Universitat Oberta de Catalunya

uoc.edu

Tipología y ciclo de vida de los datos

PRAC 1

DESCRIPCIÓN DE LA PRÁCTICA

Sara Belén Ramos González – 42221784V

Daniel Jesús Cruz Garzón – 77158362D

Contexto

Llevaremos a cabo un estudio de datos referentes a la agricultura ecológica para poder dar solución a varias cuestiones que han surgido en una empresa dedicada a la biotecnología agroalimentaria que investiga y desarrolla soluciones científicas. Para este estudio, necesitaremos recolectar información "externa" a la empresa, que tenga carácter público, a la que podamos acceder mediante técnicas de web scrapping, que permitan automatizar la extracción y almacenamiento de la información.

Los datos que necesitamos tendrán que dar respuesta a la toma de decisiones en dicha empresa en base a un análisis de mercado sobre la agricultura ecológica que llevaremos a cabo en la Comunidad Autónoma de Andalucía. Investigaremos entonces para encontrar información sobre los operadores que intervienen en agricultura ecológica en esta área geográfica, así como de las empresas a las que pertenecen y dónde se ubican dentro de la geografía andaluza.



Realizaremos un modelado de la arquitectura de la solución tecnológica necesaria para llevar a cabo los pasos de este estudio del dato, usando las distintas herramientas y lenguajes de programación que conocemos.

También trabajaremos en el modelado de los datos en sí mismos, realizando su análisis funcional correspondiente para entenderlos y estructurarlos de la mejor manera posible. Consecuentemente, el resultado será un dataset preparado para continuar con las tareas de análisis que conlleva la ejecución del estudio completo.

1. Título

Operadores Andaluces en Agricultura Ecológica (segmentados según Provincias).

2. Descripción del dataset

Se trata de un conjunto de datos público, alojado en una base de datos que conforman el Registro General de Operadores Ecológicos (REGOE).

El estudio de estos datos permitirá identificar a los operadores ecológicos según su actividad, ubicación y producción, entendiendo por operador ecológico cualquier persona física o jurídica que tenga interés en producir, transformar, elaborar o envasar alimentos de origen agrario con el fin de comercializarlos utilizando los términos ecológico, biológico u orgánico y sus abreviaturas.

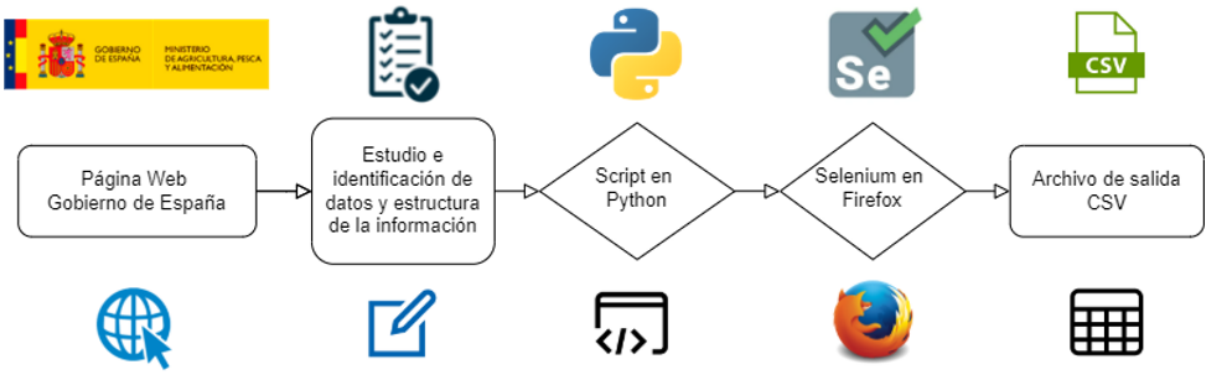
En este caso, los datos se distribuyen en la web del ministerio en forma de varias tablas paginadas en las que se nos permite hacer filtrado sobre los datos según el nombre de la empresa, la Comunidad Autónoma donde opera, Provincia y/o Actividad a la que se dedica dicho operador.

Para llevar a cabo esta práctica, aplicaremos un filtro en la búsqueda basada en la Comunidad Autónoma, en el cual definiremos "Andalucía" como opción elegida, y tras éste, todas sus provincias correspondientes (que es lo que nos interesa para nuestro estudio).

Mediante la extracción y transformación que llevamos a cabo con nuestro script en Python, obtendremos una única tabla de datos que aglutinará todos los valores de manera ordenada en sus filas y columnas correspondientes.

3. Representación gráfica

Arquitectura de la solución



4. Contenido

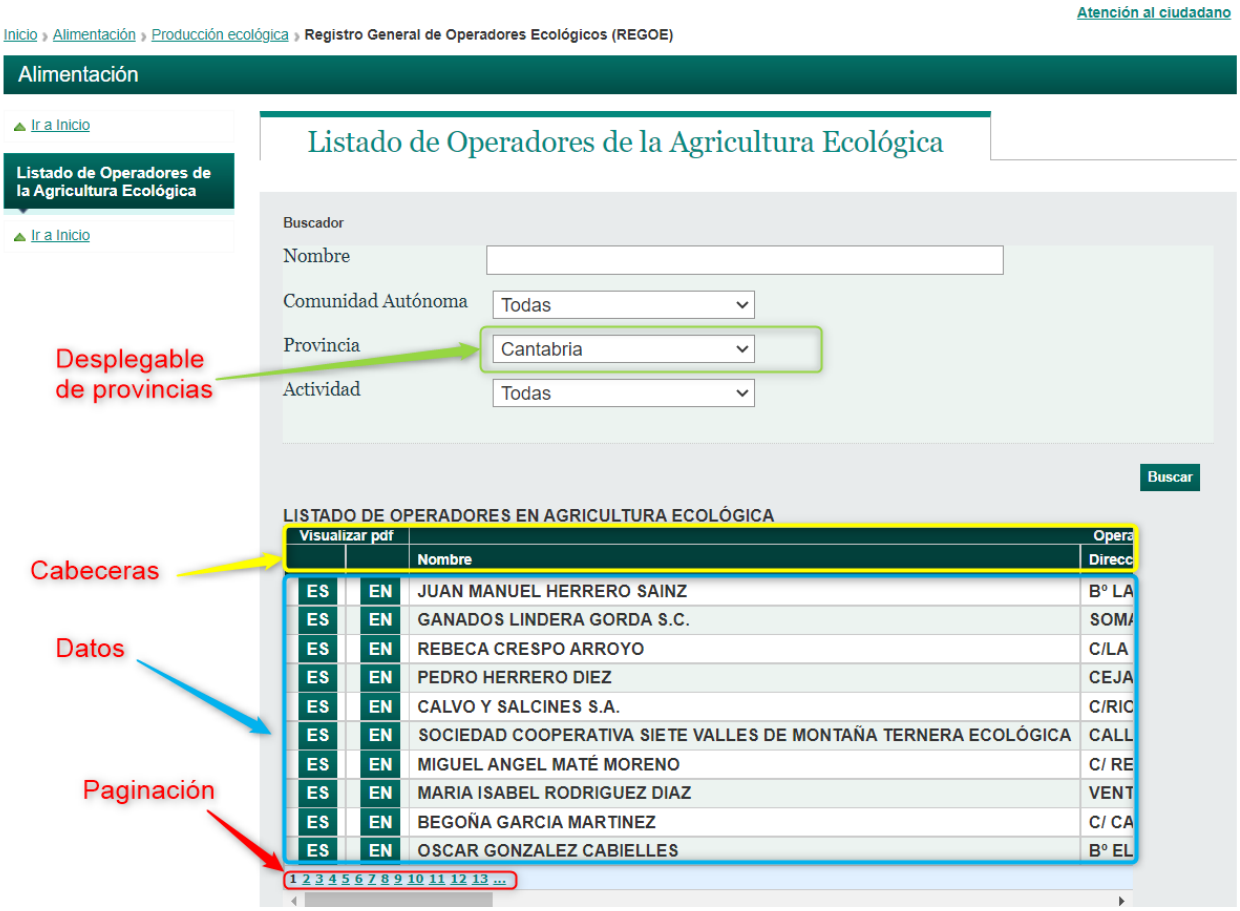
El dataset que hemos descrito en el apartado anterior consta de datos que han sido cargados a lo largo de los dos últimos años (2021, 2022).

A continuación detallamos qué datos podemos encontrar, tanto si visitamos la página web, como si consultamos nuestro archivo de extracción de datos generado a partir de nuestro script.

4.1. Contenido de la web

Podemos acceder a la página web que contiene la información que estudiaremos pinchando en el siguiente enlace: https://servicio.mapama.gob.es/regoe/Publica/Operadores.aspx

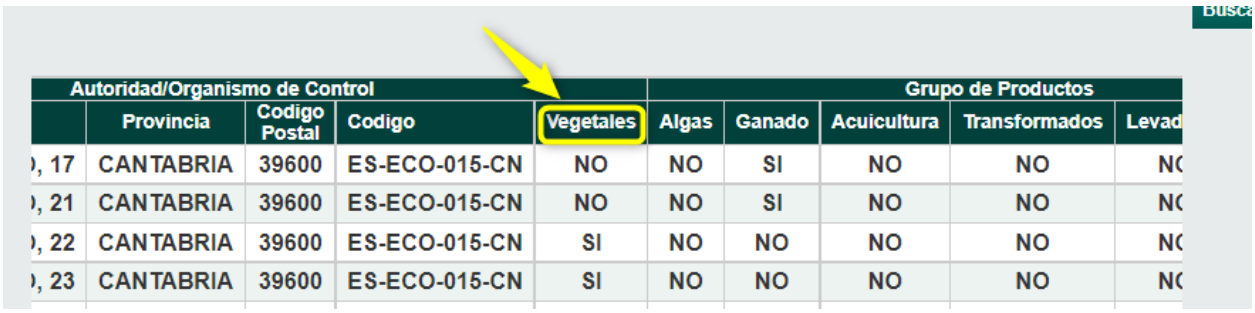
En esta web encontraremos, en cada página, un conjunto de valores estructurados en forma de un tabla, en función de la provincia elegida en el desplegable correspondiente, siendo la apariencia tal y como se muestra en la siguiente imagen:



Esta tabla está compuesta por:

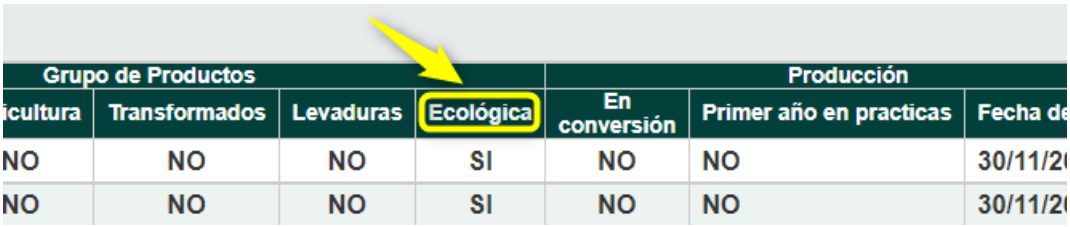
- 27 columnas: son los distintos campos que definen la información que contendrá cada registro. Cada campo describirá una característica distinta de cada operador de agricultura.
- 2 registros de cabecera: concretamente las dos primeras filas, que recogen los nombres de los campos.

Prestar especial atención a que, en un primer análisis visual de los datos, nos damos cuenta de que hay un error estructural en la cabecera (primer registro), ya que no aglutina bien los valores del segundo registro de la cabecera. Por ejemplo, el campo <<Vegetales>> del segundo registro, debería corresponder al campo (del primer registro) <<Grupo de Productos>>, y en cambio, aparece dentro del campo << Autoridad/Organismo de Control>> (del primer registro también). Esto lo podemos observar en la captura que se muestra a continuación:



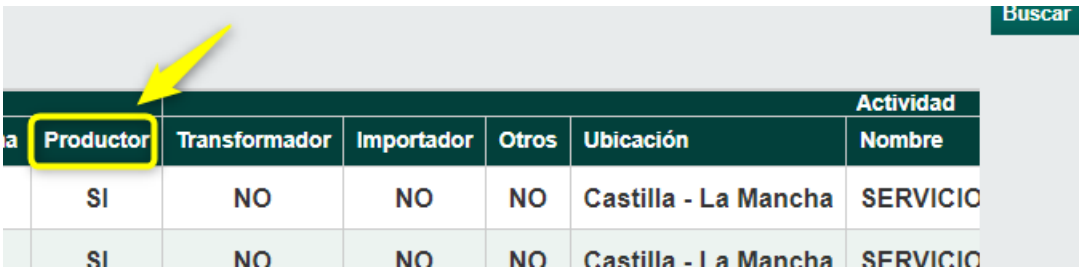
Autoridad/Organismo de Control				Grupo de Productos					
	Provincia	Codigo Postal	Codigo	Vegetales	Algas	Ganado	Acuicultura	Transformados	Levad
), 17	CANTABRIA	39600	ES-ECO-015-CN	NO	NO	SI	NO	NO	NO
), 21	CANTABRIA	39600	ES-ECO-015-CN	NO	NO	SI	NO	NO	NO
), 22	CANTABRIA	39600	ES-ECO-015-CN	SI	NO	NO	NO	NO	NO
), 23	CANTABRIA	39600	ES-ECO-015-CN	SI	NO	NO	NO	NO	NO

Ocurre lo mismo con el campo <<Ecológica>> del segundo registro, que no está bien relacionado con su campo correspondiente del primer registro de datos, ya que debería de estar incluido dentro del campo <<Producción>> (del primer registro). Podemos observarlo en la siguiente imagen:



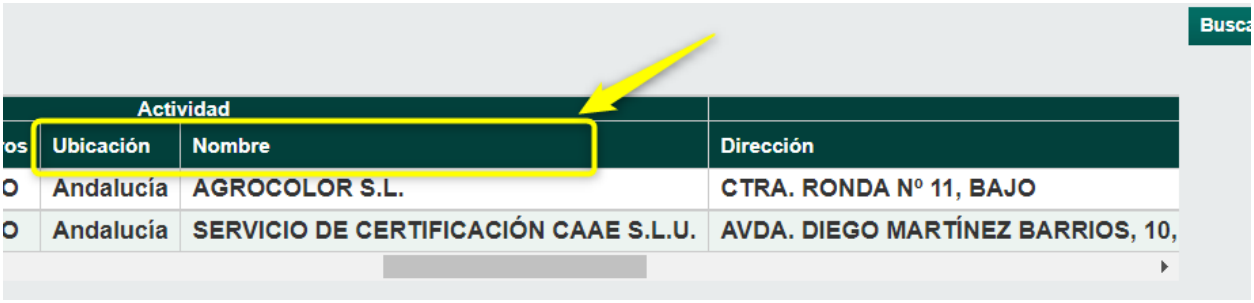
Grupo de Productos				Producción		
icultura	Transformados	Levaduras	Ecológica	En conversión	Primer año en practicas	Fecha de
NO	NO	NO	SI	NO	NO	30/11/2(
NO	NO	NO	SI	NO	NO	30/11/2(

También con el campo <<Productor>>, tal y como mostramos en esta captura:



		Actividad				
a	Productor	Transformador	Importador	Otros	Ubicación	Nombre
	SI	NO	NO	NO	Castilla - La Mancha	SERVICIO
	SI	NO	NO	NO	Castilla - La Mancha	SERVICIO

Y en cuanto a los campos <<Ubicación>> y <<Nombre>> que aparecen en la cabecera <<Actividad>>, deberían de aparecer en la cabecera << Autoridad/Organismo de Control>>



Actividad			
os	Ubicación	Nombre	Dirección
O	Andalucía	AGROCOLOR S.L.	CTRA. RONDA Nº 11, BAJO
O	Andalucía	SERVICIO DE CERTIFICACIÓN CAAE S.L.U.	AVDA. DIEGO MARTÍNEZ BARRIOS, 10,

- 10 registros: el resto de filas de la tabla de cada página, conformados por los valores informados de los campos definidos en su cabecera correspondiente. Cada registro albergará un operador de agricultura distinto.

4.2. Contenido de datos descargados

Nuestro dataset final, resultante de aplicar el script de extracción, tendrá la apariencia que se muestra a continuación:

	A	B	C	D	E	F	G	H	I	J	
1	Nombre	DirecciÃ³n	Provincia	Comunidad AutÃ³noma	Productor	Transformadc	Importador	Otros	UbicaciÃ³n	Nombre	DirecciÃ³n
8	JUAN BAUTISTA GAZQUEZ MARTINEZ	C/ AMISTAD, !	ALMERÃA	ANDALUCÃA	SI	NO	NO	NO	AndalucÃ-a	SERVICIO DE CERTIFICACIÃ"N CAAE S.L.U.	AVDA. DIEGO MARTÃNEZ BARRI
9	JOSEFA ARRIETA GARCIA	C/ CANONIGC	ALMERÃA	ANDALUCÃA	SI	NO	NO	NO	AndalucÃ-a	SERVICIO DE CERTIFICACIÃ"N CAAE S.L.U.	AVDA. DIEGO MARTÃNEZ BARRI
10	MARIA JESUS PARIS BELTRAN	PARAJE LA FU	ALMERÃA	ANDALUCÃA	SI	NO	NO	NO	AndalucÃ-a	AGROCOLOR S.L.	CTRA. RONDA NÂº 11, BAJO
11	PABLO IGLESIAS REAL	C/ LUSSEX, 10	ALMERÃA	ANDALUCÃA	SI	NO	NO	NO	AndalucÃ-a	SERVICIO DE CERTIFICACIÃ"N CAAE S.L.U.	AVDA. DIEGO MARTÃNEZ BARRI
12	ANDRES MONTIEL MORATA	C/ VALERO RI	ALMERÃA	ANDALUCÃA	SI	NO	NO	NO	AndalucÃ-a	SERVICIO DE CERTIFICACIÃ"N CAAE S.L.U.	AVDA. DIEGO MARTÃNEZ BARRI

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Provincia	Codigo Postal	Codigo	Vegetales	Algas	Ganado	Acuicultura	Transformadc	Levaduras	EcolÃ³gica	En conversiÃ³	Primer aÃ±o e	Fecha de carga	Fecha de control	
8	SEVILLA	41013	ES-ECO-001-AN	SI	NO	NO	NO	NO	NO	NO	SI	NO	27/12/2021	25/11/2021	
9	SEVILLA	41013	ES-ECO-001-AN	SI	NO	NO	NO	NO	NO	SI	NO	NO	27/12/2021	23/04/2021	
10	ALMERÃA	4004	ES-ECO-003-AN	SI	NO	NO	NO	NO	NO	SI	NO	NO	27/12/2021	14/05/2021	
11	SEVILLA	41013	ES-ECO-001-AN	SI	NO	NO	NO	NO	NO	SI	NO	NO	27/12/2021	01/12/2020	
12	SEVILLA	41013	ES-ECO-001-AN	SI	NO	NO	NO	NO	NO	SI	NO	NO	27/12/2021	21/12/2021	

Este dataset, estará alojado en un archivo de extensión csv, e incluirá lo siguiente:

- Un total de 25 campos, ya que prescindiremos de las dos primeras columnas de la tabla (porque son campos que no están informados con valores, sino con los enlaces url que redirigen a la ficha de características del operador correspondiente, bien en español, o bien en inglés).
- Un primer registro en el archivo resultante, que contendrá los valores de cabecera (coincidentes con el segundo registro de la tabla que nos proporciona la web).

Haciendo la extracción solamente con el segundo registro para definir la cabecera, evitamos la confusión que podría ocasionar la correspondencia incorrecta de los campos que explicábamos en el contenido de la web (apartardo 4.1), sobre todo a la hora de interpretar los resultados.

- El resto de registros contenidos en el fichero resultante serán los datos arrojados en la tabla de cada página, y constituyen los valores que tendremos que analizar.
- Descripción de los campos:

  ✔ Información relativa al operador:

  - Nombre: nombre del operador concreto sobre el que se vuelca la información en cada celda del dataset.
  - Dirección: la dirección donde está registrado el operador concreto.
  - Provincia: provincia donde está registrado el operador concreto.
  - Comunidad autónoma: Comunidad Autónoma donde está registrado el operador concreto.

  ✔ Información relativa al tipo de actividad del operador:

- Productor: Campo informado con los valores SI/NO, que indica si el operador es productor o no.
- Transformador: Campo informado con los valores SI/NO, que indica si el operador es transformador, o no.
- Importador: Campo informado con los valores SI/NO, que indica si el operador es importador, o no.
- Otros: Campo informado con los valores SI/NO, que indica si el operador se dedica a otra actividad o no.

✔ Información relativa a la Autoridad/Organismo de Control (entidad que se encarga de que se cumplan los requisitos de aplicación de la producción ecológica) que supervisa al operador:

- Ubicación: Comunidad Autónoma donde se ubica la Autoridad/Organismo de Control.
- Nombre: Nombre de la Autoridad/Organismo de Control.
- Dirección: Dirección donde se ubica registrada la Autoridad/Organismo de Control.
- Provincia: Provincia donde se ubica la Autoridad/Organismo de Control.
- Codigo Postal: Código postal correspondiente a la zona donde se ubica la Autoridad/Organismo de Control.
- Codigo: Código que identifica a la Autoridad/Organismo de Control.

✔ Grupo de productos evaluados del operador determinado:

- Vegetales: Campo informado con los valores SI/NO, que indica si el operador trabaja vegetales, o no.
- Algas: Campo informado con los valores SI/NO, que indica si el operador trabaja con algas, o no.
- Ganado: Campo informado con los valores SI/NO, que indica si el operador trabaja ganado, o no.
- Acuicultura: Campo informado con los valores SI/NO, que indica si el operador trabaja acuicultura, o no.
- Transformados: Campo informado con los valores SI/NO, que indica si el operador trabaja productos transformados, o no.
- Levaduras: Campo informado con los valores SI/NO, que indica si el operador trabaja levaduras, o no.

✔ Datos relativos al estado de la producción del operador estudiado:

- Ecológica: la producción del operador es ecológica.
- En conversión: la producción del operador está en conversión para ser ecológica.
- Primer año en practicas: Campo informado con los valores SI/NO, que indica si la producción se encuentra en el primer año en prácticas, o no.

✔ Fechas:

- Fecha de carga: no será relevante para nuestro estudio.
- Fecha de control: fecha en la que se ha producido el último control.

4.3. Contenido del fichero “log” del script

Hemos incluido en nuestro script la exportación de un fichero log donde se va almacenando cada paso que se va ejecutando en el código para controlar su correcto funcionamiento y capturar los posibles errores.

5. Agradecimientos

Agradecemos la publicación de los datos al “Ministerio de Agricultura, Pesca y Alimentación” del Gobierno Español, que es la entidad que posibilita la consulta de estos datos.

También agradecemos la colaboración a las Autoridades Competentes de las Comunidades Autónomas que son quiénes suministran la información que recoge esta base de datos que hemos consultado.

Y por supuesto, no podemos olvidarnos de transmitir nuestros agradecimientos a Jason Huggins, quien ha sido el creador de la librería Selenium que hemos implementado en el script codificado en Python para acceder al contenido de la página web donde se alojan los datos que necesitamos.

6. Inspiración

6.1. Por qué estos datos

- El primer motivo que nos ha llevado a estudiar estos datos, es la creciente demanda de productos ecológicos, biológicos y orgánicos que se está produciendo en nuestra sociedad actual.
- También ha resultado inspirador el hecho de que se trate de un tema que nos preocupa bastante hoy en día como seres humanos, ya que está muy ligado a nuestra salud y a la longevidad.

- La tercera causa que nos ha movido a llevar a cabo este estudio, es que trata sobre un sector en auge que será susceptible de ser analizado por los especialistas en datos en los próximos años y quizás podremos usarlo para un estudio futuro.
- Por último, y no por eso menos importante, trabajamos y vivimos en la ciudad de Almería, una provincia bastante involucrada en el sector de la agricultura y la biotecnología, lo cual despierta en nosotros cierta curiosidad sobre qué conclusiones podemos sacar si analizamos datos relacionado con este ámbito en nuestra Comunidad Autónoma.

6.2. Preguntas a resolver

Como ya explicamos anteriormente, necesitamos hacer un análisis de mercado sobre la agricultura ecológica en Andalucía. Con este análisis podemos responder a preguntas tales como:

- ¿En qué provincia de Andalucía hay más operadores cuya producción es ecológica y son clientes potenciales para la venta de productos biotecnológicos? Respondiendo a esta pregunta podremos saber dónde focalizar las ventas de las soluciones.
- ¿Es Andalucía una Comunidad Autónoma en la que relativamente se puede considerar que sus operadores de agricultura son en su mayoría ecológicos? Según nuestra respuesta, decidiremos si merece la pena en un futuro expandirnos a otros territorios nacionales, o no.
- ¿Qué tipo de producción es la mayoritariamente ecológica? ¿Qué tipo de producción aún está a la cola de ser ecológica en nuestro país? Con esto podremos focalizar la investigación biotecnológica en la producción de soluciones orientadas al tipo de producción dominante. Y también, podremos proponer nuevas soluciones para los tipos de producción donde no hay bastante desarrollo aún y ser pioneros.
- ¿Cuál es el organismo de control preferido para cada zona geográfica en España? Con esto podremos ponernos en contacto con ellos y guiarnos sobre qué reglamentaciones y mecanismos debe pasar la solución que hayamos creado en el laboratorio de I+D.
- Otras preguntas que aporten soluciones para basar en ellas la toma de decisiones de la empresa biotecnológica.

6.3. Elección del método de web scrapping

Hemos escogido Selenium porque, además de que se adapta a nuestra disposición en crudo de los datos, creemos que es un método que podríamos aplicar a descargas de datos que necesitemos en un futuro para otros análisis y proyectos, ya que es un método capaz de simular un comportamiento humano de selección de datos según nuestros intereses de selección en páginas web.

7. Referencias

https://www.gobiernodecanarias.org/agp/icca/servicios/rope/

https://www.aragon.es/-/produccion-ecologica

https://www.boe.es/buscar/act.php?id=BOE-A-2014-10522

8. Licencias

En el supuesto que tratamos, al tratarse de una recopilación de datos para una posterior toma de decisiones empresariales, estos son altamente sensibles y por lo tanto propiedad de la empresa. El acceso de terceros podría comprometer la estrategia de la compañía, es por ello, que se ha optado por una licencia para el datasheet privada (individual contents under Database Contents License) donde los acuerdos de distribución de este estarán formulados según los intereses y normas de la empresa.

Sin embargo, para agilizar la práctica se ha creado con una licencia libre.

9. Código, enlace a GitHub

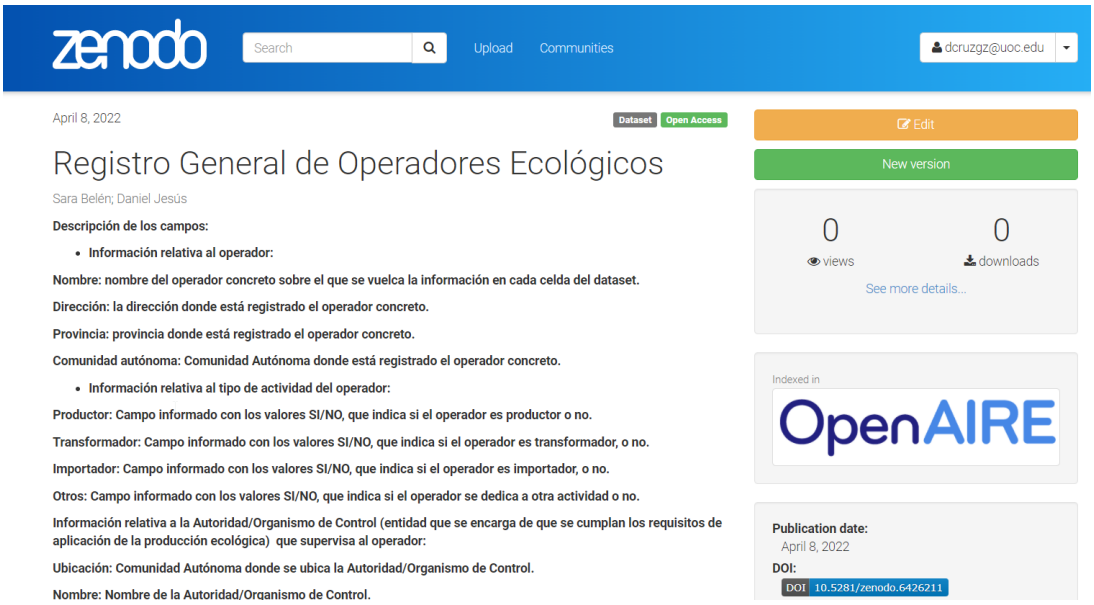
Se crea el repositorio en GitHub, accediendo al siguiente enlace. (Es privado, teniendo sólo acceso los autores y el profesor a través del correo dperez1@uoc.edu)

Enlace al repositorio en GitHub

10. Datasheet

Se sube el datasheet a zenodo. DOI: 10.5281/zenodo.6426211

https://zenodo.org/record/6426211#.YlCjAyhBxPY



CONTRIBUCIONES	FIRMA
Investigación previa	SBRG, DJCG
Redacción de las Respuestas	SBRG, DJCG
Desarrollo del código	SBRG, DJCG