Máster Universitario en Ciencia de Datos
Cristopher Patricio Valenzuela Espinoza
Lydia López Sánchez

Tipología y ciclo de vida de los datos

PRA1: ¿Cómo podemos capturar los datos de la web?

CONTEXTO

Este proyecto se centra en la obtención y publicación de un conjunto de datos en formato reutilizable y abierto obtenido a partir de la web https://www.polenes.com/home. Esta permite consultar de forma gráfica los niveles de pólenes de diferentes tipos de plantas alergénicas por provincias y ubicaciones en España. Ofrece comparaciones y series históricas, pero no es posible descargar los datos para su explotación.

CONJUNTO DE DATOS

Nombre

Recuentos semanales y acumulados de pólenes por estación (España, 2025)

Descripción

Breve resumen: dataset tabular donde cada fila contiene un registro (estación × semana × tipo de polen) con valores numéricos de recuento acumulado, etiquetas de estación y semana, y metadatos básicos (fecha inicio, fecha fin del intervalo, semana ISO). Está pensado para análisis temporales de temporada de pólenes, comparaciones entre estaciones y modelado estacional de riesgo alérgico.

Estructura propuesta (ejemplo de columnas):

- Estacion (string)
- Desde (fecha, DD/MM/YYYY)
- Hasta (fecha, DD/MM/YYYY)
- Semana (int / ISO week)
- Polen (string)
- Acumulado (string presentación)
- Acumulado_num (float)
- CSV_path (string, local) — opcional para referencia

Cada fila representa la medición correspondiente al periodo semanal indicado en esa estación y para ese tipo de polen.

Campos:

- Estación: nombre textual de la estación (p. ej. "Alicante")
- Desde: fecha de inicio del intervalo (DD/MM/YYYY)
- Hasta: fecha final del intervalo (DD/MM/YYYY)
- Semana: número ISO de la semana (dos dígitos)
- Polen: tipo de polen medido (ej. Gramineas, Olea, Cupresáceas...)
- Acumulado: valor tal como aparece en la web (texto)
- Acumulado_num: valor numérico normalizado (float) para análisis cuantitativo

Periodo de tiempo: en el proyecto se extrajeron semanas desde septiembre hasta octubre de 2025 (ejemplo). El intervalo es configurable en el script para cualquier periodo disponible en la web. El código subido está parametrizado por semanas (intervalos de 7 días) para cubrir la temporada y evitar solapamientos. Esto se configura en la función recorrer_semanas_septiembre_octubre_2025, en la variable fecha.

ASPECTOS LEGALES

Propietario de los datos

El portal www.polenes.com es mantenido por el Comité de Aerobiología Clínica de la Sociedad Española de Aerobiología e Inmunología Clínica (en adelante, SEAIC), uno de los organismos de referencia en lo que al estudio de la calidad del aire, los niveles de pólenes y la aerobiología en España se refiere.

En su apartado QUIÉNES SOMOS (SEAIC, n.d.) se explica el origen del proyecto y su evolución a lo largo del tiempo. Según informa la web, el proyecto nace a raíz de que el Dr. J Subiza Garrido-Lestache, miembro del Comité, comenzara a medir niveles de polen en la atmósfera mediante una estación ubicada en su clínica en Madrid. Con el tiempo se fueron añadiendo nuevas estaciones medidoras y, tras la llegada de Internet, empiezan a publicarse los datos de niveles de pólenes en la web.

En el marco del análisis de la propiedad del sitio, se lanza una consulta al protocolo whois:

```
import whois

whois.whois("https://wwww.polenes.com")
```

La respuesta obtenida no ofrece datos de interés más allá de que se trata de una web alojada en el servicio Dinahosting.

Análisis realizados

Restricciones para webcrawlers

En una primera exploración, no parece encontrarse ningún fichero `robots.txt` que ofrezca indicaciones acerca de si existen limitaciones para webcrawlers,

Estructura del sitio

No se ha encontrado un fichero `sitemap.xml` o similar.

Una navegación por el sitio muestra que está compuesto únicamente de seis páginas por las que se navega mediante un menú de pestañas.

Asimismo, respecto a la estimación del tamaño, la búsqueda avanzada en Google del dominio https://www.polenes.com/home únicamente nos muestra una entrada.

El reto que plantea la estructura no está pues, en este caso, en el tamaño del sitio o en la cantidad de enlaces, si no en el hecho de que los datos que se quieren obtener se generan dinámicamente en función de las selecciones del usuario en distintas listas desplegables y filtros que le ofrece la interfaz.

Esta suposición queda confirmada tras estudiar la tecnología de implementación mediante la librería de python `builtwith`:

```
import builtwith

builtwith.parse('https://www.polenes.com/home')
```

```
{'font-scripts': ['Google Font API'],
 'javascript-frameworks': ['Meteor', 'Snap.svg']}
```

La salida del comando nos confirma que los datos se generan dinámicamente mediante el framework Meteor y, a continuación, Snap.svg se encarga de graficarlos en forma de imagen incrustada en la web.

Esta estructura ha supuesto un reto para la obtención de los datos, puesto que en ningún momento los datos figuran en el código HTML de la página que es mostrada en el navegador.

Principios éticos y legales

Propietario y consideraciones éticas y legales

El propietario de los datos publicados en la web https://www.polenes.com es el Comité de Aerobiología Clínica (SEAIC), así como las estaciones que colaboran con la red. La página actúa como agregador y publicador de esos recuentos.

Para citar trabajos previos o análisis similares, se consultaron referencias públicas de aerobiología y reportes de SEAIC (la web misma y las secciones de agradecimientos/documentación sirven como referencia inicial).

Respecto al cumplimiento ético y legal, se han aplicado las siguientes medidas:

- Solo se accedió a datos públicos ya publicados en la web; no se extrajeron ni procesaron datos personales.
- Se respetaron los términos de uso visibles en la web y se limitó la frecuencia de peticiones (pausas entre acciones) para evitar sobrecargar el servidor.
- En la memoria se documenta la fecha y hora de extracción y la metodología para asegurar reproducibilidad y transparencia.
- Si se planifica redistribución, conviene comprobar licencias y atribuciones explícitas en la web y citar correctamente la fuente (SEAIC / https://www.polenes.com) y dar crédito en el dataset.

JUSTIFICACIÓN DEL PROYECTO

Según puede leerse en su web, el proyecto www.polenes.com anima a que se sumen nuevos colectores con el fin de ampliar la cobertura geográfica, mejorar los datos obtenidos y potenciar la investigación relacionada respecto a alergias provocadas por pólenes y mohos en el aire.

Se trata de una web concebida para la visualización de los datos de manera manual y que, por tanto, dificulta su explotación automatizada. Por esto mismo, aunque su interfaz pueda resultar altamente atractiva y usable para el usuario final, su estructura impide la obtención de valor añadido a partir de los datos.

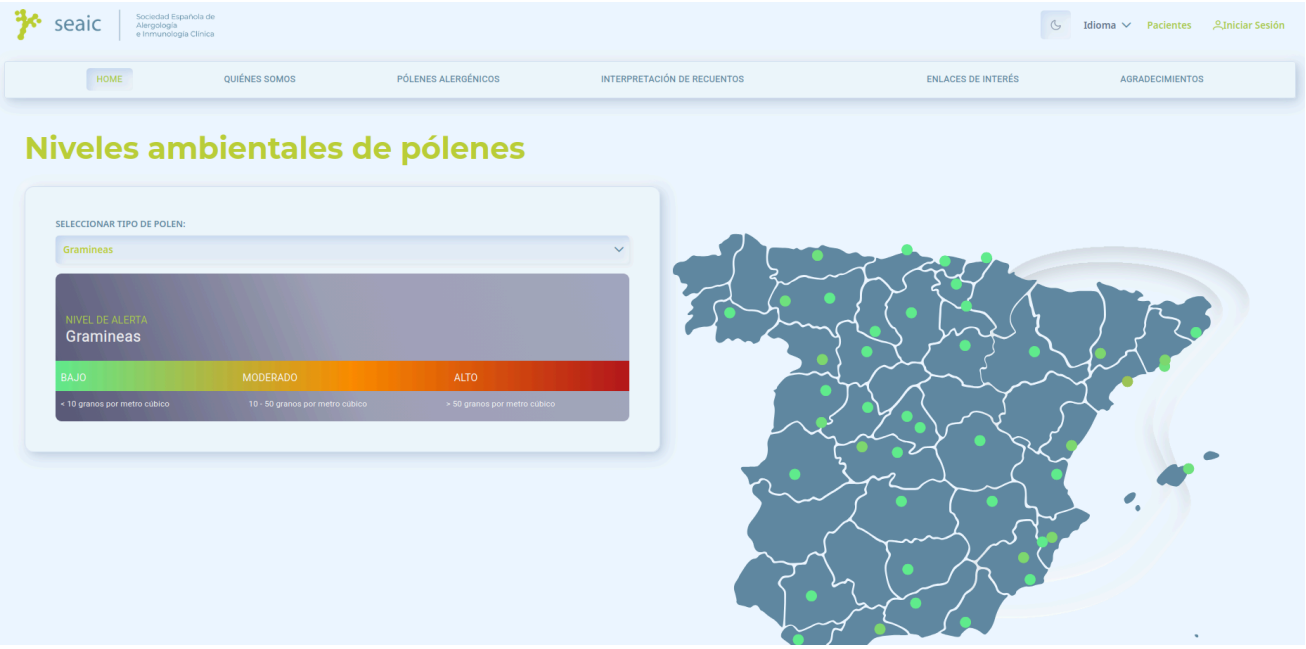

Figura 1. Estaciones de la red. Captura de pantalla de la web.

Teniendo en cuenta la filosofía de proyecto colaborativo mencionada más arriba, se ha estimado que un proyecto que permita generar un dataset reutilizable a partir de estos datos puede suponer una contribución interesante en el marco de los fines del Comité de Palinología.

Así pues, el presente proyecto se orienta a la recolección de estos datos trabajando con técnicas de web scraping y de manejo de información generada dinámicamente con javascript para generar un dataset que permita su reutilización.

La extracción de estos datos y su publicación en un dataset público y estructurado puede facilitar la implementación de casos de uso derivados tales como la predicción a futuro de niveles de pólenes en función de diversas condiciones a priori.

Asimismo, puede suponer un conjunto de datos de alto valor como fuente para estudios epidemiológicos sobre niveles registrados de asma o de urgencias neumológicas.

INFORMACIÓN TÉCNICA

Licencia del conjunto de datos

Tanto el código como el dataset se han publicado bajo licencia CC BY-NC-SA 4.0 (Creative Commons, 2013). Esta licencia únicamente impone tres restricciones a la compartición y reutilización tanto del código fuente como del dataset:

- BY - Derecho de atribución. La reutilización o redistribución de los ítems sujetos a esta licencia deben ir acompañadas de información acreditativa de los autores de ambos objetos (código y dataset), un enlace a la licencia y una descripción de aquellos cambios que, en su caso, se hayan realizado sobre los originales.
- NC - No comercial. El material no puede ser utilizado para fines comerciales.
- SA - Conservación de licencia. Se permite la transformación, reutilización, mezcla o cualquier otra modificación del material siempre y cuando la obra derivada resultante conserve el mismo licenciamiento que esta.

Asimismo, la licencia no permite la aplicación de términos legales ni medidas tecnológicas que supongan a terceros la imposición de restricciones adicionales sobre lo nada de lo que esta licencia permite.

La elección de la licencia se justifica por el propio carácter de la web origen, que parte de un trabajo colaborativo y con intención divulgativa e investigadora.

Asimismo, se estima igualmente que esta licencia es la que mejor se adapta a las características del proyecto académico cuya motivación principal ha sido la mejora de unos datos públicos para favorecer su reutilización.

Código fuente

El código fuente se encuentra disponible en el siguiente repositorio de github:

- https://github.com/lylosa/recuento-polen-spain

Información acerca del código y el proceso de recolección

Una vez se crea el driver para acceder a la página, tenemos una función `select_all_polens_once` para dejar marcados todos los tipos de pólenes, ya que queremos obtener datos de todos ellos en todas las estaciones

Las distintas estaciones se seleccionan mediante la función `select_station`. Para obtener el listado hay que emular una navegación manual en la que se fuera seleccionando cada estación en la lista desplegable.

La función `click_day_exact` controla las fechas para realizar clic en el día exacto del que se quieren obtener los valores.

La interfaz de la web va generando dinámicamente unas tablas que se muestran en la pantalla. Estos valores son extraídos por la función `extract_table`.

Dataset

En su última versión, se encuentra publicado tanto en el repositorio de Github como en zenodo con el DOI: https://doi.org/10.5281/zenodo.17575434

Presentación

Se han realizado sendos vídeos explicativos como presentación al proyecto que se corresponden, respectivamente, con una introducción teórica y una más práctica.

Parte 1:

- Introducción al proyecto, investigación previa y análisis de la web.
- Enlace: https://drive.google.com/file/d/1gPKYihEi68JQiRhpl8bF9fsN5PGM1pmj/view?usp=drive_link

Parte 2:

- Demostración del código y el proceso de web scraping
- Enlace: https://drive.google.com/file/d/1oRS0q-cr4oNo7AyNn5E9tQiUrzsiLJ-s/view?usp=drive_link

CONTRIBUCIONES

Contribuciones	Firma
Investigación previa	LLS, CPVE
Redacción de las respuestas	LLS, CPVE
Desarrollo del código	CPVE, LLS
Participación en el vídeo	CPVE, LLS

REFERENCIAS BIBLIOGRÁFICAS Y FUENTES EXTERNAS

References

Creative Commons. (2013). *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)* [Licencia]. Creative Commons. https://creativecommons.org/licenses/by-nc-sa/4.0/

SEAIC. (n.d.). *SEAIC: Quiénes somos*. SEAIC | Niveles ambientales de polen. Retrieved 10 25, 2025, from https://www.polenes.com/who-we-are