# **Práctica 1: Web scraping**

### **Descripción**

El objetivo de este documento es describir la práctica 1 “Web Scraping” de la asignatura *Tipología y ciclo de vida de los datos*, asignatura del Máster en Ciencia de Datos de la Universitat Oberta de Catalunya.

En la práctica consiste en la creación de un dataset relativo a valores climatológicos diarios de España obtenidos a partir de diversas fuentes. Se han utilizado técnicas de web scraping (selenium, parseado de URL y procesado de HTML) y consultas a API públicas mediante el lenguaje de programación Python. Las fuentes de datos son Ogimet (<https://www.ogimet.com/>) y el API de provisión de datos AEMET OpenData (<https://opendata.aemet.es/>)

**Miembros del equipo**

La práctica ha sido realizada por los estudiantes del Máster en Ciencia de Datos **Francisco Jesús Cárdenas Ruiz** y **Jesús Manuel Montero Garrido**

## **Ficheros del código fuente**

Se han generado los siguientes ficheros de código Python

**src/principal.py**: punto de entrada al programa. Inicia los diferentes procesos de scraping

**src/indicativo.py**: contiene clase *indicativo* que encapsula la funcionalidad de conexión a través de Selenium a la siguiente dirección [https://ogimet.com/indicativos.phtm](https://ogimet.com/indicativos.phtml)l. Automatiza la entrada de datos a un formulario para la obtención de los indicativos sinópticos de las estaciones meteorológicas de Agencia Estatal de Meteorología (AEMET). El resultado es un listado con los indicativos sinópticos de las estacions meteorológicas.

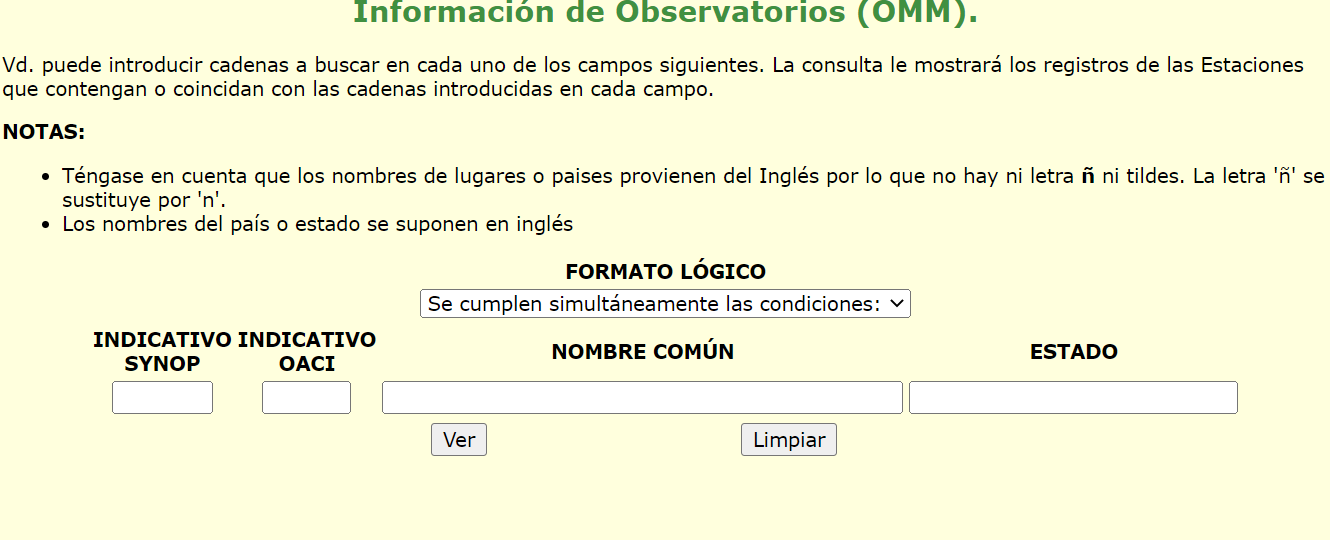


Figura 1: Formulario que es automatizado por selenium para la obtención de los indicativos sinópticos.

**src/url.py:** contiene clase url que encapsula la funcionalidad de la técnica de webscraping denominada variación de parámetros en una URL. En particular, esta clase encapsula la construcción y variación de parámetros de esta URL:

https://ogimet.com/cgi-bin/gsynres?ind=<indicativo>&ndays=<num\_dias>&ano=<anio>&mes=<mes>&day=<dia>&hora=<hora>&ord=REV&enviar=Ver

Donde:

Indicativo es el indicativo sinóptico

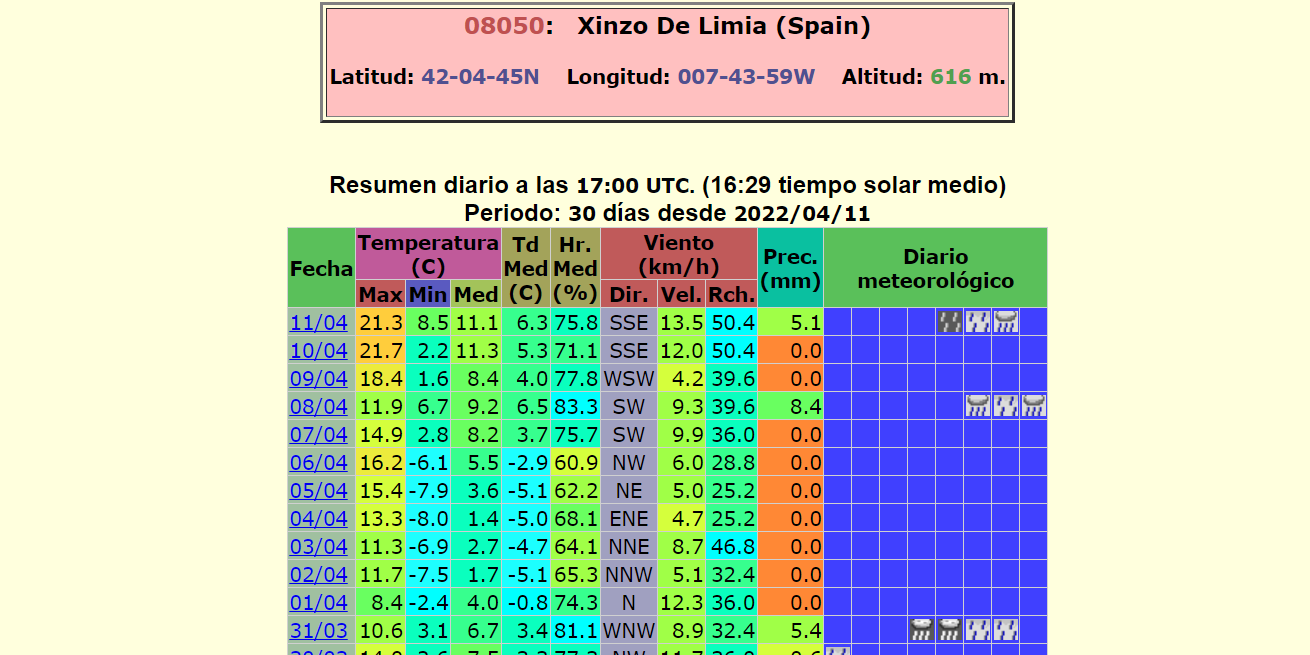
num\_dias: número de días para los que se piden datos

anio: año del que se piden datos

mes: mes del que se piden datos

hora: hora desde la que se piden datos.

El resultado de una petición a través de este formulario es la siguiente página web que es parseada para obtener los datos de los valores climatológicos.



**src/conexion.py:** contiene la clase conexion que encapsula la funcionalidad para acceder a las diferentes URL. Tiene implementado el algoritmo Exponential Backoff (<https://en.wikipedia.org/wiki/Exponential_backoff>) con el objeto de separar los sucesivos intentos de conexión fallidos y evitar una posible denegación de servicio (DoS) de los recursos de OGIMET.

**src/ogimet.py:** contiene la case ogimet que encapsula los diferentes métodos que posibilitan el parseado del HTML de la página donde están los valores climatológicos diarios. El resultado de este análisis es una lista de valores climatológicos diarios a partir de los indicativos sinópticos obtenidos por la clase indicativo.

**src/openData.py:** contiene la clase openData que encapsula las peticiones al API AEMET OpenData (<https://opendata.aemet.es>). Tiene implementado el algoritmo Exponential Backoff (<https://en.wikipedia.org/wiki/Exponential_backoff>) con el objeto de separar los sucesivos intentos de conexión fallidos y evitar una posible denegación de servicio (DoS) de los recursos de AEMET OpenData.

**src/observacion.py**: contiene la clase *observacion* que se trata de un objeto plano que contiene todas las propiedades que una observación diaria puede tener. Es utilizada para encapsular los valores climatológicos diarios obtenidos por la clase ogimet y por la clase opendata. Posibilita la homogeneización de todas los valores procedentes de todas las estaciones.

**src/utilidades.py:** contiene funciones auxiliares utilizadas por el resto de clases. Se trata de una especie de cajón desastre. En particular y entre otras funcionalidades, contiene una rutina para escribir las instancias de la clase observación al fichero de salida.

src/vcde.py: clase que contiene un método denominado obtención\_registros\_diarios del que se obtienen todos los valores climatológicos diarios de España procedentes de OGIMET y AEMET OpenData

**Recursos**