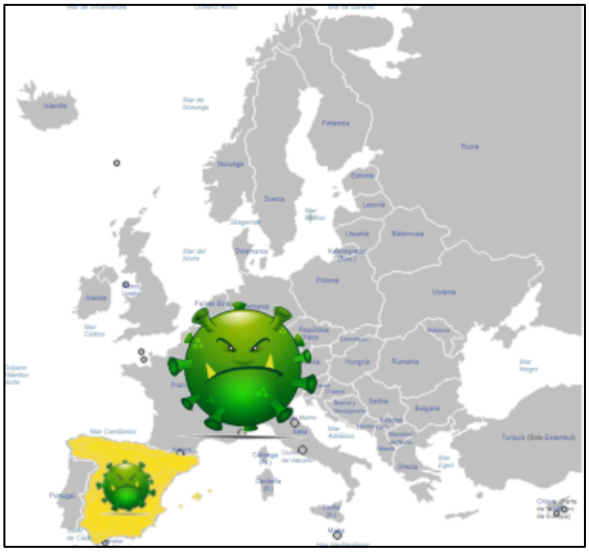
Amb la finalitat de poder fer una comparativa temporal de la incidència de la Covid19-Sars2 a l'Estat Espanyol respecte del conjunt dels països d'Europa, hem recollit les dades diàries del 20 d'Octubre reportades a ambdós territoris pel que fa a:

- *Total cases* (Número total de casos): sumatori de casos de positius acumulats.
- *New Cases* (Casos nous): número diari de casos positius reportats.

- *Total Deaths* (Número total de morts): sumatori de defuncions per Covid19 acumulades.
- *New Deaths* (Morts noves): número diari de defuncions per Covid19 reportats.
- *Total Recovered* (Número total de recuperats): sumatori de casos de pacients recuperats.
- *New Recovered* (Recuperats nous): número diari de casos de pacients recuperats.
- *Active Cases* (Casos actius): sumatori del número total de casos actius de Covid19.
- *Serious, Critical* (Casos crítics): sumatori del total de malalts de Covid19 en UCIs.

## 4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment

El data set que hem produït està representat per la següent pictograma, que fa referència a la presència del virus al continent europeu i a l'Estat Espanyol:



Per a transmetre aquesta idea, hem recorregut a un mapa del continent europeu en què l'Estat Espanyol apareix destacat de la resta del continent i on dues animacions, fàcilment identificables amb el virus de la Covid19, se situen sobre els territoris d'on hem recollit les dades.

## 5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

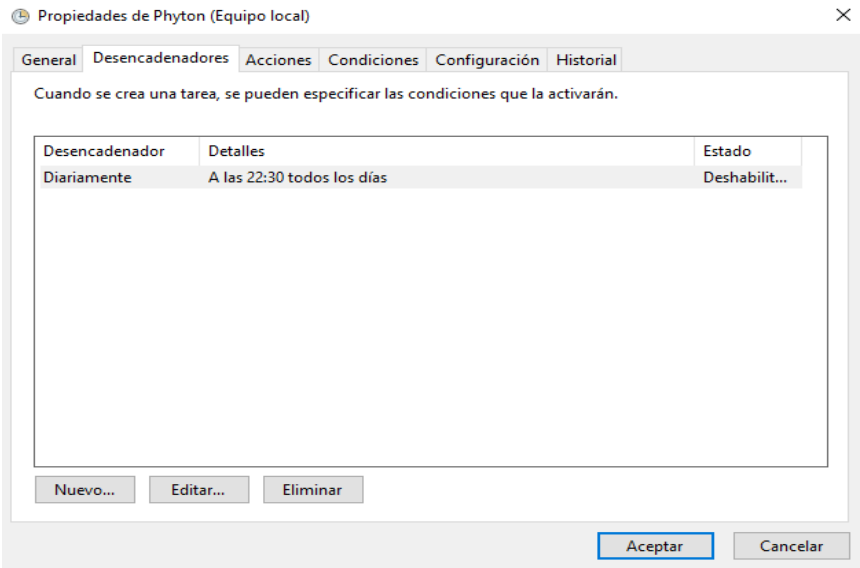
El data set que hem proporcionat consta de 16 fileres per 16 columnes agrupades en dos factors que corresponen a un conjunt de 8 variables (*8 columnes*) per territori (*2 factors: Europa*

*i Estat Espanyol)* recollides diàriament entre el 20 d'octubre i el 4 de novembre (16 dies, per tant *16 fileres*) pel que fa a:

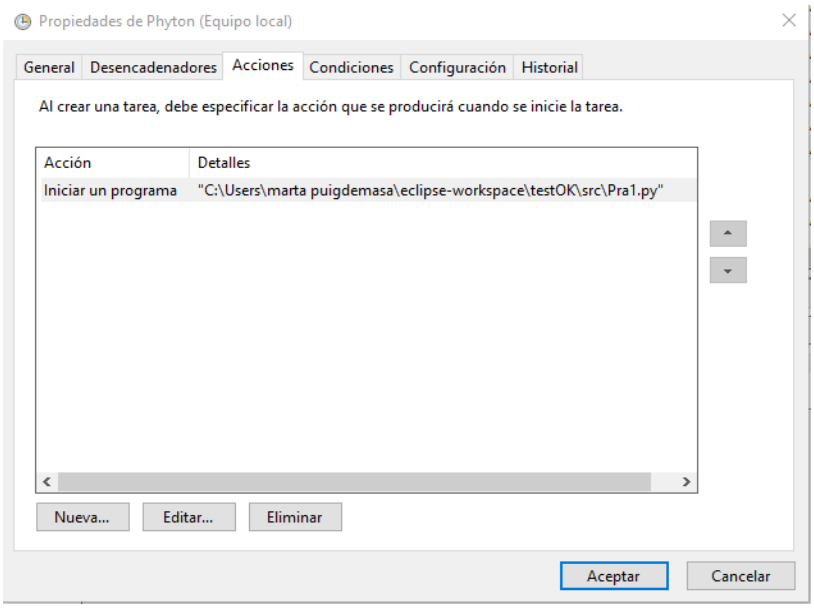
- *Total cases* (Número total de casos): sumatori de casos de positius acumulats.
- *New Cases* (Casos nous): número diari de casos positius reportats.
- *Total Deaths* (Número total de morts): sumatori de defuncions per Covid19 acumulades.
- *New Deaths* (Morts noves): número diari de defuncions per Covid19 reportats.
- *Total Recovered* (Número total de recuperats): sumatori de casos de pacients recuperats.
- *New Recovered* (Recuperats nous): número diari de casos de pacients recuperats.
- *Active Cases* (Casos actius): sumatori del número total de casos actius de Covid19.
- *Serious, Critical* (Casos crítics): sumatori del total de malalts de Covid19 en UCIs.

Aquest recull s'ha fet mijtançant l'execució de l'script de *Python* ***Pra1_Covid_MP_AG_Web_Scraping.py*** de forma diària mitjançant una tasca programada de Windows que s'ha executat cada dia a les 22:30. Per tal d'il·lustrar el procediment, adjuntem una captura de pantalla de la tasca programada:

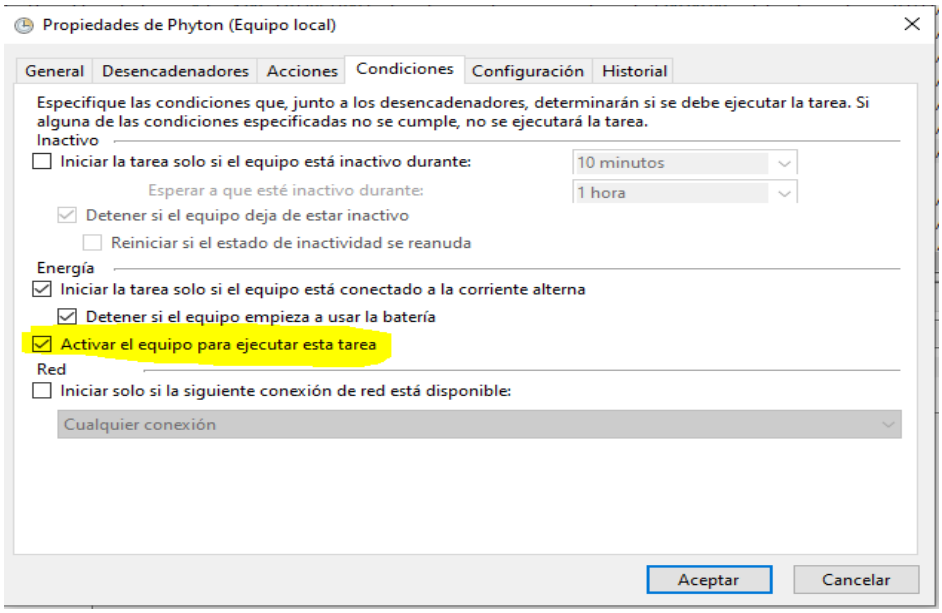
1. Definim la periodicitat en que s'ha d'executar: s'executa diàriament a les 22:30:




2. Seleccionem la ubicació del .py que guarda les dades:




3. Marquem la opció de que activi l'equip en cas de que estigui en supensió




4. Comprovem que es va executant correctament i el resultat al .fitxer csv es l'esperat

## 6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Volem agrair el web *Worldometers* per la seva dedicació desinteressada en la tasca de recollir i integrar dades i fer-les accessibles al públic en general.

## 7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

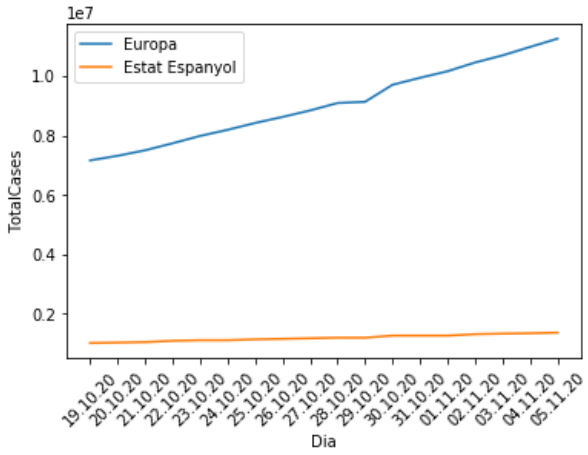
El continent europeu ha sigut un dels més afectats per la Covid-Sars2, sobretot degut a la greu afecció que han sofert els països del sud, especialment l'Estat Espanyol i Itàlia, durant el que es coneix com la *Primera Onada*. Malgrat el significatiu alentiment en l'increment de casos que es va aconseguir mitjançant estrictes mesures de confinament, la relaxació de mesures, el retorn a les rutines socials i laborals i la participació en activitats en espais interns ha comportat un important repunt en el número de casos de del mes de setembre. Conseqüentment, diferents organismes i mitjans de comunicacions ja apunten a una *Segona Onada*. De fet, l'Estat Espanyol va superar el 21 d'octubre el milió de contagiats (https://www.elpuntavui.cat/societat/article/14-salut/1867547-espanya-supera-el-milio-de-contagis-del-coronavirus.html).

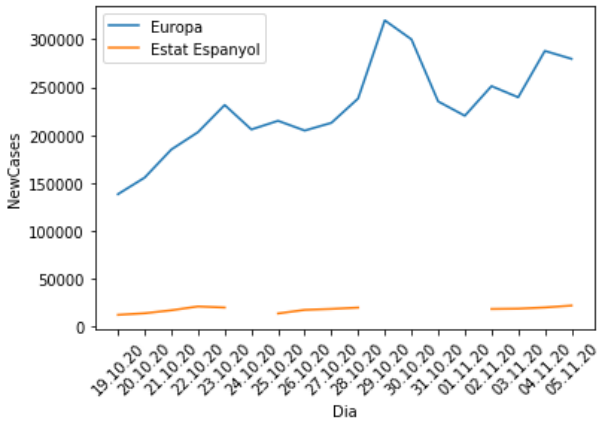
Des del punt de vista sanitari i informatiu la situació a l'Estat Espanyol és molt preocupant i ha obert una sèrie de preguntes, com ara: L'evolució a l'Estat Espanyol segueix la mateixa tendència que a la resta del continent o és un cas aïllat? Quants dels casos que es reporten diàriament al continent provenen de l'Estat Espanyol?

Fer una comparativa directa entre països o un *ranking* és bastant senzill ja que *Worldometers* ofereix les dades normalitzades per milió d'habitants, però inferir tendències no és tan intuïtiu. Tanmateix, atès que el web ofereix l'increment diari i total de casos i defuncions, podem descarregar-les i fer-hi una visualització. Per tant, per a poder resoldre les preguntes adès esmentades durem a terme un protocol de *web scraping* amb finalitats de recerca que monitoritzi diàriament les dades totals recollides per *Worldometers* a Europa i a nivell particular a l'Estat Espanyol entre el dia 20 d'octubre i 4 de novembre.

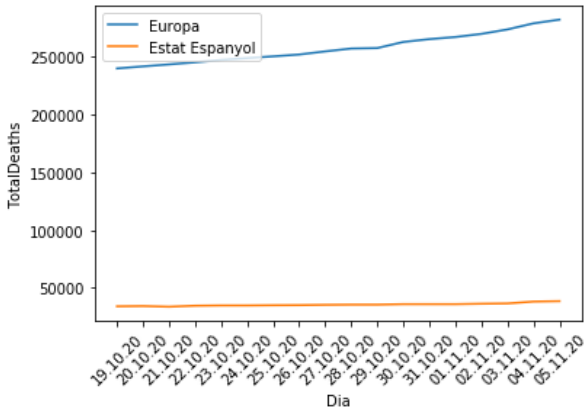
Si tenim en compte que dels 48 estats considerats a l'espai del continent europeu l'Estat Espanyol és el 6è en població, la contribució que fa en nombre de casos de Covid19 és substancial. L'Estat Espanyol aporta al voltant d'un 14% de tots els casos positius i defuncions que es registren a Europa, en canvi els increments diaris en ambdós paràmetres representen entre 6-10% dels casos continentals. La diferència de vora dues vegades entre els casos totals i els increments diaris mostren una primera conclusió: **el gran nombre de caos acumulats prové dels casos de la primera onada**.

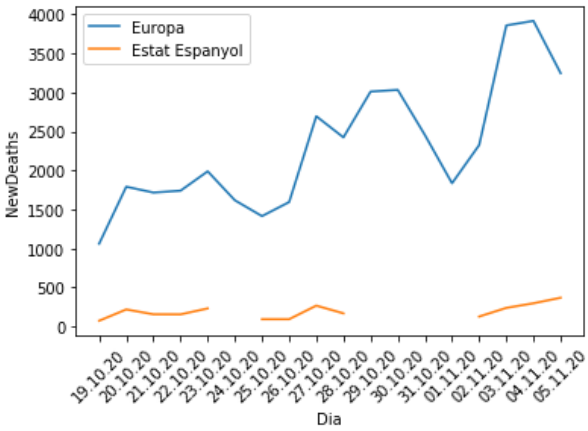
Per saber si el patró de casos, atribuïble a l'avanç de la pandèmia, és comparable a la de la resta d'Europa o si és deguda a un determinada política social o sanitària, hem representat gràfics comparatius per a cadascuna de les variables mitjançant un senzill script de *Python* (***Pra1_Covid_MP_AG_Grafics_Comparatius.py***, veure segona part del codi aportat al **punt 9**):



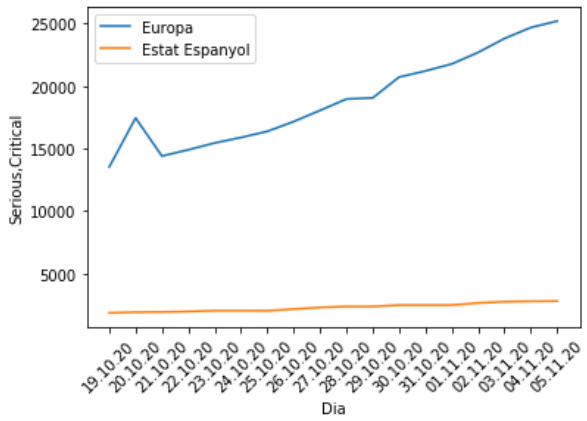




La primera comparativa mostra com tant a Europa com a l'Estat Espanyol hi ha hagut una tendència a l'alça en el nombre de casos totals de Covid19. La incidència acumulada diària presenta un patró d'alt-i-baixos, mentre que a l'Estat Espanyol té una aparença més constant. Hem de remarcar que aquest fet es pot deure a què molts països no actualitzen les dades durant el cap de setmana, tal i com podem veure que passa en el cas de l'Estat Espanyol.




El que hem comentat anteriorment per als casos totals és igualment de vàlid per als casos de les morts. De fet, en aquest cas es veu molt més clarament que les davallades en la corba de noves morts coincideix plenament en els caps de setmana.




Finalment, el casos seriosos per Covid19 són els que presenten una tendència més marcada a l'alça en tots dos contextos.

Per tant, podem concloure que, tot i ser molt notòries pel volum de casos que representen, **els patrons que descriuen les dades segueixen la mateixa tendència que a la resta del continent**.

## 8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

**- Released Under CC0: Public Domain License**

**- Released Under CC BY-NC-SA 4.0 License**

**- Released Under CC BY-SA 4.0 License**

**- Database released under Open Database License, individual contents under Database Contents License**

**- Other (specified above)**

**- Unknown License**

Seguint les indicacions *Creative Commons* (https://creativecommons.org/licenses/?lang=ca), ens hem decantat per la llicència **Reconeixement-NoComercial (CC BY-NC)**. Aquest llicència permet que altres usuaris modifiquen, adapten i utilitzen les dades, sempre que ens citen, però no amb finalitats comercial . Considerem que així aconseguirem una màxima difusió de la nostra recerca en l'aspecte acadèmic i alhora evitarem que algú en traga profit econòmic d'una treball que s'ha elaborat mitjançant l'ús gratuït de dades acumulades per un tercer.

## 9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

En les següents línies incloem el codi que s'ha emprat per a la *Web Scraping* i per a l'elaboració dels gràfics presentats en aquesta pràctica. El codi també s'ha adjuntat en format de codi *py* per a una consulta més òptima:

**Pra1_Covid_MP_AG_Web_Scraping.py**

```
# -*- coding: utf-8 -*-

#    ------------------ IMPORTACIÓ LLIBRERIES ----------------------
from bs4 import BeautifulSoup
import requests
import time
import os

#    ------------------  WEB SCRAPING  ----------------------
# Seleccionem la pàgina:
url = "https://www.worldometers.info/coronavirus/#countries"

# Baixem el codi de la pàgina:
req = requests.get(url)
soup = BeautifulSoup(req.content, 'html.parser')

# Definim el fitxer .csv on guardarem les dades i definim variables:
csv = "C:\python\Covid.csv"
fichero = open(csv, 'a+')

#Comprovem si el fitxer existeix, si no  es crearà i s'afegirà la capçalera amb el títol del camps
EscriureTitol=False
if os.stat("C:\python\Covid.csv").st_size == 0:
   EscriureTitol =True

# Afegim la data del dia en què hem baixat les dades, si no hem d'escriure la capçalera:
if EscriureTitol == False:
   fichero.write(time.strftime("%d/%m/%y")+";")

# Cerquem la capçalera de la taula a "rascar" i baixem les dades per a Europa:
taula_europa = soup.findAll("table", id="main_table_countries_today" )
for i in taula_europa:
   #capçalera del .csv només quan es fitxer no te res
```

```
    if EscriureTitol == True:
        titol = i.find('tr').get_text()
        fichero.write(";Europe;;;;;;;;Spain;;;;;;;")
        fichero.write('\n')
        titol_net= titol.split(sep='\n')
        fichero.write("Date;")
        titol_net= titol_net[3:13]
        text =""
        for j in range(8):
            text = text + titol_net[j]+";"
        fichero.write(text)
        fichero.write(text)
        fichero.write('\n')
        fichero.write(time.strftime("%d/%m/%y")+";")
    resum = i.find('tr', attrs={'data-continent':"Europe"}).get_text()
    totalsEuropa = resum.split(sep='\n')

# Extraiem les dades del total d'Europa que ens interessen i en netegem el format:
totalsEuropa_netes = totalsEuropa[5:13]
for j in range(8):
    text=totalsEuropa_netes[j]
    text=text.replace('\n',"")
    text=text.replace(',',"")
    text=text.replace('+',"")
    fichero.write(text+';')

# Extraiem les dades per a l'Estat Espanyol que ens interessen i en netegem el format:
espanya = soup.find(href="country/spain/").parent.find_next_siblings()
for j in range(8):
    text =espanya[j].get_text()
    text=text.replace('\n',"")
    text=text.replace(',',"")
    text=text.replace('+',"")
    fichero.write(text+';')

fichero.write('\n')
```

**Pra1_Covid_MP_AG_Grafics_Comparatius.py**

```
# ------------ GENERACIÓ DE GRÀFIQUES COMPARATIVES ----------------
# Importem llibreries
```

```
import pandas as pd
import matplotlib.pyplot as plt
# Carreguem amb pandas el fitxer Covid.csv com un dataframe:
dadescovid=pd.read_csv("Covid.csv", sep=';', index_col=0, header=1)
# Seleccionem el nom de les fileres, la data, com a variable per a l'eix de les x:
x=dadescovid.index
# Iteracionem amb un for-loop sobre els primeres 8 columnes per a recòrrer totes les variables:
for i in list(range(8)):
    # Seleccionem la columna corresponent al conjunt d'Europa:
    y=dadescovid.iloc[:,i]
    # Seleccionem la columna corresponent a l'Estat Espanyol:
    y2=dadescovid.iloc[:,(i+8)]
    # Muntem un gràfic de línies per a ambdues variables:
    fig = plt.figure()
    ax = fig.add_subplot(1,1,1)
    ax.plot(x, y, label='Europa')
    ax.plot(x, y2, label='Estat Espanyol')
    plt.legend(loc='upper left')
    plt.ylabel(dadescovid.columns[i])
    plt.xticks(rotation=45)
    plt.xlabel('Dia')
    plt.savefig(str(dadescovid.columns[i]), bbox_inches='tight')
    plt.show()
    plt.close()
```

## 10. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

El data set Covid.csv ha sigut desat al repositori Zenodo amb el *Digital Object Identifier* (doi) següent: https://doi.org/10.5281/zenodo.4263296.

| Contribució | Signatura |
|---|---|
| Disseny i concepció de la pràctica | Marta Puidemasa i Antoni Garcia |
| Recerca prèvia de la temàtica | Marta Puidemasa i Antoni Garcia |
| Redacció i elaboració de respostes | Marta Puidemasa i Antoni Garcia |
| Disseny de la imatge representativa | Marta Puidemasa i Antoni Garcia |
| Elaboració i desenvolupament del codi | Marta Puidemasa i Antoni Garcia |
| Execucions i discussions sobre la pràctica | Marta Puidemasa i Antoni Garcia |