
# MÁSTER UNIVERSITARIO EN BIG DATA. TECNOLOGÍA Y ANALÍTICA AVANZADA

TRABAJO FIN DE MÁSTER

# EXTRACCIÓN/ADQUISICIÓN DE DATOS DE DIVERSAS FUENTES E INTEGRACIÓN EN LA PLATAFORMA DE IMEUREKA

Autor: Miguel Enrile Fernández de Arévalo

Director: Pablo Collado Puerta

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Extracción / adquisición de datos de diversas fuentes e integración en la plataforma de IMeureka en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2020/21 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

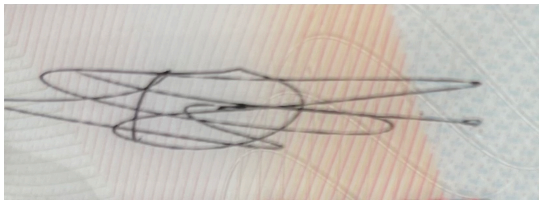
tomada de otros documentos está debidamente referenciada.

Fdo.: Miguel Enrile Fernández de Arévalo Fecha: 17/ 05/ 2021

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo: Pablo Collado Puerta Fecha: 17/ 06/ 2021



**AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO**

***1º. Declaración de la autoría y acreditación de la misma.***

El autor Ilmo.Sr.D. Miguel Enrile Fernández de Arévalo

DECLARA ser el titular de los derechos de propiedad intelectual de la obra Extracción / adquisición de datos de diversas fuentes e integración en la plataforma de IMeureka, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

***2º. Objeto y fines de la cesión.***

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

***3º. Condiciones de la cesión y acceso***

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

***4º. Derechos del autor.***

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

***5º. Deberes del autor.***

El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.
d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción

de derechos derivada de las obras objeto de la cesión.

***6º. Fines y funcionamiento del Repositorio Institucional.***

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 17 de Junio de 2021

**ACEPTA**

Fdo MIGUEL ENRILE FERNÁNDEZ DE ARÉVALO

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:


# MÁSTER UNIVERSITARIO EN BIG DATA. TECNOLOGÍA Y ANALÍTICA AVANZADA

TRABAJO FIN DE MASTER

# EXTRACCIÓN/ADQUISICIÓN DE DATOS DE DIVERSAS FUENTES E INTEGRACIÓN EN LA PLATAFORMA DE IMEUREKA

Autor: Miguel Enrile Fernández de Arévalo

Director: Pablo Collado Puerta

Madrid

# Agradecimientos

A mis padres y a mi familia por el apoyo y la confianza a lo largo de estos años.

# EXTRACCIÓN/ADQUISICIÓN DE DATOS DE DIVERSAS FUENTES E INTEGRACIÓN EN LA PLATAFORMA DE IMEUREKA

**Autor:** Enrile Fernández de Arévalo, Miguel
Director: Collado Puerta, Pablo
Entidad Colaboradora: Insurance Manager SL (IMeureka)

## RESUMEN DEL PROYECTO

El proyecto consistirá en la extracción de datos de diferentes fuentes tanto públicas como privadas y su tratamiento de tal manera que los datos puedan ser usados por los siguientes miembros de la cadena de trabajo de IMeureka. Deberán desarrollarse API's y utilizar las técnicas necesarias para extraer los datos e integrarlos en las bases de datos existentes en la empresa.

**Palabras clave**: Transformación digital, insurtech, IMeureka, Web scraping.

### 1. Introducción

La empresa Insurance Manager SL [1], ubicada en Madrid, es la propietaria de la plataforma de gestión de riesgos www.imeureka.com. Sus objetivos son dos, por un lado habilitar un entorno transparente de libre comercio, abierto a la contraoferta, en el contexto de la intermediación de pólizas de seguros, para ello, proporcionan su plataforma de gestión de riesgos e intermediación. Por otro lado, facilitar la comunicación entre la compañía aseguradora y la empresa.

La extracción e integración de los datos planificados para este trabajo permitirán crear paquetes de servicios en venta que faciliten la incorporación de nuevos corredores al Marketplace de IMeureka. Adicionalmente, estos datos ayudarán a la propia correduría IMeureka a aumentar su facturación mediante el descubrimiento de nuevas oportunidades.

### 2. Definición del proyecto

2.1 El primer objetivo del proyecto será la extracción de datos de las fuentes seleccionadas por la empresa, con los métodos que se estimen oportunos (descarga completa de Base de Datos, conexión con API Near Real Time …). Las fuentes de datos serán:

2.1.1 Catastro[2]: Los datos del catastro deberán ser accesibles desde la web de IMeureka, para ello, se llevarán a cabo las implementaciones necesarias para que los datos del catastro estén disponibles en Near-Real Time a los usuarios de la plataforma. En el momento en que los usuarios quieran la información de un inmueble, se realizará una petición al catastro, se almacenará en DB y se le mostrará.

2.1.2 Licitaciones: La aplicación deberá estar completamente integrada en la plataforma de IMeureka a través de API suficientemente generales para que puedan

ser usado por otros proyectos que requieran información de la DB. Se creará una herramienta que permita a los corredores el aprovechamiento de estos datos.

2.1.3 Fichero de Multas: Los empleados de IMeureka podrán comprobar si algún usuario/cliente tiene multas públicas pendientes.

2.1.4 Empresas españolas: Se llevará a cabo una extracción de una lista de datos de empresas españolas, almacenamiento en base de datos y se creará una herramienta que permita a los corredores de la plataforma el aprovechamiento de estos datos.

2.2 El segundo objetivo es que los desarrollos realizados queden completamente integrados dentro de la plataforma, tanto a nivel visual como funcional. Se desarrollarán funciones claras, con programación todo lo simple y estructurada posible, y se documentarán estas funciones para que el código sea reutilizable por otros programadores que no conozcan necesariamente las técnicas empleadas.

## 3. Descripción del modelo/sistema/herramienta

La extracción de datos está explicada en detalle en el ensayo, pero a modo de resumen, los pasos que se han seguido son los siguientes:

0. RGPD (Reglamento General de Protección de Datos) [2]: ¿Es legal la extracción?

1. Comprobar si existe API ya creada para facilitar la extracción / Comprobar si los datos están a la venta en internet (soluciones preexistentes)

2. Comprobar los términos y condiciones del sitio web y la propiedad intelectual de los datos que se van a extraer

3. Comprobar si se pueden interceptar peticiones XHR

4. Comprobar si se pueden simular las peticiones y extraer los datos del html u otras fuentes semi-estructuradas

5. Navegación simulada en caso de que todo fracase

## 4. Resultados

Se han extraído los datos de los objetivos salvo los del fichero de multas. Tras una investigación se concluyó que estos datos no aportaban utilidad suficiente para justificar el trabajo.

La herramienta de licitaciones permite a los corredores encontrar oportunidades de negocio en el sector público. Está diseñada de tal manera que el corredor no tiene necesidad de acceder a la página original en la cual se publican los datos.


*Ilustración 1. Visualización de una licitación en la plataforma*

La herramienta de empresas permite a los corredores filtrar entre casi un millón de empresas activas españolas. Se les ha proporcionado una manera sencilla para incrementar su productividad.

El catastro ha quedado implementado de manera transparente para los corredores, no como paquete aparte. Es un servicio adicional que se ofrece para las pólizas con inmuebles. Al escribir una dirección el corredor / cliente puede seleccionar sus datos catastrales, ahorrando tiempo y ganando en conveniencia.

## 5. Conclusiones

Los datos y las herramientas desarrolladas aportan valor a los corredores externos y a los comerciales de IMeureka. Se pretenden, en un futuro, monetizar mediante la creación de un formato de suscripción para las mismas.

## 6. Referencias


[1] IMeureka, Página web y contacto. https://www.imeureka.com/

[2] Sede Electrónica del catastro. https://www.sedecatastro.gob.es/

[3] Plataforma de contratación del Estado.
https://contrataciondelestado.es/wps/portal/licitaciones

[4] Tablón edictal único (multas):
https://sede.dgt.gob.es/es/multas/consulta-multa-tablon-edictal-unico/

[5] Empresas españolas: http://www.infocif.es/

# DATA EXTRACTION / ACQUISITION FROM VARIOUS SOURCES AND INTEGRATION INTO THE IMEUREKA PLATFORM

**Author: Miguel Enrile Fernández de Arévalo**
Supervisor: Pablo Collado Puerta
Collaborating Entity: Imeureka

## ABSTRACT

The project will consist of the extraction of data from different sources, both public and private, and its treatment in such a way that the data can be used by the following members of the IMeureka chain of work. APIs should be developed, and the necessary techniques used to extract the data and integrate it into the company's existing databases.

Keywords: Digital transformation, insurtech, IMeureka, Web scraping.

### 1. Introduction

The company Insurance Manager SL [1], located in Madrid, is the owner of the risk management platform www.imeureka.com. Its objectives are twofold, on the one hand to enable a transparent environment of free trade, open to counter offer, in the context of the intermediation of insurance policies, for this, they provide their risk management and intermediation platform. On the other hand, facilitate communication between the insurance company and the company.

The extraction and integration of the data planned for this work will allow the creation of packages of services for sale that facilitate the incorporation of new brokers to the IMeureka Marketplace. Additionally, this data will help IMeureka brokerage itself to increase its turnover by discovering new opportunities.

### 2. Project definition

2.1 The first objective of the project will be the extraction of data from the sources selected by the company, with the methods deemed appropriate (complete database download, connection with Near Real Time API ...). The data sources will be:

2.1.1 Cadastre [2]: The cadastre data must be accessible from the IMeureka website, for this, the necessary implementations will be carried out so that the cadastre data is available in Near-Real Time to the users of the platform. . As soon as users want information on a property, a request will be made to the cadastre, it will be stored in a DB and it will be shown to them.

2.1.2 Public contracts: The application must be fully integrated into the IMeureka platform through sufficiently general APIs so that they can be used by other projects that require information from the DB. A tool will be created that allows runners to take advantage of this data.

2.1.3 Fines File: IMeureka employees will be able to check if any user / client has pending public fines.

2.1.4 Spanish companies: An extraction of a list of data of Spanish companies will be carried out, storage in a database and a tool will be created that allows the brokers of the platform to take advantage of this data.

2.2 The second objective is that the developments made are fully integrated into the platform, both visually and functionally. Clear functions will be developed, with programming as simple and structured as possible, and these functions will be documented so that the code is reusable by other programmers who do not necessarily know the techniques used.

## 3. Description of the system

Data extraction is explained in detail in the trial, but in summary, the steps that have been followed are as follows:

0. RGPD (General Data Protection Regulation) [2]: Is extraction legal?

1. Check if there is an API already created to facilitate extraction / Check if the data is for sale on the internet (pre-existing solutions)

2. Check the terms and conditions of the website and the intellectual property of the data to be extracted

3. Check if XHR requests can be intercepted

4. Check if requests can be simulated and extract data from html or other semi-structured sources

5. Simulated navigation in case everything fails

## 4. Results

The data of the objectives have been extracted except those of the file of fines. After an investigation, it was concluded that these data did not provide sufficient utility to justify the work.

The bidding tool allows brokers to find business opportunities in the public sector. It is designed in such a way that the broker does not need to access the original page on which the data is published.


*Illustration 2. Viewing of a public contract on the platform*

The companies tool allows brokers to filter among almost one million active Spanish companies. They have been provided with an easy way to increase their productivity.

The cadastre has been implemented in a transparent way for brokers, not as a separate package. It is an additional service that is offered for policies with real estate. By entering an address, the broker / client can select their cadastral data, saving time and gaining convenience.

## 5. Conclusions

The data and tools developed add value to IMeureka's external brokers and sales representatives. They are intended, in the future, to be monetized by creating a subscription format for them.

## 6. Bibliography


[1] IMeureka, Website and contact. https://www.imeureka.com/

[2] Electronic Office of the cadastre. https://www.sedecatastro.gob.es/

[3] State contracting platform. https://contrataciondelestado.es/wps/portal/licitaciones

[4] TEU (fines):

https://sede.dgt.gob.es/es/multas/consulta-multa-tablon-edictal-unico/

[5] Spanish companies: http://www.infocif.es/

# *Índice de la memoria*

# *Índice de figuras*

# *Índice de tablas*

# Capítulo 1. INTRODUCCIÓN

La empresa Insurance Manager SL [1], ubicada en Madrid, es la propietaria de la plataforma de gestión de riesgos www.imeureka.com. Sus objetivos son dos: por un lado, habilitar un entorno transparente de libre comercio, abierto a la contraoferta, en el contexto de la intermediación de pólizas de seguros. Para ello, proporcionan su plataforma de gestión de riesgos e intermediación. Por otro lado, facilitar la comunicación entre la compañía aseguradora y la empresa.


*Ilustración 3: Logo de IMeureka*

IMeureka es una correduría de seguros centrada en la transformación digital, gracias a la tecnología, están colaborando en la evolución del sector de los seguros hacia modelos más transparentes y eficientes. Ofrecen servicios como la verificación de riesgos, la gestión de siniestros, Asesoría jurídica, Ciberseguridad ...

El sector asegurador en España tiene un tamaño bastante considerable (cerca de 60.000 millones de euros anuales), disfruta de un crecimiento constante y sostenible y se ha mantenido extremadamente resiliente a través de los ciclos económicos. Sin embargo, el sector ha demostrado ser difícil de penetrar tecnológicamente debido a su resistencia al cambio y ahora está listo para la puesta en marcha de tecnologías disruptivas para conseguir

las eficiencias que tanto necesita y traer de vuelta el centro de atención sobre el principal activo de la industria aseguradora: los datos.

Las características del mercado asegurador (mercado grande y segmentado, crecimiento constante y rentable, altos márgenes, bajas barreras de entrada, gran estructura heredada de grandes jugadores, etc.) presentan una oportunidad única para que surja un jugador de Marketplace en línea, y nos trae a la declaración de la visión de IMeureka: "Convertirnos en el Marketplace de referencia en el sector asegurador. Tecnología, transparencia, eficiencia y calidad de servicio, serán la base de cualquier relación de confianza en los negocios." [2]

Las deficiencias existentes en la intermediación de pólizas a menudo encarecen el precio de las primas de seguro manera innecesaria. Además, el bróker, debido a sus intereses económicos y facilitado por la opacidad del proceso, tiende a cerrar la operación con la compañía aseguradora que le genera mayor ingreso. Esto es malo para el cliente, ya que encarece la prima que paga, pero es tremendamente negativo para el mercado asegurador, que pierde el tiempo cotizando riesgos que nunca va a ganar ya que el bróker no lo va a permitir. La comisión del bróker es, en algunos casos de malas prácticas, arbitraria, orientado a conseguir "el máximo beneficio sin perder el cliente". Algunas aseguradoras incluso permiten encarecer de manera artificial las pólizas; al terminar una cotización, la aseguradora le indica al bróker un precio mínimo, y éste puede subirlo a lo que quiera para aumentar su comisión. Por otro lado, existen muchas pólizas que se van renovando año a año sin volverlas a cotizar. En muchos casos estas pólizas antiguas tienen sobreprecio debido a que las condiciones de los riesgos hace años no eran las mismas, ni la competitividad de cotización entre aseguradoras era tan agresiva como lo es actualmente.

Para solucionar esto, se creó el Marketplace de IMeureka, en el cual los usuarios (gestores de riesgos de las empresas y clientes particulares) pueden gestionar su programa de seguros (acceder a la información de sus riesgos digitalizada, comunicar siniestros y contratar pólizas, pedir cotización pólizas existentes) y las compañías aseguradoras pueden cotizar las pólizas que quieran (lanzar ofertas) sin que el bróker se interponga por intereses económicos.

Recientemente, la empresa ha abierto otra línea de negocio, y es que todas las ventajas competitivas que estaba adquiriendo a través de la innovación tecnológica en el sector las pondrá a disposición de las correduría actuales existentes del mercado español. De esta manera IMeureka abre otro frente, aparte de la correduría de seguros orientada simultáneamente al liderazgo en costes y a la diferenciación, inicia un nuevo modelo de negocio con el objetivo de transformarse en el "idealista de los seguros". La empresa pondrá a disposición de las corredurías actuales las herramientas que vayan desarrollando, especialmente las creadas en este ensayo, de tal manera que éstas no la vean como la competencia si no como todo lo contrario, un apoyo en su negocio.

Es común que las corredurías pequeñas no tengan presupuesto ni conocimientos suficientes para crear herramientas tecnológicas que les ayuden a aumentar su facturación. Por ello, IMeureka planea crear herramientas que puedan ser utilizadas bajo un modelo de suscripción.

Paralelamente al trabajo en este ensayo, la empresa está trabajando en temas de NLP, reconocimiento de imágenes, extracción de datos de fuentes externas, lectura e interpretación automática de PDF … Todo ello orientado a aportar valor a todos los clientes de la empresa, tanto clientes de la correduría como los clientes del Marketplace explicado anteriormente.

Es importante resaltar la diferencia entre IMeureka y un simple comparador de seguros, en palabras de IMeureka "No somos un comparador. Ni mucho menos. Los comparadores son herramientas que ofrecen tarifas pre acordadas con distintas compañías. Precios fijos. IMeureka es una herramienta de gestión de riesgos, un Marketplace donde las compañías tendrán acceso a tu información y podrán ir mejorando sus ofertas para tu beneficio. Precio dinámico." [2]

En resumen, la misión de IMeureka es generar valor para los clientes [2] proporcionando las herramientas indicadas para ello.

IMeureka ha cerrado ya acuerdos con decenas de aseguradoras, confirmando así el interés del nicho de mercado que quieren explotar. Algunas referentes del sector como Mapfre,

Generali, Plus Ultra, Axa, Helvetia , Cáser Seguros… La idea es que, al sacar un riesgo al Marketplace, sean todas las aseguradoras las que compitan (tanto en precio como en coberturas ofrecidas) entre ellas, en un entorno anónimo entre ellas para garantizar la transparencia.

En el contexto previamente explicado surge la idea de este ensayo en cuestión, las corredurías existentes no cuentan con los medios ni la tecnología para aprovechar los datos útiles que son generados en internet. Hay una gran cantidad de datos públicos que pueden ayudar a las corredurías de seguros a encontrar más negocio o a ofrecer un mejor servicio.

IMeureka se dio cuenta de la necesidad de emplear más datos en su negocio, muchos de ellos se podían obtener de manera automática y de fuentes públicas. Con ellos, podrán ofrecer un mejor servicio tanto a clientes finales como a aseguradoras.

Desde el punto de vista de negocio, sus principales objetivos con la extracción de datos son los siguientes:

1. Catastro [3]: Inventario de inmuebles. Todos deben estar obligatoriamente inscritos. Los clientes encontrarán en la web directamente los datos de los inmuebles a asegurar, en lugar de tener que buscarlos manualmente o responder a preguntas a través de emails, que sean públicos. Se debe facilitar a los usuarios la obtención de estos datos sin que los mismos entren a la web del catastro a buscarlos.
2. Licitaciones [4]: Sistema público de adjudicación de contratos de todo tipo (obras, servicios de mantenimiento, servicios informáticos ...) para todo tipo de organizaciones públicas (ayuntamientos, ministerios, consorcios...), esta plataforma es actualizada cada vez que se publica una nueva licitación. El objetivo es tener un sistema para ver las nuevas licitaciones que se van publicando de la categoría de seguros, directamente en la plataforma de IMeureka, para poder ser los primeros en cotizar la póliza con diferentes compañías.

3. Fichero de Multas [5]: El TEU (tablón edictal único) almacena los datos de multas para que puedan ser consultados por la ciudadanía. Este apéndice tratará de extraer esos datos si resultan ser útiles para la empresa.
4. Empresas españolas [6]: Datos de todas las empresas españolas posibles. Será una herramienta que ayude a los comerciales a buscar empresas con características determinadas, mejorando eficiencias y abriendo puertas de negocio.

El proyecto se abordará teniendo excepcional respeto por las leyes referentes a GDPR [7]. El Reglamento General de Protección de Datos de la Unión Europea se centra en la protección de los datos personales (información relativa a una persona física con la que se puede identificar, especial cuidado con recopilaciones de datos que pueden llevar a identificar a personas).

El sector asegurador se encontraba en un estancamiento tecnológico hasta mediados de la década de 2010, momento en el que el valor de las insurtech empezó a dispararse. En los cinco primeros meses de 2021, las inversiones en insurtech multiplicaron por 13 el total de inversiones en estas mismas empresas en 2016. [8] El grueso de estas inversiones ha ido a parar a Reino Unido, Alemania y Francia (85%), mientras que en España se ha invertido un modesto 2%.

El primer paso de Imeureka para ser una empresa con tecnología puntera pasa por la adquisición de datos que pueda emplear en el negocio. Al inicio de este proyecto, sólo contaban con los propios datos que su plataforma generaba. Escasos. Por ello, plantearon este proyecto de adquisición de datos públicos para integrarlos en la plataforma actual y abrir nuevas oportunidades de negocio.

Este proyecto tratará sobre la extracción de datos de diversas fuentes, y su integración con los sistemas preexistentes en IMeureka. A sí mismo, se crearán herramientas para que los usuarios designados puedan visualizar la información de manera eficiente.

# Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

Para el desarrollo de la aplicación se han usado cuatro tecnologías principalmente.

Selenium [9] - Framework de automatización gratuito y de código abierto. Permite controlar navegadores de manera automática y realizar tareas como el web scraping. Es muy potente ya que usa navegadores reales, y se encarga automáticamente de la gestión de elementos de navegación como cookies, timeouts … Es la manera más versátil de realizar extracción de datos, pero también es la más lenta y consume muchos recursos.

Se usó la librería "php-webdriver-Selenium WebDriver bindings for PHP" [10], que permite controlar navegadores web automatizados de Selenium desde PHP. La elección del lenguaje PHP fue para facilitar  la integración con los sistemas existentes en la empresa.

Postman [11] – Herramienta para simplificar el desarrollo e implementación de Apis. Con esta herramienta se pueden realizar peticiones http/https de todo tipo, fue usada el desarrollo para depurar la API y la conexión con el backend.

JSON [12] – Es un formato ligero de intercambio de datos. En la aplicación fue usado para la comunicación entre la BD y la aplicación.

Git & Bitbucket [13 , 14] – Git es un software de control de versiones. En la aplicación fue utilizada para tener versiones del código externas a la propia máquina en la que se desarrollaba. Bitbucket permite integrarse con Git para almacenar el código en un repositorio en línea

PHP Storm [15] – IDE para la programación de la API en PHP, agiliza el desarrollo.

XAMPP [16] – "XAMPP es un paquete de software libre, que consiste principalmente en el sistema de gestión de bases de datos MySQL, el servidor web Apache y los intérpretes para lenguajes de script PHP y Perl. El nombre es en realidad un acrónimo: X, Apache,

MariaDB/MySQL, PHP, Perl." XAMPP fue usado para la gestión de bases de datos MYSQL en IMeureka, y la ejecución del código en localhost.

# Capítulo 3. ESTADO DE LA CUESTIÓN

En los últimos datos los temas relacionados con la privacidad de los usuarios y la protección de los derechos de autor en internet han tomado mucha relevancia. El Reglamento General de Protección de Datos es el reglamento europeo que protege a personas físicas en el tratamiento de sus datos personales. Se aplicó el 25 de mayo de 2018 y desde entonces se ha realizado multitud de multas millonarias a empresas por no respetarla. Sin ir más lejos, España es el país en el que más multas se pusieron en el primer trimestre de 2021 por incumplir la GDPR [17], superando los 15 millones de Euros, y castigando a empresas como Vodafone o Caixabank.

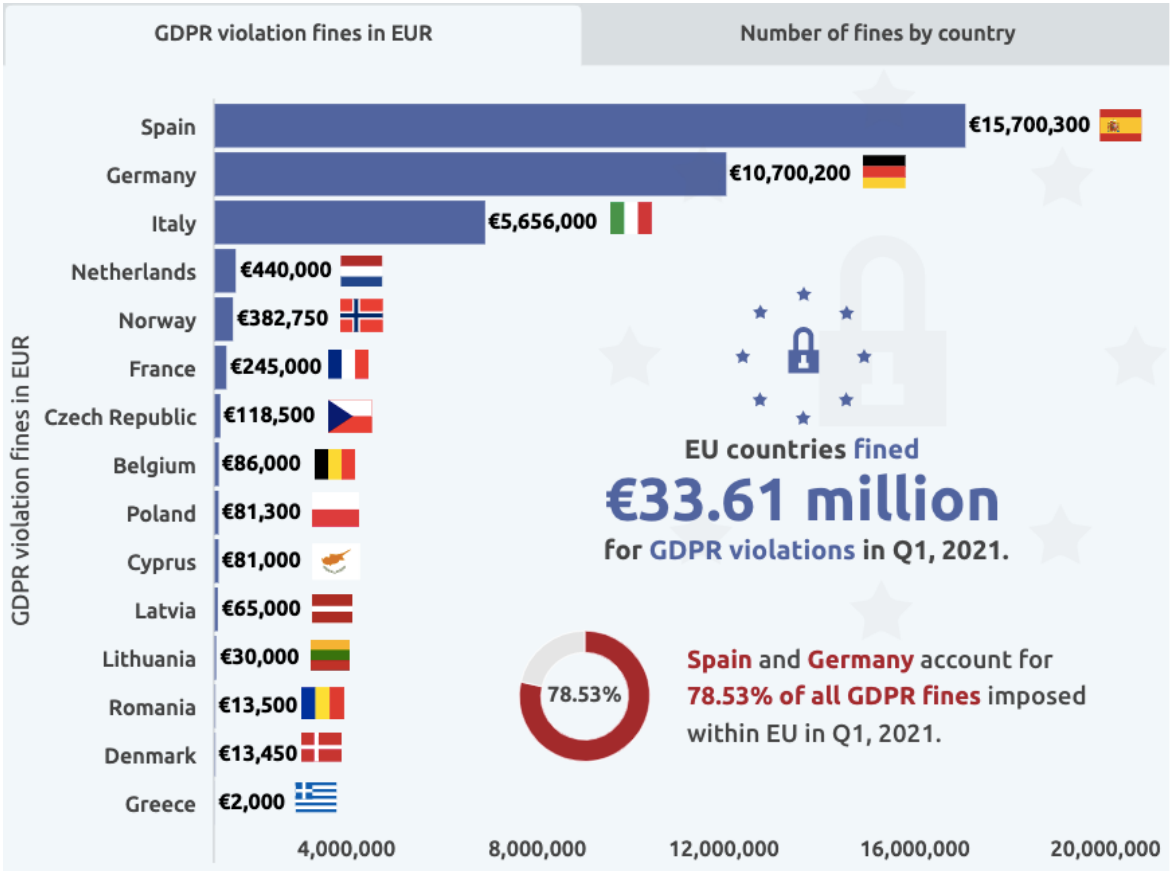



*Ilustración 2: Multas RGPD Q1 2021 en Europa [17]*

Esta es la motivación por la cual se hace imperativo realizar un breve estudio de legalidad antes de realizar cualquier extracción, IMeureka no tiene capacidad para asumir multas de estos calibres.

En esencia, RGPD prohíbe, salvo excepciones, el tratamiento de datos personales de residentes en la Unión Europea sin su consentimiento explícito. Esto supone que no se pueden trabajar con datos personales en tareas de web scraping, ya que no es factible conseguir permiso. Se consideran datos personales, entre otros, el nombre de una persona, la dirección, el email, el número de teléfono, sus datos bancarios, su dirección IP, su fecha de nacimiento, su número de seguridad social, su información médica… Es destacable que a lo largo del proyecto fue necesario leer en detalle la RGPD e incluso para la línea de investigación que se estaba siguiendo.

Por otro lado, en materia de propiedad intelectual también hay que extremar las precauciones. Las grandes empresas en las cuales sus datos son su mayor activo pueden y llevan a cabo denuncias. El caso de HiQ Labs vs LinkedIn es bastante famoso. La red social acusó a HiQ Labs por extraer datos de su plataforma sin consentimiento. Aunque la sentencia dictaminó que HiQ Labs tenía derecho a extraer los datos, cambió las normas de linkedin para siempre, que ya no deja sus datos a público.[19]

En resumen para este apartado, si este ensayo se hubiese realizado hace 5 años, es probable que esta sección de RGPD no hubiese sido necesaria (los datos de las personas no estaban tan protegidos), pero con la situación actual se hace imprescindible.

Por otro lado, se debe tener cuidado con el contenido de propiedad intelectual. Según el artículo 274 de nuestro código penal. “Será castigado con la pena de seis meses a dos años de prisión y multa de 12 a 24 meses el que, con fines industriales o comerciales, sin consentimiento del titular de un derecho de propiedad industrial registrado conforme a la legislación de marcas y con conocimiento del registro, reproduzca, imite, modifique o de cualquier otro modo utilice un signo distintivo idéntico o confundible con aquel, para distinguir los mismos o similares productos, servicios, actividades o establecimientos para los que el derecho de propiedad industrial se encuentre registrado.”[20]

Una vez que está claro que no es recomendable iniciar una extracción masiva de datos sin realizar un estudio previo de la legislación, podemos pasar al estudio del mercado asegurador.

El mercado asegurador es maduro y con alta competencia, goza de grandes economías de escala, por lo que diferenciarse es fundamental. En la siguiente imagen podemos ver el número de aseguradoras privadas en España por año:


*Ilustración 3: Número de empresas de seguros vs año*

El número de empresas aseguradoras lleva cayendo 10 años de manera consecutiva (salvo en 2017), por lo que podría pensarse que es un mercado que está muriendo. Nada más lejos de la realidad como demuestra la siguiente imagen:

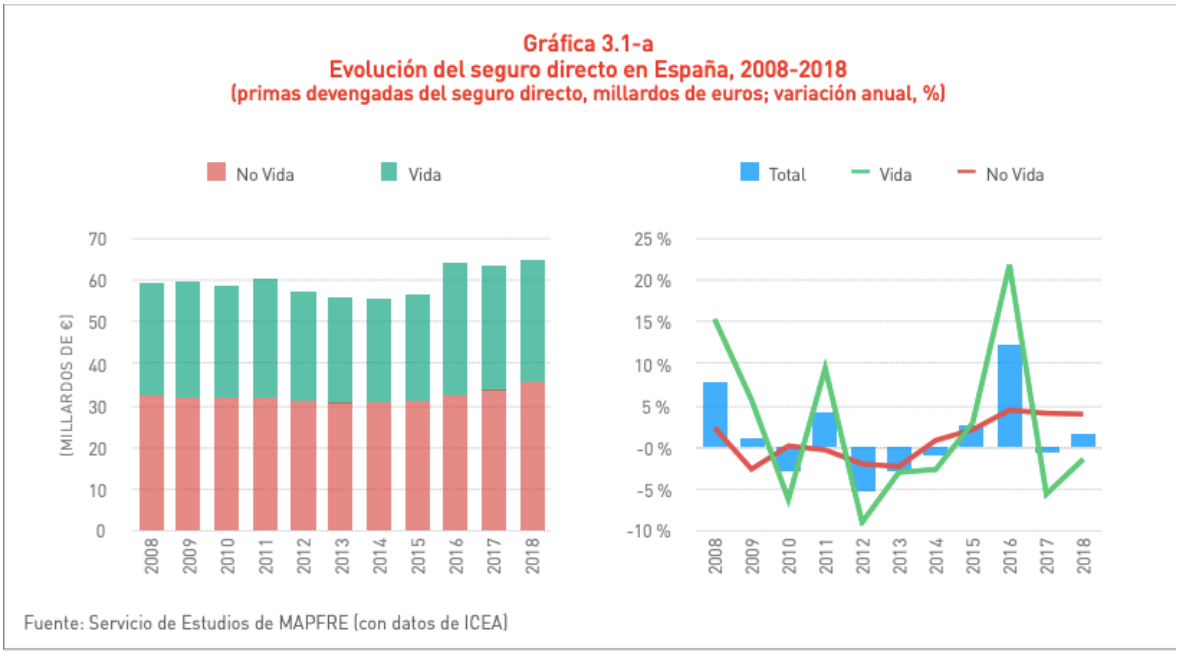



*Ilustración 4: Facturación seguros españoles[21]*

Podemos observar que, a pesar de que el volumen de las primas se mantiene estable e incluso crece en años recientes, la cantidad de aseguradoras disminuye. Cada vez, las aseguradoras grandes son más grandes, pueden ofrecer mejores precios y se van creando enormes barreras de entrada que expulsa a los más ineficientes o menos innovadores.

El índice de Herfindahl [22] es una medida empleada en economía que informa sobre la concentración de un mercado. Un índice alto >1000 supone que el mercado se reparte entre muy pocas empresas grandes. Un índice bajo lo contrario.

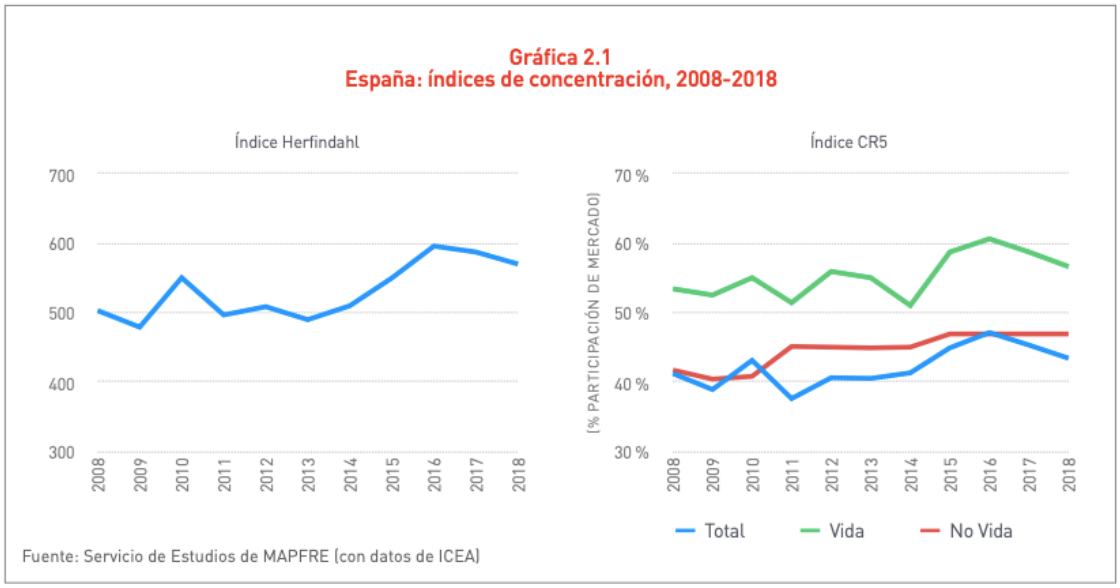



*Ilustración 5: El índice de Herfindahl del mercado asegurador [21]*

El índice está aumentando, y esto demuestra lo expuesto antes, que el mercado se está concentrando en cada vez menos manos.

Parece observarse una tendencia alcista en el índice de Herfindahl, lo que significa que el mercado se está volviendo cada vez menos competitivo (cuadra con lo visto de la disminución del número de empresas, cada vez las empresas que sobreviven tienen más cuota de mercado)

Por tanto, ante una industria madura y de decreciente competitividad (antes eran más competitivos, pero se está concentrando en pocas manos), se hace necesaria aportar valor diferencial más allá de la reducción en costes. Es por ello por lo que IMeureka decide llevar adelante este proyecto y muchos otros, las nuevas tecnologías serán su factor diferencial para entrar en el mercado de las corredurías.

Para la búsqueda de los datos, se siguió el procedimiento que se va a explicar a continuación:

El paso cero es comprobar si los datos requeridos ya existen en bases de datos públicas o privadas en internet, ya sean de pago o gratuitas. En caso de que sean de pago, habría que hacer un estudio para averiguar qué es lo más rentable para la empresa.

El primer paso es comprobar si los datos que se buscan existen ya a través de una API. En caso de que el organismo en cuestión los facilite, ésta será la manera más sencilla, ya que la propia web de la que se desea extraer estará poniendo de su parte

El segundo paso, en caso de que no exista API, es comprobar la legalidad de la extracción automatizada que se va a realizar. Es posible que la web en cuestión no se haya ocupado en hacer un API, o también que en sus términos y condiciones especifiquen que los datos no deben ser extraídos de manera automatizada (Derechos de autor)

El tercer paso es comprobar si se pueden interceptar las peticiones XHR. La tecnología más extendida de este tipo es AJAX, interceptarlo permite obtener los datos en formato semi-estructurado como XML o JSON

El cuarto paso es tratar de replicar las peticiones manualmente, requiere conocimientos sobre los protocolos de internet. Se replican las peticiones que realiza el navegador de manera automatizada y se extraen los datos de las repuesta

Finalmente, si todos los métodos anteriores fallan (salvo el punto 2, la legalidad es un facto bloqueante que si no se cumple no nos permite avanzar) se realizará navegación simulada con Selenium. Es una solución "todoterreno", ya que consiste en replicar las peticiones con un navegador automatizado. En contrapartida, consume más recursos y es más lento.

# Capítulo 4. DEFINICIÓN DEL TRABAJO

## *4.1 JUSTIFICACIÓN*

El proyecto se va a llevar a cabo debido a que se estima que existe un gran nicho de mercado tanto para clientes de IMeureka como para corredurías de seguros. La idea de es extraer estos datos para ponerlos al servicio de los corredores para que utilicen los paquetes que proporcione IMeureka. Este proyecto se plantea orientado a corredores pequeños y medianos, que no tienen ni los fondos ni el conocimiento tecnológico para implementar las soluciones, pero que bien pueden utilizar las mismas para aumentar su volumen de negocio. Ofreciéndoles más ventajas aparte de la diferenciación en costes se pretende conseguir que acepten este cambio.

Para las corredurías que utilicen los servicios de IMeureka, la ventaja es clara. Más tiempo para buscar negocio y menos tiempo "desaprovechado" en tareas de gestión y búsqueda en sitios web de diversas fuentes.

## *4.2 OBJETIVOS*

Este trabajo tendrá dos objetivos principales; el primero será la obtención del dato en sí, como se mencionó antes, las fuentes de las que se van a obtener los datos serán:

1. Catastro: Los datos del catastro deberán ser accesibles desde la web de IMeureka, para ello, se llevarán a cabo las implementaciones necesarias para que los datos del catastro estén disponibles en Near-Real Time a los usuarios de la plataforma. En el momento en que los usuarios quieran la información de un inmueble, se realizará una petición al catastro, se almacenará en DB y se le mostrará.
2. Licitaciones: El objetivo con las licitaciones será la creación de un robot (webscrapping) que diariamente visite la web de licitaciones del estado y publique en la plataforma las novedades. Para visualizar estos datos, se debe crear una herramienta transparente para el usuario que permita lanzar al mercado las licitaciones aprovechando la infraestructura existente en IMeureka. Además, se debe avisar a los corredores de que nuevas licitaciones han sido publicadas, para que no se olviden.
3. Fichero de multas: Los empleados de IMeureka podrán comprobar si algún usuario/cliente tiene multas públicas pendientes. Investigar estos datos para comprobar su utilidad.
4. Empresas españolas: Se pretende guardar en la base de Datos de IMeureka una lista de empresas españolas para facilitar la labor de los comerciales. Los comerciales deben poder realizar una búsqueda por diversos filtros. Además, serán capaces de descargarse sus resultados, exportándolos a Excel o algún formato similar, de tal manera que puedan guardar los mismos en local sin necesidad de repetir la búsqueda.
5. Enriquecer todo lo posible la base de datos de empresas

El segundo objetivo es que los desarrollos realizados queden completamente integrados dentro de la plataforma, tanto a nivel visual como funcional. Se desarrollarán funciones claras, con programación todo lo simple y estructurada posible, y se documentarán estas funciones para que el código sea reutilizable por otros programadores que no conozcan necesariamente las técnicas empleadas.

Los datos deben quedar incluidos en la plataforma de tal manera que puedan ser consideradas "herramientas de trabajo" para los distintos corredores. Especialmente la parte de empresas y de licitaciones, se pretenden vender como paquetes aparte. El catastro y el fichero de multas deben usarse de manera transparente, para que el usuario no tenga siquiera que saber que se están realizando peticiones ahí.

## *4.3 METODOLOGÍA*

En primer lugar, se comenzará por el aprendizaje del estado del arte actual de las corredurías de seguros, su antiguo modelo de negocio, las novedades en el modelo que plantea implementar IMeureka, el lenguaje propio de la intermediación de seguros ...

Tras la introducción a la empresa, se aprenderá el lenguaje requerido para desarrollar la programación. En este caso, el lenguaje PHP ya se conoce, se realizará un estudio de Selenium PHP-webdriver [5] ya que el backend de la empresa está codificado con php, y aunque las labores de extracción de datos el lenguaje más extendido sea Python, para facilitar su integración, se realizará en PHP.

La manera de afrontar la extracción de datos de sitios web son siguiendo los pasos explicados en el estado de la cuestión. Se seguirán estrictamente ese orden, significando que, para una determinada web (simplificación de lo expuesto en el estado de la cuestión):

0. Se investigará si estos datos se pueden descargar / comprar
1. Se intentarán extraer los datos de una API pública
2. Se investigará la legalidad de la extracción (ciertos datos pueden estar protegidos por derechos de autor)
3. Se intentarán extraer las peticiones XHR
4. Se intentarán replicar las peticiones HTTP/HTTPS
5. Se usarán técnicas de navegación simulada (selenium)

## *4.4 PLANIFICACIÓN Y ESTIMACIÓN ECONÓMICA*

La planificación temporal a lo largo del curso académico es la siguiente:

La planificación temporal para seguir será la siguiente:

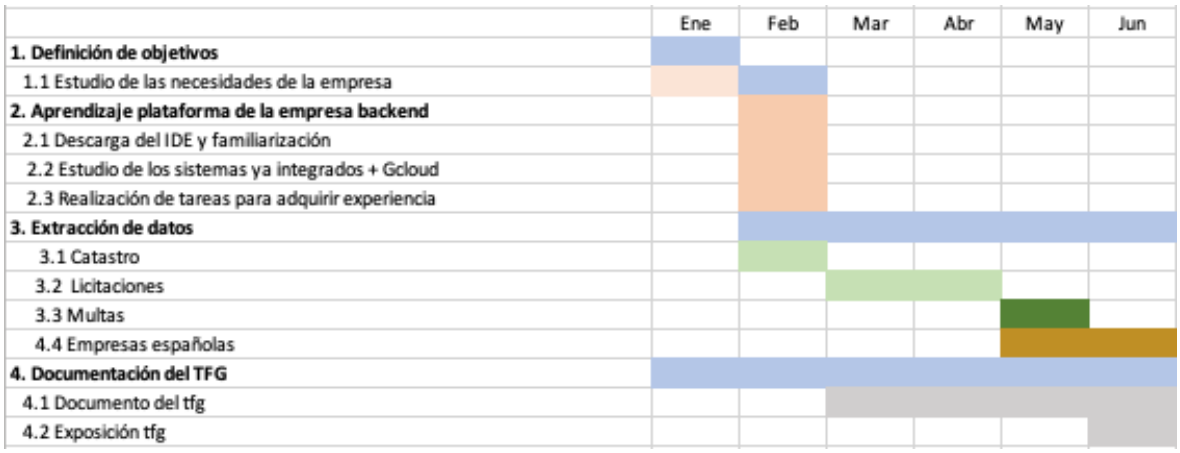



*Ilustración 6: Planificación del TFM*

Las licitaciones y las empresas serán lo que lleve más tiempo debido a que se pretende no sólo extraer los datos, si no crear herramientas versátiles que posteriormente puedan ser vendidas a corredores.

Por otro lado, se le debe dar una justificación económica a la extracción de datos y a las herramientas creadas.

Inicialmente las herramientas serán gratuitas, lo que se pretende es con unos primeros clientes probarla, para, una vez que esté puesta a punto y se consiga una buena cantidad de negocio a través de ella, empezar a monetizarla mediante un modelo de suscrición. Por otro lado, se pretende traer más corredores a la plataforma de Imeureka por:

- Incremento de clientes de la plataforma: Los datos extraídos ayudarán a simplificar la labor comercial (catastro, multas) y a conseguir más negocio (empresas, licitaciones). Con esto se pretende que sea un factor diferencial que convezca a clientes a que usen toda la funcionalidad de la paltaforma
- Incrementos de clientes del Marketplace: Abriendo la aplicación al Marketplace en general, cualquier correduría podrá utilizarla con sus propios clientes, lo que les ayudará a reducir trámites innecesarios y en consecuente a aumentar su rentabilidad. Las corredurías pagarán un canon por usar los servicios desarrollador por IMeureka, que aún está por determinar.

Los costes estimados para la empresa serán:

| Producto | Coste estimado |
|---|---|
| Becario 6 meses | 1800 € |

Ya que el resto de los costes (dominio / servidores) no son imputables a la aplicación en sí porque ya existían antes de ésta, e iban a seguir haciéndolo. La aplicación sólo se apoya en ellos, pero no le supone un costo extra.

Todos los desarrollos se rentabilizarán cuando los corredores empiecen a usar la plataforma de IMeureka. Para el acceso a las herramientas de empresas y de licitaciones, se pagará un canon mensual (estimado en torno a 100€ al mes). El objetivo de IMeureka es tener a 200 corredores a tres años vista, por lo que estas herramientas podrían facturar unos 200000€ al año para la empresa en tres años (sin apenas coste para la misma).

# Capítulo 5. SISTEMA/MODELO DESARROLLADO

## *5.1 EXTRACCIÓN DE DATOS*

En esta sección se va a detallar, para cada caso, el proceso de extracción de datos. Es decir, desde la ubicación de los datos en internet hasta el almacenamiento en la base de datos de IMeureka.

1. Catastro

El catastro dispone de una API [24] sin límite de peticiones y gratuita, por lo que no se han investigado si estos datos ya existen o son ofertados por alguna empresa.

La documentación de la API es pública y se encuentra en el siguiente enlace [24]


*Ilustración 7: Consultas en Postman*

En esencia, cuenta con 7 métodos a través de los cuáles se pueden acceder a todos los datos públicos del catastro (que en esencia son todos salvo la titularidad y el valor catastral)

- Consulta provincia : Este método devuelve las provincias de España.
- Consulta DNPPP: Datos No Protegidos de un inmueble según su Polígono y Parcela
- Consulta DNPLOC: Datos No Protegidos de un inmueble según su LOCalización
- Consulta Número : Consulta los números de inmuebles existentes en una vía
- Consulta DNPRC : Consulta de Datos No Protegidos según la referencia catastral

- Consulta Via : Servicio para consultar las vías de un municipio
- Consulta Municipio : Consulta los municipios de una provincia
- Consulta RCCOOR: Consulta de Referencia Catastral por COORdenadas

El primer paso para entender correctamente la API es implementar estos datos en Postman, ya que hay múltiples maneras de acceder a los mismos datos.

Por ejemplo, el método de consulta DNPLOC, tiene los siguientes datos de entrada:

| | KEY | VALUE |
|---|---|---|
| ☑ | Provincia | Sevilla |
| ☑ | Municipio | écija |
| ☑ | Sigla | CL |
| ☑ | Calle | merinos |
| ☑ | Numero | 30 |
| ☑ | Bloque | |
| ☑ | Escalera | |
| ☑ | Planta | |
| ☑ | Puerta | |

*Ilustración 8: Argumentos petición DNPLOC*

Con los datos que se especifican, se obtiene la respuesta en XML, en caso de que el inmueble sea único, se devuelven sus detalles (p ej: una casa).

```
<?xml version="1.0" encoding="utf-8"?>
<consulta_dnp xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.catastro.meh.es/">
    <control>
        <cudnp>1</cudnp>
        <cucons>2</cucons>
        <cucul>0</cucul>
    </control>
    <bico>
        <bi>
```

```
<idbi>
    <cn>UR</cn>
    <rc>
        <pc1>6375015</pc1>
        <pc2>UG1567N</pc2>
        <car>0001</car>
        <cc1>L</cc1>
        <cc2>I</cc2>
    </rc>
</idbi>
<dt>
    <loine>
        <cp>41</cp>
        <cm>39</cm>
    </loine>
    <cmc>39</cmc>
    <np>SEVILLA</np>
    <nm>ECIJA</nm>
    <locs>
        <lous>
            <lourb>
                <dir>
                    <cv>161</cv>
                    <tv>CL</tv>
                    <nv>MERINOS</nv>
                    <pnp>30</pnp>
                    <snp>0</snp>
                </dir>
                <loint>
                    <es>T</es>
                    <pt>OD</pt>
                    <pu>OS</pu>
                </loint>
                <dp>41400</dp>
                <dm>1</dm>
            </lourb>
        </lous>
    </locs>
</dt>
<ldt>CL MERINOS 30 41400 ECIJA (SEVILLA)</ldt>
<debi>
    <luso>Residencial</luso>
    <sfc>312</sfc>
    <cpt>100,000000</cpt>
    <ant>1989</ant>
</debi>
```

```
        </bi>
        <lcons>
            <cons>
                <lcd>VIVIENDA</lcd>
                <dt>
                    <lourb>
                        <loint />
                    </lourb>
                </dt>
                <dfcons>
                    <stl>232</stl>
                </dfcons>
            </cons>
            <cons>
                <lcd>ALMACEN</lcd>
                <dt>
                    <lourb>
                        <loint />
                    </lourb>
                </dt>
                <dfcons>
                    <stl>80</stl>
                </dfcons>
            </cons>
        </lcons>
    </bico>
</consulta_dnp>
```

En caso de que al realizar la petición de localización haya múltiples inmuebles (unidades habitables, pej un bloque), se devuelven los datos resumidos. En esencia, por inmueble se devuelve su dirección completa (piso puerta escalera… etc) y su referencia catastral. Para acceder a los datos habrá que realizar una petición concreta sobre esa referencia catastral. (Ver ANEXO : Petición con múltiples inmuebles).

Y en general el resto de los métodos la implementación es similar, se pueden ver en el siguiente enlace [25].

La interpretación (de ahora en adelante parseo) de los datos de estos xml de respuesta fueron bastante más complejos de lo que debieran debido a su enorme profundidad. Y es que en

ocasiones llegan a los 7 niveles. Por otro lado, no hay enlace a la documentación de la api en la web de la misma. Por lo que es complicado saber que existe.

Se encontró una implementación en Python de una API al catastro, pero debido a las contrains de tecnologías (en aquel momento en la empresa sólo se trabajaba con PHP, montar un servicio en Python suponía un gran trabajo de infraestructura), no se usó.

https://pycatastro.readthedocs.io/en/latest/api_reference.html

Posteriormente, se procedió a su implementación para la plataforma de IMeureka. Cuando un cliente lo necesite, simplemente escribirá su dirección en un buscador de google Mapas y éste extraerá la información del catastro. En caso de que la dirección no sea exacta, se mostrarán todas las unidades habitables que pudieran existir.

Cada edificio buscado en el catastro quedará almacenado en la base de datos de IMeureka, de tal manera que quede registro: ANEXO – DATOS CATASTRO

Los datos almacenados son:

- Referencia catastral: identificador único de la web del catastro
- Uso: A qué se destina esa unidad (Residencial, Comercial, Oficinas, Almacén…)
- Tamaño: Metros cuadrados de la unidad
- Antigüedad: Fecha de construcción
- Tipo: Tipo de unidad (vivienda, comercio…)
- Municipio: Municipio en el que se encuentra la unidad
- Provincia
- Código municipio: Codigo interno catastro para el municipio
- Código Provincia: 2 primeras cifras del código postal
- Código Postal
- Tipo vía: Tipo de vía en el que se encuentra (calle,carretera…) Almacenado como ID referenciada a la tabla master_tipo_via
- Kilómetro

- Bloque, planta, escalera, puerta, vía, número: Todo lo necesario para identificar inequívocamente una unidad
- Latitud, Longitud: Coordenadas
- Información Extra

Los métodos anteriormente mostrados fueron implementados en PHP:

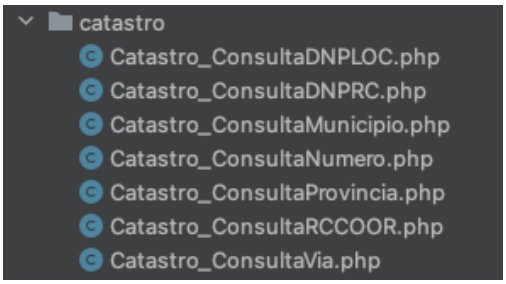



*Ilustración 9: Métodos implementados en PHPStorm*

La implementación se hizo usando los estándares de programación ya utilizados en la empresa, se explica el código como comentarios:

```
//MasterWs incluye funciones para realizar peticiones via Curl.
//Catastro
      class Catastro_ConsultaDNPLOC extends MasterWs {
/*
   * Data(parámetro de entrada a la función request) deberá contener los
      siguientes campos
   *
   * array(
      'Provincia' => 'sevilla',
      'Municipio' => 'écija',
      'Sigla' => 'CL',
      'Calle' => 'merinos',
      'Numero' => '30',
      'Bloque' => '',
      'Escalera' => '',
      'Planta' => '',
      'Puerta' => ''
      )
  Los 4 primeros obligatoriamente rellenos
```

```
  Consulta ejemplo:
   $prov = new Catastro_ConsultaDNPLOC($data);

*/
       //En este caso, el link de prod y el de dev es el mismo, pero se pone por
convención con otros métodos
    public $link_dev    =
"http://ovc.catastro.meh.es/ovcservweb/ovcswlocalizacionrc/ovccallejero.asmx/Cons
ulta_DNPLOC";

    public $link_prod  =
"http://ovc.catastro.meh.es/ovcservweb/ovcswlocalizacionrc/ovccallejero.asmx/Cons
ulta_DNPLOC";

       //La petición al catastro usará el método POST
    public $method     = "POST";

    //método que crea una request
       public function request($data) {
         return array("post_data" => http_build_query($data), "http_headers" =>
array('Content-Type: application/x-www-form-urlencoded'));
    }

       //Método que parsea una respuesta
    function parseResponse($response) {

         $result =
json_decode(json_encode(simplexml_load_string($response)),true);
         return $result;
    }


}
```

El código anterior extiende a la clase masterWS.

Para realizar una petición, simplemente crearemos la clase de la siguiente forma:

```
$inmueble = new Catastro_ConsultaDNPLOC($this->data);
```

El resultado de inmueble es un array con los mismos campos que el XML devuelto por la API del catastro.

El resto de los métodos (visibles en la ilustración X), fueron implementados de manera similar. Una vez que se tienen todos, el siguiente paso es crear un método que aglutine todas las posibilidades de extraer información del catastro (ya sea mediante latitud / longitud, referencia catastral, o los datos de localización explicados anteriormente)

Finalmente, se desarrolló el método consulta_catastro que hace un best_effort de todas las combinaciones de datos que haya introducido el usuario (o que se hayan autogenerado, por ejemplo, si el usuario introduce una dirección con múltiples viviendas). En caso de no conseguirlo se le informa al usuario. El método para consultar al catastro hace de wrapper de los métodos originales de la API del catastro.

```
function Consulta_Catastro(){
    //$f3 es el framework usado por la empresa
    $f3 = F3::instance();
    $_REQUEST = $f3->clean($_REQUEST);
    $this->result = array();
    //Se extraen los datos de la request
    //Se puede observar que contienen info para LOC, COOR y RefCat
    $this->data = array(
        'Provincia'     => isset($_REQUEST['Provincia'])? $_REQUEST['Provincia']
: '',
        'Municipio'     => isset($_REQUEST['Municipio'])? $_REQUEST['Municipio']
: '',
        'Sigla'         => isset($_REQUEST['Sigla'])    ? $_REQUEST['Sigla'] :
'',
        'Calle'         => isset($_REQUEST['Calle'])    ? $_REQUEST['Calle'] :
'',
        'Numero'        => isset($_REQUEST['Numero'])   ? $_REQUEST['Numero'] :
'',
        'Bloque'        => isset($_REQUEST['Bloque'])   ? $_REQUEST['Bloque'] :
'',
        'Escalera'      => isset($_REQUEST['Escalera']) ? $_REQUEST['Escalera'] :
'',
        'Planta'        => isset($_REQUEST['Planta'])   ? $_REQUEST['Planta'] :
'',
        'Puerta'        => isset($_REQUEST['Puerta'])   ? $_REQUEST['Puerta'] :
'',
        'RC'            => isset($_REQUEST['RC'])       ? $_REQUEST['RC'] : '',
        'Coordenada_Y'  => isset($_REQUEST['Latitude']) ? $_REQUEST['Latitude'] :
'',
        'Coordenada_X'  => isset($_REQUEST['Longitude'])? $_REQUEST['Longitude']
: '',
    );
    //Primero, se busca por localización (los datos de prov, municipio, calle...)
    //En caso de que el resultado sea más de uno, se realiza una petición por
    //cada referencia catastral para extraer los detalles
```

```
    $Consulta_direccion = $this->Consulta_direccion();

    //En caso de no extraer nada, se intenta a través de coordenadas
    if($Consulta_direccion["status"]=="ERROR"){
        //Si se entra aquí, es porque no se ha podido conseguir el inmueble por
DNPLOC, se intentará por latitud longitud
        $Consulta_coordenadas = $this->Consulta_coordenadas();
        if($Consulta_coordenadas["status"]=="ERROR"){
            //Si no se ha obtenido el inmueble por lat long, probamos por RC si
es que no está empty
            if(isset( $_REQUEST['RC']) && !empty( $_REQUEST['RC'])){
                $consulta_por_RC = $this->Consulta_por_RC($_REQUEST['RC']);
            }
        }
    }
    //Si no hay resultados, se le informa al usuario de que no se han encontrado
    if(empty($this->result)){
        $this->json(array("status"  => "ERROR"
                        , "message" => "No se han encontrado resultados"
                        , "result"  => false));
    }else{
    //En caso contrario se les muestran
        $this->json(array("status"  => "OK"
                        , "message" => "Se han encontrado resultados"
                        , "result"  => $this->result));
    }
}
```

En el siguiente apartado se explicará la UI que se desarrolló para que los usuarios usen este código.

2. Licitaciones

La investigación respecto a los contratos públicos del estado se inició mediante la búsqueda de herramientas pre-existentes que volcasen datos sobre la plataforma de IMeureka. En este caso, sí hay empresas dedicadas exclusivamente a la gestión de licitaciones.

Estas empresas no se ajustan a nuestro caso de uso ya que se dedican o a aportar buscadores más eficientes (la web pública de licitaciones funciona bastante mal y es incómoda) o a realizar alertas por email.


*Ilustración 9: Servicios de infoconcurso [26]*

En la imagen anterior vemos que ofrecen precios baratos y alertas personalizables, pero no los datos en crudo actualizados instantáneamente, que es lo que se busca. Además, muchas de estas plataformas cobran arbitrariamente por sus filtros (filtrado por 3 provincias un precio, filtrado por 10 otro superior…) cuando el coste es exactamente el mismo.

Al no encontrar un servicio de pago que cumpla con nuestras necesidades, iniciamos el proceso descrito en el estado de la cuestión.

Para las licitaciones (contrataciones del Estado) se realizó el proceso de búsqueda descrito en el estado de la cuestión. Tras constatar que no existe API ni pública ni de pago, se procede a investigar la legalidad de la automatización.

Se decide seguir adelante con el proyecto porque los datos no son personales y ni en su sección de aviso legal ni en protección de datos se prohíbe ni explícita ni implícitamente la extracción. El siguiente paso es comprobar si la web usa XHR, tras la negativa, queda intentar replicar las peticiones de manera directa a través de Postman. Resulta que esta web utiliza ciertas cookies para las redirecciones, de tal manera que, al copiar y pegar un enlace en el buscador de google, no nos lleva siempre a la misma página (en esencia, esto significa que aparte del recurso son necesarias ciertas cookies que es más sencillo obtenerlas mediante la navegación simulada). Además, se cuelga frecuentemente. Llegamos al último caso, es necesario la navegación simulada.

Para realizar el proceso de navegación simulada, primero es necesario navegar manualmente por la página para entender el funcionamiento de ésta (además de estudiar cómo se distribuyen las etiquetas HTML de la página, ya que es clave para la implementación de Selenium).

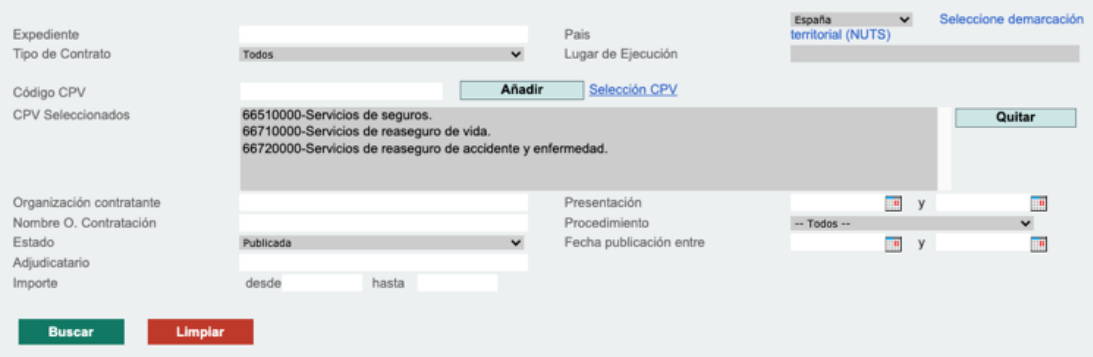



*Ilustración 10: Petición que se desea replicar [27]*

Tras realizar la búsqueda, obtenemos la siguiente tabla, ordenadas por defecto por fecha, desde la más reciente a la más antigua. Ésta navegación será la necesaria para obtener las novedades que vayan surgiendo en la plataforma:

| Expediente | Tipo de Contrato | Estado | Importe | Presentación | Órgano de Contratación |
|---|---|---|---|---|---|
| @2021011SER<br>Seguro por riesgos de responsabilidad civil del Ayuntamiento de Las Rozas de Madrid | Servicios<br>Servicios financieros: a) servicios de seguros; b) servicios bancarios y de inversión | Publicada | 65.000,00 | 15/07/2021 | Junta de Gobierno del Ayuntamiento Las Rozas |
| CONTR 2021 0000277539<br>suscripcion póliza seguros asist.sanit. y repatriación auxiliares conversac | Privado | Publicada | 212.268,00 | 08/06/2021 | Consejería de Educación y Deporte. Secretaría General Técnica |
| 194/2021<br>Prestació del servei d'assegurança mèdica quin col·lectiu beneficiat són el personal laboral i funcionarial de l'Ajuntament d'Alcarràs, així com els representants polítics al Ple municipal. La classificació CPV (Vocabulari Comú de Contractes públics de la Comissió Europea) nomenclatura principal és la següent: 66512220-0 Serveis d'assegurances mèdiques. | Servicios | Publicada | 211.560,00 | 15/07/2021 | Ajuntament d'Alcarràs |

*Ilustración 11: Respuesta del buscador de licitaciones [27]*

Se da la particularidad de que también hay archivos en las licitaciones, por lo que se aprovecha el google cloud Storage (BD noSQL ya usada previamente por la empresa) y se descargan. Es necesario, para cada fila de la tabla, entrar a los dos enlaces para almacenar por un lado la información de los detalles de la licitación, y por otro la organización. En base de datos la información anterior queda reflejada en las siguientes tablas:

Tabla Licitaciones:

- ID_licitaciones
- Codigo_expediente
- ID_órgano_contratación
- ID_master_tipo_contrato_licitacion
- Estado_licitacion
- Objeto_contrato
- Presupuesto_base_licitacion_sin_impuestos
- Valor_estimado_contrato
- Codigo_CPV
- Lugar_de_ejecucion
- Procedimiento_de_contratacion
- Fecha_fin_presentacion_oferta
- ID_producto
- URL
- ID_broker
- Enviado
- Fec_intro

La tabla órgano de contratación consta de los siguientes campos:

- ID_organo_contratacion
- Organo_contratacion
- Organización_contratante
- NIF
- Idioma
- Direccion_site_organo
- URL
- Actividad
- Via,CP,Poblacion,Pais
- Tlf,Fax;email

Para esta parte se creó un método llamado get_nuevas_licitaciones, el cual, mediante la API para PHP de selenium, extrae las novedades.

```
public function get_nuevas_licitaciones(){
    set_time_limit(0);
    try {
        //Developed by Miguel Enrile
        //Esta función obtiene las licitaciones nuevas que salen a mercado, y
extrae los datos del expediente y del organismo de contratación
        //Es un programa de selenium con php web-driver
        //doc : https://github.com/php-webdriver/php-webdriver

        //Objeto licitaciones creado para la interacción con la base de datos
        $Lic = new Licitaciones();
        //Creación de un navegador simulado, se almacena en la variable $driver

        $driver = $this->createDriver();
        //1. Extracción de los expedientes que ya existen en DB, de esta manera,
no se busca 2 veces un expediente que ya existe.
        $licitaciones_db_completas = $Lic->getAll("ID_licitaciones", 'asc');
        $licitaciones_db = array();

        //Objeto órgano de contratación creado para la interacción con la base de
datos
        $org = new Organo_contratacion();
        //Se modifica el array para que concuerde con lo extraído de la tabla
licitaciones en la web del estado
        foreach ($licitaciones_db_completas as $completa) {
            $nombre = $org-
>getById($completa["ID_organo_contratacion"])[0]["Organo_contratacion"];

            $aux = array("Codigo_expediente" => $completa["Codigo_expediente"],
"Organo_contratacion" => $nombre);
            array_push($licitaciones_db, $aux);
```

```
        }
        //2. Se ejecuta un proceso que extrae todos los datos de las licitaciones
activas
        /*
         * ACCESO FORMULARIO LICITACIONES
         */
        $driver->get("https://contrataciondelestado.es/wps/portal/licitaciones");
        //Este es uno de los motivos por los que se usa selenium, sólo se puede
acceder a la raíz de las licitaciones
        //(el link de arriba). Por ello, cada vez es necesario encontrar el
elemento que nos lleva al buscador y pulsarlo.
        //Afortunadamente, el nombre de los elementos no cambia.
        $licitaciones = $driver-
>findElement(WebDriverBy::id("viewns_Z7_AVEQAI930OBRD02JPMTPG21004_:form1:logoFor
mularioBusqueda"));
        $licitaciones->click();
        /*
         * RELLENO FORMULARIO LICITACIONES, aquí se buscan las licitaciones a
alto nivel por CPV
         */
        try {
            //Select pais
            $test = $driver-
>findElement(WebDriverBy::cssSelector('select[id="viewns_Z7_AVEQAI930OBRD02JPMTPG
21004_:form1:menu1MAQ1"] option[value="ES"]'));
            $test->click();
            //CÓDIGOS CPV : Se añaden al buscador los códigos CPV
            foreach ($this->codigos_CPV as $cod) {
                $CPV = $driver-
>findElement(WebDriverBy::id("viewns_Z7_AVEQAI930OBRD02JPMTPG21004_:form1:cpvMult
iple:codigoCpv"));

                $CPV->sendKeys($cod);
                $ButtonAddCPV = $driver-
>findElement(WebDriverBy::id("viewns_Z7_AVEQAI930OBRD02JPMTPG21004_:form1:cpvMult
iplebuttonAnyadirMultiple"));
                $ButtonAddCPV->click();
            }
            $test = $driver-
>findElement(WebDriverBy::cssSelector('select[id="viewns_Z7_AVEQAI930OBRD02JPMTPG
21004_:form1:estadoLici"] option[value="PUB"]'));
            $test->click();

        } catch (\Exception $e) {
            Logging::write($e->getTraceAsString(),'error_licitaciones');
            echo $e;
            $driver->close();
            $driver->quit();
            die();
        }
        //Se realiza la búsqueda
        $Buscar = $driver-
>findElement(WebDriverBy::id("viewns_Z7_AVEQAI930OBRD02JPMTPG21004_:form1:button1
```

```
"));
        $Buscar->click();
        //Se obtiene el número de páginas, que será usado para la iteración en un
for loop de cada una de ellas y la
        //extracción al completo de la tabla de resultados.
        $NumPaginas = $driver-
>findElement(WebDriverBy::id("viewns_Z7_AVEQAI930OBRD02JPMTPG21004_:form1:textfoo
terInfoTotalPaginaMAQ"));
        $paginas = intval($NumPaginas->gettext());
        $expedientes_completo = array();
        //Se itera entre las páginas
        //Con esta combinación de índices, se obtienen los datos clave de un
expediente, su id y su org contratacion
        for ($i = 1; $i < ($paginas + 1); $i++) {
            if ($i > 1) {
                $Siguiente = $driver-
>findElement(WebDriverBy::id("viewns_Z7_AVEQAI930OBRD02JPMTPG21004_:form1:footerS
iguiente"));
                $Siguiente->click();
            }
            //extractAllBasicDataLicitaciones extrae los datos de la tabla de
licitaciones
            $expedientes_completo = array_merge($expedientes_completo, $this-
>extractAllBasicDataLicitaciones($driver));
        }

        sleep(1);
        $driver->close();
        $driver->quit();

        //3.Se buscan las diferencias entre el array extraído de BD y el extraído
de la web
        // Si no hay diferencias, significa que no han habido actualizaciones,
get_diferencias_array
        //devolverá vacío.
        // Lo común es que la diferencia sea de 1 o 2 licitaciones, salvo en la
primera ejecución
        $expedientes_a_extraer = $this-
>get_diferencias_array($expedientes_completo, $licitaciones_db);
        //Actualización organismos de contratación
        //bucle, si no existe lo inserta en organismo de contratación
        foreach ($expedientes_a_extraer as $expediente) {
            //Aunque la licitación sea nueva, se buscan los organismos de
contratación por si ya existiesen
            //en base de datos.
            $org = new Organo_contratacion();
            //Si extraido = 0 o si no existe ...
            $organo = $org->getByColumn("Organo_contratacion",
$expediente["Organo_contratacion"]);
            if ($organo[0]["Extraido"] == 0 || empty($organo)) {
                //Se realiza una primera inserción, de la info básica, si no
existía antes en db
                if(isset($organo[0]["ID_organo_contratacion"])){
```

```
                    //Si ya existe en db, se busca el id
                    $id_asig = $organo[0]["ID_organo_contratacion"];
                }else{
                    $id_asig = $org->insertArray(array("Organo_contratacion" =>
$expediente["Organo_contratacion"]));
                }
                //En este momento, también extraemos la información en detalle
del organismo de contratación
                $this->extract_details_organoContratacion($id_asig,
$expediente["Organo_contratacion"]);
            }
        }
        //PARA SACARLOS COMO JSON POR PANTALLA
        foreach ($expedientes_a_extraer as $expediente_basico) {
            //Extracción de la info completa, rellena la parte de licitación de
DB
            $Lic = $this-
>extractAllCompleteDataLicitaciones($expediente_basico['Codigo_expediente'],
$expediente_basico['Organo_contratacion']);

            if(!empty($Lic)) {
                //Extracción de los archivos de las licitaciones
                $this->extractFilesFromUrlLicitacion($Lic['link'],
$Lic['id_asignado']);
            }
        }
    }catch (\Exception $e){
        //Imprime excepciones en el sistema de login de la empresa
        Logging::write($e->getTraceAsString(),'error_licitaciones');
        echo $e;
    }
}
```

Este servicio debe estar siempre levantado, y se parametrizará el tiempo de refresco. Empíricamente hemos visto que de media surgen unas 4 licitaciones al día de los CPV que nos interesan. Tiene sentido realizar actualizaciones periódicas, pero sin exceso. A final se decidió por una búsqueda a la hora.

```
$cron->set('cronLicitaciones'
,'com\controllers\LicitacionesController->get_nuevas_licitaciones'
,'5 * * * *');      //Cada hora, a las y 5
```

Cada hora a las y 5 se realiza la búsqueda. Se estima un buen compromiso entre tiempo de refresco y no saturación del servidor de la empresa.

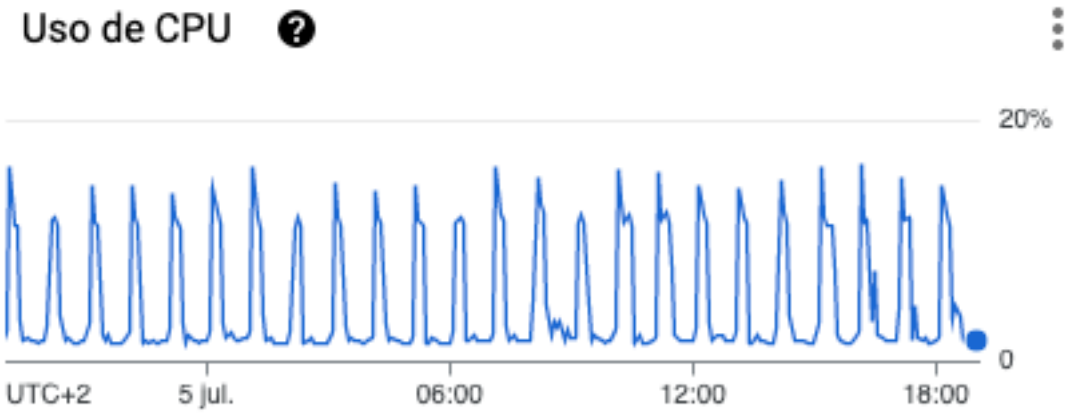



*Ilustración 12: Uso del CPU del servidor – Periodo de 24h [27]*

El servidor que aloja el extractor de licitaciones también aloja la web. Podemos observar 24 picos, coincidiendo con las 24 ejecuciones programadas diarias. Se incluye esta imagen para demostrar el gran consumo de recursos que supone Selenium, argumentando los motivos por los cuáles esta opción debe ser la última.

El proceso anterior se lleva ejecutando meses, con los siguientes resultados:

469 licitaciones & órganos de contratación

Si desea observar un ejemplo de los datos de licitaciones almacenados en base de datos, diríjase al anexo de licitaciones situado al final del documento.

Por últimos, las licitaciones son asignadas a corredores de IMeureka en función de

Las licitaciones, en función a su lugar de origen, son añadidas a diferentes personas.

```
if(in_array($_POST['Cliente']['ID_prov'], array(4, 14, 18, 23, 29, 6, 10, 11, 21,
41))){
    $gestorIM = 150; // Almeria, Cordoba, Granada, Jaen, Malaga, Badajoz,
Caceres, Cadiz, Huelva, Sevilla a Sara
    $mail_gestor = email_gestor@imeureka.com';
}/*elseif(in_array($_POST['Cliente']['ID_prov'], array(6, 10, 11, 21, 41))){
    $gestorIM = 125; // Badajoz, Caceres, Cadiz, Huelva, Sevilla a Carlos
    $mail_gestor = email_gestor@imeureka.com';
}*/elseif(in_array($_POST['Cliente']['ID_prov'], array(4, 14, 18, 23, 29))){
    $gestorIM = 788; // Alava, Guipúzcoa, La Rioja, Navarra, Vizcaya a Iker
    $mail_gestor = 'email_gestor@imeureka.com';
}elseif(in_array($_POST['Cliente']['ID_prov'], array(22, 44, 50, 28, 7, 46, 5, 9,
24, 34, 37, 40, 42, 47, 49))){
    $gestorIM = 469; // Huesca, Teruel, Zaragoza, Madrid, Baleares, Valencia,
Avila, Burgos, Leon, Palencia, Salamanca, Segovia, Soria, Valladolid, Zamora a
Jorge
    $mail_gestor = email_gestor@imeureka.com';
}else{
    $gestorIM = 176; // El resto a Pedro
    $mail_gestor = email_gestor@imeureka.com';
}
```

3. Empresas

Para la extracción de los datos de las empresas se llevó a cabo una búsqueda exhaustiva de la información, tanto de pago como gratuita, existente en internet. En este caso, sí se venden bases de datos como tal, por lo que había que hacer una comparativa entre precios y los datos que incluían.

La extracción de datos mediante las técnicas ya utilizadas anteriormente para el caso del ranking de empresas es de dudosa legalidad, debido a que en la mayoría de webs explícitamente se indica que los datos tiene copyright y que no se pueden extraer de manera automatizada sin el consentimiento explícito.

Por tanto, se decide buscar otros métodos de conseguir la información.

Empresa 1. elInforma

elInforma es una web que almacena información sobre empresas, el problema es que protegen estos datos con recelo ya que son su modelo de negocio. Es sin duda la base de datos de empresas españolas más completa de Internet.


*Ilustración 13: Oferta de elInforma*

elInforma extrae datos de lugares públicos como el BOE y el borne y crea sus listados e informes. Se pidió precio por el listado completo. El precio que dan a cada dato concreto es de 0.05€. Existiendo 1 millón de empresas y 20 datos por empresa, esto nos sale aproximadamente 1 millón de euros. Hablando por correo con ellos nos hicieron oferta por su base de datos al precio de 300.000€. Por desgracia, no tenemos presupuesto (ni lo vale, ya que se puede obtener muchísimo más barato).

Habría dos maneras, la primera sería intentar copiar a elInforma en su labor, ya que los datos que obtenga deben ser públicos, aunque esto parece muy complicado (las fuentes públicas pueden ser pdf's del BORME, anuncios del BOE…).

Hay muchas más bases de datos en venta aparte de elInforma, se detalla a continuación el estudio realizado.

Empresa 2. Emailsgo [29]

La empresa emailsgo, tras pedirle información sobre su base de datos, nos informaron que contaban con los siguientes datos

"Cada registro incluye los siguientes datos: Actividad, Nombre Empresa, Dirección, Población, Provincia, Código Postal, Teléfono1, y en algunos casos Teléfono2 y Fax."

El precio de su base de datos es de 650€. Al final no se decidió por esta base de datos debido a que no nos dio buena impresión (una landing muy anticuada para una empresa tecnológica, informales por email...).

Empresa 3. Bigdatalowcost.com [30]

La empresa bigadtalowcost nos ofrecía más de 1.200.000 compañías con 800.000 direcciones de email. Los campos eran nombre de la compañía, email, dirección, localización, código postal, teléfono, web, actividad …

El precio de la base de datos es de 150€, de las opciones vistas, es la que parece a priori mejor en cuanto a calidad precio.

Empresa 4. centraldecomunicacion.com [31]

Esta empresa ofrecía una base de datos muy barata (57€) pero no incluía en CNAE (información de la actividad) que es un dato clave que se busca, por lo que fue descartada.

Empresa 5. Infocif [32]

Tras buscar por la web de Infocif, no se encontró ningún punto en el que vendiesen sus datos, por lo que se decidió contactar con ellos por teléfono, ya que se observó que su base de datos era muy completa.

Cabe destacar que en sus términos y condiciones no prohíben explícitamente el web scraping, aunque sí la reproducción no autorizada.

Tras hablar con Infocif, nos informaron por email acerca de su política con sus datos. Nos informaron de que disponen de un buscador del que se pueden descargar los datos, limitados a 50 empresas cada vez. Al preguntarle explícitamente por permiso para utilizar esos datos, simplemente nos informaron de que, en caso de realizar demasiadas peticiones, nuestra IP sería bloqueada, pero no emprenderían más acciones ya que no era su plan.

Al preguntarles explícitamente sobre la disponibilidad del buscador para descargar sus empresas, ésta fue su respuesta:

“En estos momentos, estamos terminando de desarrollarlo, por lo que la exportación de resultados está limitada a 50 empresas, no obstante, próximamente estará abierto completamente.”

Con lo que entendemos que Infocif no vende sus datos, si no que tienen otro modelo de negocio, más parecido a “red social de empresas”. Por lo que podemos extraer sus datos de manera gratuita.

Empresa 6. Datacertia [33]

El precio de la base de datos era de 259€, con lo que excedía el presupuesto. Cabe destacar que la empresa nos pareció bastante fiable.

Empresa 7. Databusiness [34]

Esta empresa sólo contaba con una base de datos de 350k empresas, que no eran suficientes para nuestros objetivos.

Empresa 8. Publiemail [35]

Base de datos pequeña y cara

Empresa 9. BoldData [36]

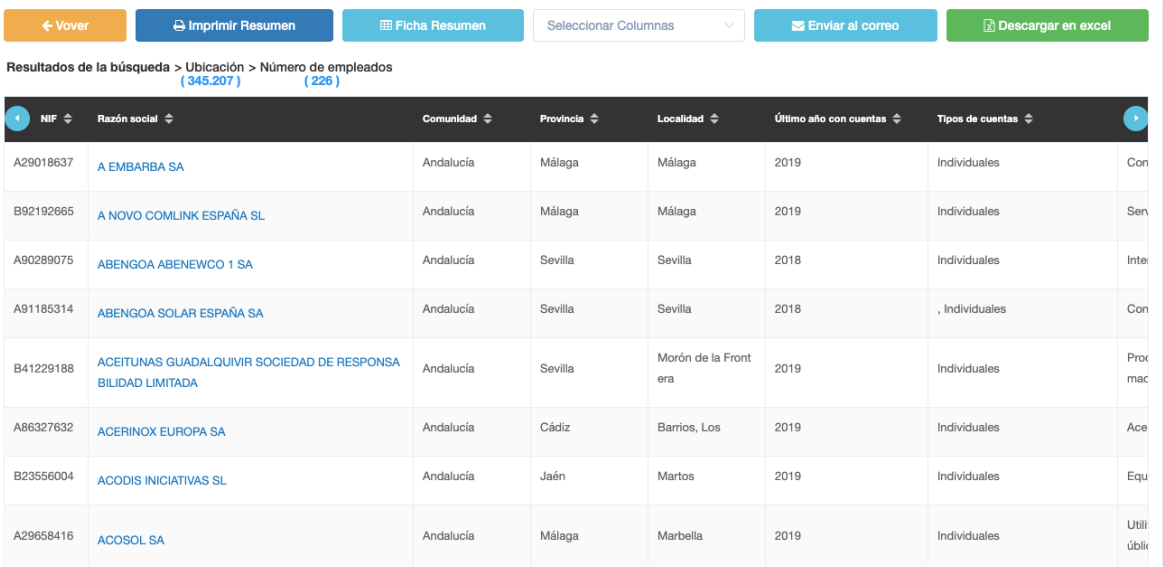
Esta empresa no contestó a nuestros correos

Tras hablar con el mercado de venta de bases de datos, se decidió extraer la base de datos de Infocif, ya que a priori parece la más completa y no hay riesgo asociado a ello.

La página web del buscador [39] usa una petición para cargar los datos que podemos ver en la siguiente imagen:


*Ilustración 14: Filtros del buscador de empresas de Infocif*

La imagen anterior es su buscador, se puede ver que En Andalucía el número de empresas con más de 250 empleados es de 226.




| NIF | Razón social | Comunidad | Provincia | Localidad | Último año con cuentas | Tipos de cuentas | |
|---|---|---|---|---|---|---|---|
| A29018637 | A EMBARBA SA | Andalucía | Málaga | Málaga | 2019 | Individuales | Con |
| B92192665 | A NOVO COMLINK ESPAÑA SL | Andalucía | Málaga | Málaga | 2019 | Individuales | Ser |
| A90289075 | ABENGOA ABENEWCO 1 SA | Andalucía | Sevilla | Sevilla | 2018 | Individuales | Inte |
| A91185314 | ABENGOA SOLAR ESPAÑA SA | Andalucía | Sevilla | Sevilla | 2018 | , Individuales | Con |
| B41229188 | ACEITUNAS GUADALQUIVIR SOCIEDAD DE RESPONSABILIDAD LIMITADA | Andalucía | Sevilla | Morón de la Frontera | 2019 | Individuales | Proc mac |
| A86327632 | ACERINOX EUROPA SA | Andalucía | Cádiz | Barrios, Los | 2019 | Individuales | Ace |
| B23556004 | ACODIS INICIATIVAS SL | Andalucía | Jaén | Martos | 2019 | Individuales | Equ |
| A29658416 | ACOSOL SA | Andalucía | Málaga | Marbella | 2019 | Individuales | Utili úbli |

*Ilustración 15: Resultados búsqueda Infocif*

Y al buscar, el listado obtenido es el anterior. Podemos observar que cuenta con un botón público de "descargar listado" pero la información ya se muestra por pantalla.

Analizando las conexiones que realiza, se aisló la siguiente en Postman:

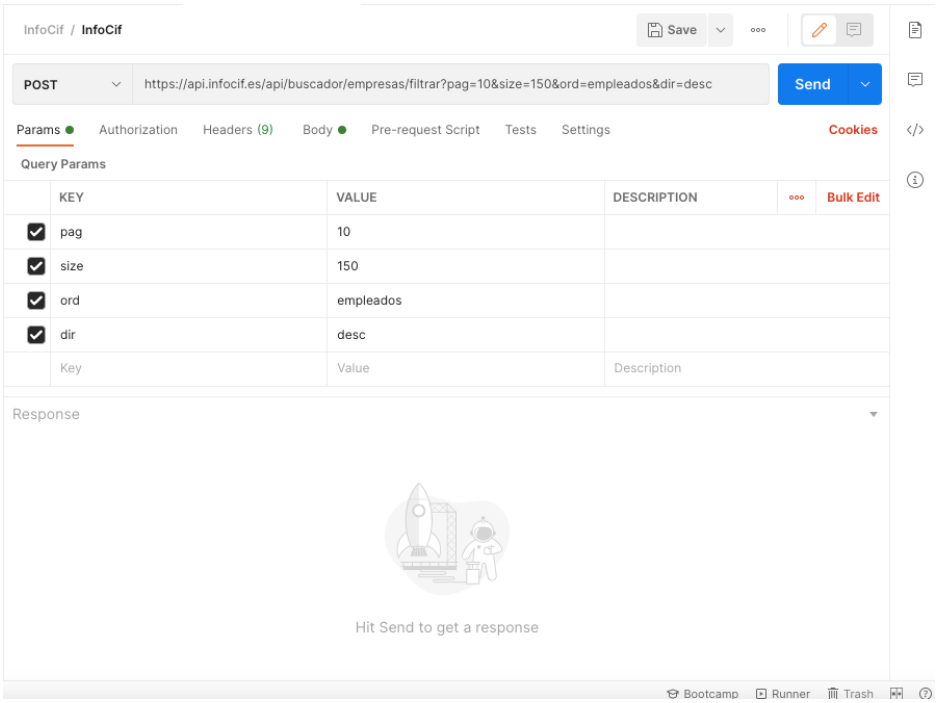



*Ilustración 16: Petición al buscador de empresas*

Como vemos, podemos añadir como parámetros la cantidad de empresas por petición (150 en este caso), el orden …. El argumento de páginas existe ya que esta herramienta está pensada para mostrar un sistema de paginación en su web (en la misma está limitado a 100). Por ejemplo, con los argumentos que se están usando, se recibirían las empresas 1500 a 1650 ordenadas por número de empleados. A la hora de la extracción fue necesario tener en cuenta estos índices para extraer todas las empresas.

La respuesta a la petición anterior es:

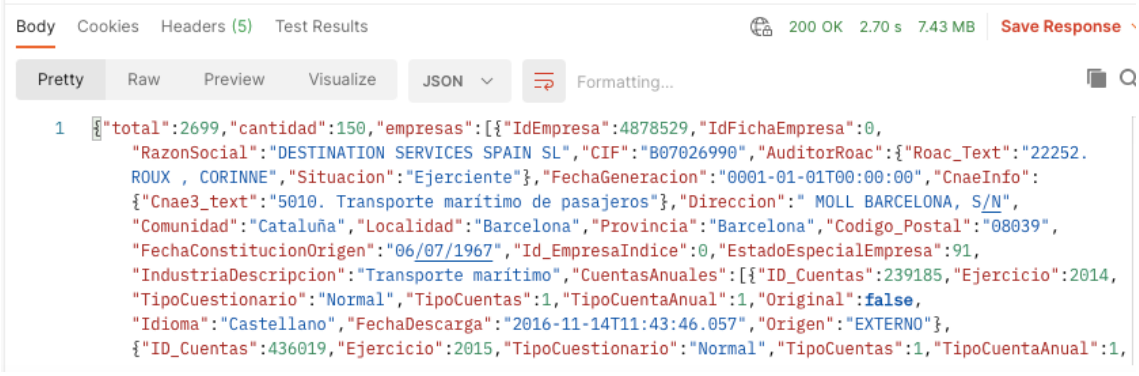



*Ilustración 17: Respuesta petición buscador*

Por lo que se va a desarrollar un programa que usando sus filtros extraiga empresas. El principal filtro que se utilizó fue el de empresas activas, con lo que la cantidad caía a algo menos de 1 millón.

Para realizar las peticiones, se decidió dividir por códigos postales ya que el número máximo de empresas por listado mostradas es de 10000, y se observó que la mejor manera de dividir para que quedasen repartidas sin llegar a esta cifra era por CP. Modificando la petición se consigue que en vez de 150 en 150 se puedan hacer de 1000 en 1000.(Se podría realizar para más, pero el tamaño del archivo descargado es muy grande y la página (y los proxies, gratuitos y de dudosa reputación) colapsan).

A continuación, se va a discutir sobre el tiempo de extracción que se iba a tardar a priori:

Teniendo en cuenta que no se paralelizó la extracción para no sobrecargar los servidores de infocif (un gesto de cortesía por su buen trato), las cuentas sobre la extracción son las siguientes.

11752 códigos postales

Un 10% de los códigos postales (aprox) tienen más de 1000 empresas por código postal.

20s por petición (incluye sleeps, tiempo de descarga de json, y tiempo de espera para no sobrecargarles)

Resultado, 71h de descarga de JSON ininterrumpidas. Finalmente se tardó 11 días ya que inicialmente la extracción se detenía en ocasiones (timeouts, la máquina se paraba…) , y esto era ejecutado sobre una máquina física que además tenía otras labores, con lo que no podía estar 24/7.

11 752 códigos postales.

11752 * 1,1 * 20s = 258544 s = 71h initerrumpidas

También se desarrolló un algoritmo que se llamó el "pool de proxis", ya que nos la única condición que pusieron de Infocif era que en caso de hacer demasiadas peticiones la IP sería baneada.

El algoritmo funciona de la siguiente manera:

Al principio, se lleva a cabo un web scraping de una web de proxis gratuitos [40]:

| IP Address | Port | Code | Country | Version | Anonymity | Https | Last Checked |
|---|---|---|---|---|---|---|---|
| 202.29.241.214 | 4153 | TH | Thailand | Socks4 | Anonymous | Yes | 20 seconds ago |
| 131.108.60.34 | 3629 | BR | Brazil | Socks4 | Anonymous | Yes | 20 seconds ago |
| 170.244.64.12 | 31476 | BR | Brazil | Socks4 | Anonymous | Yes | 20 seconds ago |
| 203.128.72.62 | 4145 | ID | Indonesia | Socks4 | Anonymous | Yes | 20 seconds ago |
| 89.201.194.102 | 1080 | HR | Croatia | Socks4 | Anonymous | Yes | 20 seconds ago |
| 194.143.251.73 | 4145 | HU | Hungary | Socks4 | Anonymous | Yes | 20 seconds ago |
| 152.32.84.108 | 4153 | PH | Philippines | Socks4 | Anonymous | Yes | 20 seconds ago |
| 117.211.166.168 | 33451 | IN | India | Socks4 | Anonymous | Yes | 20 seconds ago |
| 211.24.105.19 | 53598 | MY | Malaysia | Socks4 | Anonymous | Yes | 20 seconds ago |
| 36.66.177.168 | 43340 | ID | Indonesia | Socks4 | Anonymous | Yes | 20 seconds ago |
| 202.4.107.69 | 5678 | BD | Bangladesh | Socks4 | Anonymous | Yes | 20 seconds ago |
| 45.4.51.68 | 63253 | BR | Brazil | Socks4 | Anonymous | Yes | 20 seconds ago |
| 95.169.187.58 | 9053 | DE | Germany | Socks4 | Anonymous | Yes | 20 seconds ago |
| 54.82.228.163 | 1080 | US | United States | Socks4 | Anonymous | Yes | 20 seconds ago |

*Ilustración 18: Lista de proxis gratuitos*

Esta web ofrece una gran cantidad de proxis, pero de muy baja calidad (muchos o no responden o dejan de funcionar al poco tiempo). A esta web se le extraen los proxis de su

HTML (300 en total) ya que en sus condiciones no dicen lo contrario. En su HTML tienen lo siguiente:

```
Free proxies from free-proxy-list.net Updated at 2021-07-06 16:12:04 UTC.
186.219.96.47:54570 46.229.187.169:53281 109.70.189.70:56408 51.79.203.183:8080
185.162.142.81:53281 187.217.81.233:3128 191.100.28.115:8080 212.77.138.161:41258
122.15.131.65:57873 81.252.38.12:8080 185.37.211.222:50330 175.100.72.95:57938
67.55.185.240:1888 168.138.211.5:8080 79.111.13.155:50625 52.78.172.171:80
50.246.120.125:8080 27.123.1.46:3128 51.222.21.95:32768 51.222.21.93:32768
47.75.254.142:8888 …Muchos omitidos …46.237.255.10:3128 124.219.176.139:39589
14.207.120.81:8080 218.147.194.26:80 2.188.184.40:3128 5.56.133.127:3128
2.188.184.34:3128 118.175.207.180:40017 95.165.187.202:45396 177.66.112.221:57945
3.6.251.241:80 192.109.165.47:80 46.101.49.62:80 192.109.165.5:80 192.109.165.144:80
193.149.225.40:80 119.28.68.91:8080 100.20.170.195:80 203.189.156.149:80
176.110.121.90:21776 78.140.7.239:33943 45.121.216.218:55443 41.191.226.86:55443
60.251.183.62:80 113.254.228.146:8888 191.252.38.6:80 207.157.25.41:80
51.222.21.94:32768 187.217.81.229:3128 131.108.185.130:443 217.168.76.230:59021
177.99.206.82:8080 154.72.199.202:41201 140.227.76.44:6000 176.9.164.139:3128
185.175.119.113:47174 14.97.2.106:80 192.109.165.42:80 157.230.255.230:8118
140.227.77.186:6000 90.189.110.170:3128 192.109.165.61:80 192.109.165.115:80
91.89.89.10:8080 193.149.225.10:80 31.173.94.93:43539
```

Con lo cual cada vez que se realiza una petición a esta web se obtienen en total 300 proxis.

Una vez se tienen los proxis, empíricamente se comprobó que funcionaban regular, por lo que se creó la función checkproxies, que sirve para comprobar si una lista de proxis (definidos como ip:puerto) funcionan:

```
function checkProxy($proxies)
{
    //Array de valid proxies, contendrá los proxies que se haya comprobado que
funcionan correctamente
    $valid_proxies = array();
    array_push($valid_proxies, time());
    foreach ($proxies as $proxy) {
        // You can use any web site here to show you real ip such as
       //http://whatismyipaddress.com/
        // Se comprueba contra esta web, si devuelve respuesta
        $url = "https://whatismyipaddress.com/";
        // Get current time to check proxy speed
        $loadingtime = time();
        echo " TESTANDO PROXI" . $proxy . " <br>";

        $curlResponse = requestWeb($url, $proxy, "GET", null, 3000);
```

```
        //Itera por los proxies y almacena en una lista los que funcionan
        if ($curlResponse === false) {
        } else {
            //Si el proxy tarda menos de 6s, lo damos por bueno, y lo añadimos a
a lista de valid
            if ((time() - $loadingtime) < 6) {

                array_push($valid_proxies, $proxy);
            }
        }
    }
    return $valid_proxies;
}
```

Se puede observar que para comprobar el funcionamiento de un proxi lo que se hace es realizar una petición a una web que sea muy fiable, en caso de que no responda, ese proxi no funciona.

Una vez que ya tenemos un array de proxis válidos (normalmente 20 de cada 300).

Hay que comprobar que éstos, durante la ejecución, no dejen de serlo. (o que no se acabe la lista de proxis válidos).

Si en algún momento un proxi deja de funcionar, con unset se le elimina de array:

```
if ($curlResponseInicial == false) {
    //proxy, descartado, no funciona
    echo "PROXY INVALIDO EN GET CANTIDAD DE EMPRESAS : " .
$valid_proxies[$index_ramdom_proxy_used] . " <br>";
    unset($valid_proxies[$index_ramdom_proxy_used]);
    if (count($valid_proxies) == 0) {
        //Se re-buscan proxies;
        $valid_proxies = updateProxy($valid_proxies);
    }

      //En caso de que el array quede vacío, con update proxi se vuelven a
buscar proxies válidos
if (count($valid_proxies) == 0) {
    //Se re-buscan proxies;
    $valid_proxies = updateProxy($valid_proxies);
}
```

El JSON devuelto por las peticiones se puede observar en el anexo JSON respuesta.

Con el pool de proxis, ya se pudo extraer todos los datos necesarios de Infocif (varios GB) en archivos de texto. El siguiente paso parsear estos JSONS para introducir su información en base de datos.

```
//Introduce los jsons extraídos de infocif en la base de datos de Imeureka
public function fillDatabase(){
    //El timeout por defecto de las funciones de php es de 60s
    //Para esta en concreto, es necesario aumentarlo.
    set_time_limit(5000);
    //Esta función sirve para introducir en DB los datos del json de infocif,
//simplemente apunte a la ruta de la carpeta de empresas y
    //automáticamente le incluirá todo
    $emp = new Empresas();
    //Coge los archivos de la carpeta temp/empresas (json) y los introduce en DB
    $directorio = '/Applications/XAMPP/htdocs/imeureka/temp/empresas/';
    $ficheros = array_diff(scandir($directorio), array('..', '.','.DS_Store'));
    //print_r($ficheros);

    $i=0;
    foreach ($ficheros as $fichero){
        $fichero_json = file_get_contents($directorio.$fichero, true);
        $array_empresas_cp = json_decode($fichero_json, true);

        echo "FICHERO INTRODUCIDO" . $fichero;
        ob_flush();
        flush();

        foreach ($array_empresas_cp['empresas'] as $empresa) {
            //Insertar la empresa
            $emp = new Empresas();

            if (empty($empresa['UltimaCuentaAnual'])) {
                $ultima_cuenta_anual = "0";
                $ejercicio = 0;
                $importe_neto_cifra_negocio = 0;
                $resultado_ejercicio = 0;
                $empleados = 0;
            } else {
                $ultima_cuenta_anual = "1";
                $ejercicio = $empresa['UltimaCuentaAnual']['Ejercicio'];
                $importe_neto_cifra_negocio =
$empresa['UltimaCuentaAnual']['ImporteNetoCifraDeNegocioICIF'];
                $resultado_ejercicio =
$empresa['UltimaCuentaAnual']['ResultadoEjercicio'];
                $empleados = $empresa['UltimaCuentaAnual']['SumTotalEmpleados'];
            }


            $res = $emp->getById($empresa['IdEmpresa']);

            try{
```

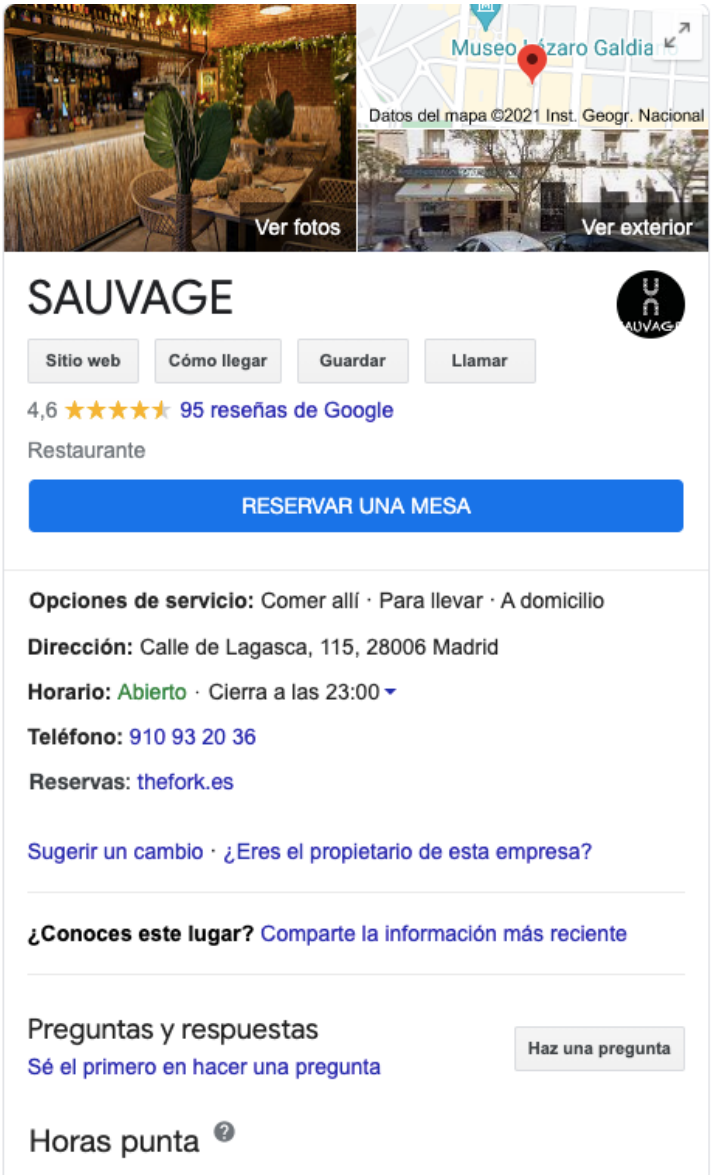
```
                $fecha_constitucion = (new DateTime(str_replace("/", "-",
$empresa['FechaConstitucionOrigen'])))->format("Y-m-d H:i:s");
            }catch (\Exception $e){
                $fecha_constitucion = '1600-01-01';
            }



            if (empty($res)) {
                //CON ID_EMPRESA
                $emp->insertArray(array(
                        "ID_empresa" => $empresa['IdEmpresa'],
                        "ID_ficha_empresa" => $empresa['IdFichaEmpresa'],
                        "Razon_social" => $empresa['RazonSocial'],
                        "CIF" => $empresa['CIF'],
                        "Cnae" => $empresa['CnaeInfo']['Cnae3_text'],
                        "Fecha_generacion" => date('Y-m-d h:i:s',
strtotime($empresa['FechaGeneracion'])),
                        "Direccion" => $empresa['Direccion'],
                        "Comunidad" => $empresa['Comunidad'],
                        "Localidad" => $empresa['Localidad'],
                        "Provincia" => $empresa['Provincia'],
                        "Cp" => $empresa['Codigo_Postal'],
                        "Fecha_constitucion_origen" =>$fecha_constitucion ,
                        "ID_empresa_indice" => $empresa['Id_EmpresaIndice'],
                        "Industria_descripcion" =>
$empresa['IndustriaDescripcion'],
                        "Ultima_cuenta_anual" => $ultima_cuenta_anual,
                        "Cargos" => count($empresa['Cargos']),
                        "Ejercicio" => $ejercicio,
                        "Importe_neto_cifra_negocio" =>
$importe_neto_cifra_negocio,
                        "Resultado_ejercicio" => $resultado_ejercicio,
                        "Empleados" => $empleados,
                    )
                );
            }
            //Insertar los cargos
            foreach ($empresa['Cargos'] as $cargo){


                try{
                    $fecha_acto = date('Y-m-d h:i:s',
strtotime($cargo['FechaActo']));
                }catch (\Exception $e){
                    $fecha_acto = '1600-01-01 01:00:00';
                }
                try{
                    $fecha_integracion = date('Y-m-d h:i:s',
strtotime($cargo['FechaIntegracion']));
                }catch (\Exception $e){
                    $fecha_integracion = '1600-01-01 01:00:00';
                }
```

```
                //var_dump($cargo['IdEmpresa']);
                //die();
                $contacto = new Contactos_empresas();
                $res = $contacto->getById($cargo['Id_CargosClasificados']);
                $res2 = $emp->getById($cargo['IdEmpresa']);
                if (empty($res) && !empty($res2)) {
                    //CON ID_EMPRESA
                    $contacto->insertArray(array(
                            "ID_contactos_empresa" =>
$cargo['Id_CargosClasificados'],
                            "ID_empresa" => $cargo['IdEmpresa'],
                            "Fecha_acto" => $fecha_acto,
                            "Fecha_integracion" =>  $fecha_integracion,
                            "Nombre" => $cargo['Nombre'],
                            "Cargo_espejo" => $cargo['CargoEspejo'],
                            "Subgrupo" => $cargo['SubGrupo'],
                        )
                    );
                }

            }
        }


    }
}
```

Los datos introducidos de Infocif en base de datos son:

- ID_empresa : único descargado de ellos
- Razón social
- CIF
- Cnae (código que indica la actividad de la empresa)
- Dirección
- Comunidad Localidad Provincia Cp
- Fecha_constitución_origen
- Industria_descripcion
- Nº Cargos
- Facturación último año
- Resultado último año
- Nº Empleados

Estos datos ya de por sí eran útiles, pero desde negocio se les ocurrió otra oportunidad para enriquecer aún más la base de datos, y es que, en muchas ocasiones al buscar los nombres de las empresas, aparecen en la barra de búsqueda las tarjetas de las empresas:



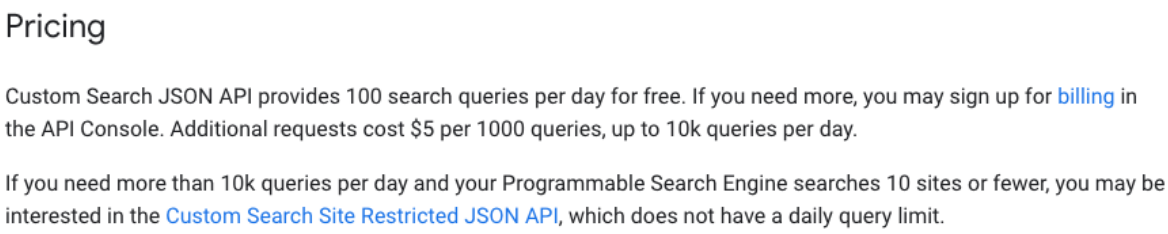
Esto incrementa los datos que podemos extraer, poniéndole valoración a las empresas en base a las opiniones de sus usuarios, pero más importante para nosotros, teléfono y enlace.

- Link
- Nombre comercial
- Puntuación
- Numero de reseñas
- Tlf

Siendo google, se era consciente de que para una máquina concreta era imposible tumbar sus servidores (como sí existía la posibilidad con Infocif). Por tanto, el primer approach que se hizo fue paralelizar las peticiones en grupos de 500 para tratar de tardar poco en extraer esas tarjetas.

Pero google no permite extracciones masivas automatizadas de su sitio web, por lo que al poco tiempo empezaron a salir captchas…

Existe una herramienta de pago para google search[41], pero es muy cara.


Pricing

Custom Search JSON API provides 100 search queries per day for free. If you need more, you may sign up for billing in the API Console. Additional requests cost $5 per 1000 queries, up to 10k queries per day.

If you need more than 10k queries per day and your Programmable Search Engine searches 10 sites or fewer, you may be interested in the Custom Search Site Restricted JSON API, which does not have a daily query limit.

*Ilustración 19: Precios Google Search*

Pero es de pago y el precio excede lo razonable.

Otro approach fue repetir las peticiones en paralelo pero con proxis, esto no funciona ya que Google conoce las ip's que son proxis y desde el primer la primera petición exige captcha para continuar.

Sin embargo, es sabido que google [42],[43] aunque activamente protege su buscador, no denuncia, por lo que finalmente se encontró una solución.

La solución final empleada fue realizarle peticiones en serie a google con tiempos entre las mismas aleatorios y con IP real.

El único problema de este approach es que el tiempo se dispara (aproximadamente 1300h), asumiendo una petición cada 5 segundos de media y un millón de empresas.

Con el método de complete_database se realiza lo anteriormente expuesto:

```
public function completeDatabase(){
    //Esta función extrae los datos de las tarjetas de google cloud y los guarda
en la base de datos
    Logging::write("El script se ha iniciado, esto sólo es un test de que el
logging funciona",'error_extraccion_empresas_google');


    do {
        $emp = new Empresas();
        //Devuelve empresas con razones sociales no extraidas
        $res = $emp->getNoExtraidoLimit(100);
        //Se cambian los ' ' por '+' para darle formato de búsqueda de google
        foreach ($res as $razon_social) {
            $emp = new Empresas();
            $temp = str_replace(' ', '+', $razon_social['Razon_social']);
            $temp2 = 'https://www.google.es/search?q=' . $temp . '&oq=' . $temp .
'';
            try {
                $datos = $this->peticion_anonimizada($temp2, 2, 8);
            }catch (\Exception
$e){Logging::write($e,'error_extraccion_empresas_google');}
            echo "PETICIÓN REALIZADA A : " . $temp2 . "<br><br><br>";
            ob_flush();
            flush();
            var_dump($datos);
            //Actualizas en db
            try {
            if($datos["link"]==null){
                $emp->edit($razon_social["ID_empresa"],array("Extraido"=>1));
            }else{
                $emp->edit($razon_social["ID_empresa"],array("Extraido"=>1,"Link"
=>
$datos["link"],"Nombre_comercial"=>$datos["nombre_comercial"],"Puntuacion"=>$dato
s["puntuacion"],"Numero_resenas"=>$datos["numero_resenas"],"Tlf"=>$datos["tlf"]
```

```
));
            }
            }catch (\Exception
$e){Logging::write($e,'error_extraccion_empresas_google');}
            //borro y recreo el archivo
            try{
                unlink($temp_route . 'cronEmpresasCheck.txt');
                fopen($temp_route . 'cronEmpresasCheck.txt', "w");

            }catch (\Exception $e){
                echo "Error en la creación de archivos en el bucle";
                echo $e;
            }
        }


    }while(count($res)!=0);
}
```

La solución final fue ejecutar la función anterior sobre el servidor de IMeureka, que está siempre en línea, para que fuese poco a poco extrayendo.

Podemos observar el resultado final con la siguiente query:

SELECT * FROM insurance_new.empresas where link is not null order by Empleados desc;

| | | | | |
|---|---|---|---|---|
| http://www.casatarradellas.es/ | Casa Tarradellas S.A. | 3 | 504 | 938 81 65 50 |
| http://www.alainafflelouoptico.es/ | Alain Afflelou Óptico | 4 | 0 | +34 911 51 77 00 |
| https://hralaluz.es/home | Hospital Residencia Asistida la Luz S. Privado | 4 | 42 | +34 947 50 01 50 |
| http://www.k08.es/ | Pizarras Los Tres CuÃ±ados | 5 | 0 | +34 988 32 47 20 |
| http://www.lawebdecanada.com/ | CANADA | 4 | 18 | +34 933 56 84 83 |
| http://www.niprorenalsolutions.com/es/contacto.... | Nipro Renal Solutions | 4 | 0 | +34 973 30 64 30 |
| https://www.carrefour.es/ | Centros Comerciales Carrefour, S.A | 3 | 206 | 914 90 89 00 |
| https://originiafoods.com/plantas/saar/ | Saar Originia Foods | 4 | 32 | +34 976 66 27 95 |
| https://www.es.issworld.com/ | Iss Facility Services Sa. | 1 | 0 | 914 84 24 00 |
| http://www.prosegur.com/ | PROSEGUR SERVICIOS INTEGRALES DE SE... | 3 | 18 | 915 89 82 40 |
| https://bit.ly/3rdpPOC | Adecco TT SA | 3 | 28 | 916 60 18 10 |
| https://www.consum.es/ | Consum | 5 | 0 | 961 97 40 00 |
| http://www.pescapuerta.es/ | Pescapuerta S.A. | 4 | 15 | +34 986 29 25 50 |
| http://www.alcampo.es/compra-online?utm_cam... | Alcampo | 3 | 204 | 915 93 98 72 |
| https://www.fecsa.net/ | Fábrica Española De Confecciones | 5 | 6 | +34 916 52 34 00 |
| http://www.manpower.es/ | Manpower | 1 | 10 | 914 54 37 50 |
| http://www.renault.es/ | Renault Espana SA | 2 | 43 | 915 06 53 58 |
| http://www.novavidaempleo.com/ | Centro Especial De Empleo Novavida S L | 4 | 13 | +34 928 53 04 92 |
| http://lamuralladeestepa.es/ | La Muralla | 4 | 33 | +34 955 91 32 40 |
| http://leroymerlinrgpd.es/ | Leroy Merlin España, S.L.U. | 3 | 10 | 900 813 344 |
| https://www.lidl.es/ | Lidl | 4 | 1 | 900 958 311 |
| http://urbegi.com/ | Urbegi | 4 | 18 | +34 946 80 19 34 |
| http://www.tiruna.com/ | Tiruña | 4 | 22 | +34 948 35 51 11 |
| http://www.emtmadrid.es/ | EMT - Sede Central | 4 | 336 | 914 06 88 10 |
| http://www.espaderosdelatlantico.com/ | Espaderos del Atlantico | 4 | 9 | +34 986 24 34 80 |
| https://www.acciona-service.com/es/areas-de-a... | Acciona Facility Services S.A | 3 | 17 | 914 71 02 48 |
| https://www.sacyrservicios.com/facilities/ | Sacyr Facilities | 3 | 53 | 915 45 53 00 |
| https://www.ikea.com/es/es/ | IKEA Ibérica, S.A. (Oficinas) | 3 | 552 | 900 400 922 |
| http://www.ingesan.es/ | OHL Servicios Ingesan | 4 | 26 | 917 74 70 00 |
| https://www.michelin.es/ | Michelin España y Portugal S.A. | 4 | 25 | 914 10 50 00 |
| https://www.eumedica-industries.com/ | Eumedica Pharmaceuticals Industries, s.l. | 4 | 9 | +34 985 26 05 04 |
| http://www.ford.es/ | Ford España S.L. | 3 | 128 | 900 807 090 |
| https://site.groupe-psa.com/madrid/es/ | Centro de Madrid de PSA Peugeot Citroën | 4 | 726 | 913 47 20 00 |
| https://www.aramark.es/ | Aramark Servicios de Catering S.L. | 3 | 42 | +34 932 40 21 41 |
| http://www.onet.es/ | Onet Iberia | 3 | 45 | +34 914 90 12 50 |
| http://www.tragsa.es/ | Tragsa | 3 | 9 | 913 96 34 00 |

*Ilustración 20: Base de datos de empresas – datos adicionales*

Si recordamos, anteriormente se mencionó que la base de datos de Infocif también devuelve los cargos. Concretamente devuelve cargos estandarizados de Infocif (presidente, apoderado… etc) y el nombre y dos apellidos de la persona.

En el JSON que devolvía Infocif también había información pública sobre los cargos que componen la empresa, se extrajeron más de 3 millones de cargos.

Esta queda almacenada en la DB con:

- ID_contactos_empresa
- ID_empresa
- Fecha_acto
- Fecha_integracion
- Nombre
- Cargo

Desde negocio se pensó que sería útil tener los emails corporativos de estas personas, ya que teniendo sus nombres se podrían realizar búsquedas por la web. Hay que tener en cuenta que la gente de negocio no conoce de programación ni de ingeniería, por lo que desconocen RGPD y piden sin tener en cuenta trabas.

Antes de la extracción de datos y ante la posibilidad de que fuese ilegal, se inició una investigación para repetir la reunión anterior con los resultados. Una de las herramientas investigadas fue LinkedIn, ya que muchas personas tienen cuenta y es una red social "corporativa".

Datos de LinkedIn: prohibidos, LinkedIn cerró sus datos y ya no son públicos. (Se pretendía extraer todo de las personas utilizando su nombre y empresa a la que pertenecen). Sin necesidad de abogados se llegó a la conclusión de que esto sería ilegal.

Pero hay una forma técnica de obtener los emails, y es que existen unos servidores denominados MX que responden si un email existe o no:

Entonces, únicamente con el nombre y apellido de la persona, se puede inferir su email de la siguiente manera

Nombre.apellido@dominiodelaempresa

Napellido@dominiodelaempresa

Debido a que muchas empresas lo hacen así.

Y a continuación realizar peticiones al servidor MX [44] ("registro de intercambio de correo" – recurso DNS que encamina correos electrónicos por internet) para ver si existe el email.

El problema es que al tener 3M de contactos, son 6M de peticiones a estos servidores. Los servidores bloquean las ip's tras aproximadamente 10 peticiones y comienzan a responder que los emails no existen.

Tras buscar acerca de la verificación de la veracidad de emails, se encontró una librería de PHP que implementaba exactamente la funcionalidad que necesitábamos [45].

Implementando esta librería, podíamos saber con certeza si un email existía o no:

```
if($mail->check($email)){
    echo 'Email &lt;'.$email.'&gt; is exist!';
    $worked=true;
    $Con = new Contactos_empresas();
    $Con->edit($ID_contactos_empresa,array("Extraido"=>1, "Email"=>$email));
}elseif(verifyEmail::validate($email)){
    $Con = new Contactos_empresas();
    $Con->edit($ID_contactos_empresa,array("Extraido"=>1,"Email"=>"NOT FOUND"));
    echo 'Email &lt;'.$email.'&gt; is valid, but not exist!';
    ob_flush();
    flush();
}else{
    $Con = new Contactos_empresas();
    $Con->edit($ID_contactos_empresa,array("Extraido"=>1,"Email"=>"NOT FOUND"));
    echo 'Email &lt;'.$email.'&gt; is not valid and not exist!';
    ob_flush();
    flush();
}
```

La librería funciona una pequeña cantidad de veces y luego deja de funcionar, esto es debido a que los servidores bloquean ip's. Además, hay que tener extremo cuidado desde el email desde el que se pregunta, porque en caso de que bloqueen varias veces se puede acabar en una lista de spam. Teniendo en cuenta que IMeureka usa mailing, terminar en una lista de spam de manera irremediable acabaría con el despido de un servidor. Por lo que se prefiere no hacerlo

Los proxis no soportan las comunicaciones con el protocolo del servidor MX (que no es http). Por lo que no se puedan usar. Se investigó usar una VPN que rotase cada 5 minutos. Se estimó que como mucho se puede hacer una petición cada 20s para no ser bloqueados.

Aunque se teorizaron sobre posibles soluciones, como contratar VPN rotativas para acelerar prevenir el bloque de IP's y randomizar el orden de extracción de los emails (para no preguntar cientos de veces seguidas al mismo dominio) tras explicarle a negocio la problemática de RGPD se decidió detener esta línea de investigación debido a las posibles consecuencias legales a las que la empresa se podría enfrentar.

Choca con RGPD y está todo atado

| ID_empresa | Razon_social | CIF | Cnae | Direccion | Comunidad |
|---|---|---|---|---|---|
| 3409152 | FELIX DE INCHAURRA( | B48690994 | 4671. Comercio al por | CL ARETXETA | País Vasco |
| 2688512 | ESTABANELL Y PAHISA | A08016057 | 4110. Promoción inmc | CL REC 26-28 | Cataluña |
| 3868160 | ALTIDA TRADE SL | B99307225 | 4690. Comercio al por | URB NUESTRASEÑOR | Aragón |
| 3279360 | VOSS SA | A08402554 | 4669. Comercio al por | Paseo comercio,90 | Cataluña |
| 3738112 | AGRO PECUARIA DE M | B08067233 | | CLMAS LA ROVIRA, S/ | Cataluña |
| 3345152 | CARNS PONT SA | A58202318 | | CALLE PUJADES Nº 68 | Cataluña |
| 2822144 | HORYFRUMA SL | B73398216 | 4631. Comercio al por | CTRA DE LAESTACION | Región De Murcia |
| 3479552 | LURPELAN TUNNELLI | A95431607 | | POLIGINDUSTRIAL SA | País Vasco |
| 6232320 | TABIQUES Y TECHOS I | B26522128 | | C/ SEGADOR - PARCEI | La Rioja |
| 2497280 | BODEGAS EMILIO MO | B47359492 | 1102. Elaboración de | CR VALORIA-PEÑAFIE | Castilla Y León |

| Localidad | Provincia | Cp | Fecha_constitucion_o | Industria_descripcion | Ultima_cuen | Cargos |
|---|---|---|---|---|---|---|
| Getxo | Vizcaya | 48992 | 1994-09-23 | Petróleo y gas | 1 | 14 |
| Granollers | Barcelona | 8401 | 1927-07-01 | Inmobiliarias y similar | 1 | 35 |
| Muela, La | Zaragoza | 50196 | 2011-01-01 | Minoristas (excepto a | 1 | 2 |
| Sabadell | Barcelona | 8203 | 2021-03-15 | Equipos industriales | 1 | 24 |
| Moià | Barcelona | 8180 | 2021-03-15 | Agricultura | 1 | 5 |
| Barcelona | Barcelona | 8005 | 1986-06-09 | Bebidas y Tabaco | 1 | 18 |
| Lorca | Murcia | 30800 | 2005-10-07 | Productos alimenticio | 1 | 5 |
| Muskiz | Vizcaya | 48550 | 2006-06-28 | Construcción y Desarr | 1 | 35 |
| Logroño | Rioja, La | 26006 | 2014-10-30 | Construcción y Desarr | 1 | 2 |
| Pesquera de Duero | Valladolid | 47315 | 1995-04-08 | Bebidas y Tabaco | 1 | 18 |

| Ejercicio | Importe_net | Resultado_e | Empleados | Extraido | Link | Nombre_comercial | Puntuacion | Numero_res | Tlf |
|---|---|---|---|---|---|---|---|---|---|
| 2018 | 15673370,9 | 986700 | 12 | 1 | http://felixinchaurrag | Félix de Inchaurraga S | 4 | 4 | +34 944 92 05 12 |
| 2018 | 7894368 | 5670674 | 1 | 1 | http://www.estabane | Estabanell y Pahisa Er | 3 | 2 | +34 938 60 91 00 |
| 2018 | 14479322 | 207148 | 6 | 1 | http://www.altidatra | Altida Trade, SL | 0 | 0 | +34 876 66 48 42 |
| 2018 | 13132348 | 513978 | 77 | 1 | https://www.voss-flu | VOSS S.A.U. | 4 | 29 | +34 937 10 62 62 |
| 2018 | 219258 | -84858 | 4 | 1 | http://www.agromoia | AgropecuÃ ria Del Mo | 4 | 31 | +34 938 30 03 70 |
| 2018 | 6330298 | -127268 | 40 | 1 | http://www.carnspon | CARNS PONT, S.A. | 5 | 0 | +34 933 00 46 60 |
| 2018 | 1525634,09 | 9755 | 98 | 1 | https://www.horyfrur | Horyfruma, S.L. | 3 | 49 | +34 968 44 12 69 |
| 2018 | 8337281,89 | -203540 | 45 | 1 | http://www.lurpelan. | LURPELAN TUNNELLIF | 4 | 8 | +34 946 40 09 89 |
| 2018 | 5208821 | 380251 | 18 | 1 | http://www.tabiques | Tabiques y Techos Re | 4 | 31 | +34 941 25 81 47 |
| 2018 | 19195755,7 | 4620534 | 81 | 1 | http://www.emiliom | Bodegas Emilio Moro | 4 | 237 | +34 983 87 84 00 |

*Ilustración 21:Algunos datos de la base de datos de empresas*

Esto ha sido un ejemplo de empresas en la base de datos. Como vemos, se han obtenido los mismos campos o más de los que ofrecían las empresas de pago, con lo que ha sido una buena decisión.

Particularmente si en un futuro se decide renovar las empresas, el proceso ha quedado descrito en este ensayo, por lo que no sería necesario comprar una base de datos de nuevo (en caso de que se hubiese seguido esta línea).

## *5.2 DESARROLLO DE HERRAMIENTAS*

En el apartado anterior se ha explicado cómo se realizó la extracción de los datos, desde los recursos de internet hasta la base de datos de IMeureka.

5.2.1 Herramienta catastro:

Para ayudar al usuario al ahora de que introduzca los datos, se integró con la API de Google Maps Geocoding[46]. Gracias a ello, el usuario introduce su dirección con una ayuda predictiva, y Google responde con una latitud y longitud para esa dirección. Esto es muy útil porque como vimos el método consulta_catastro era capaz de responder con esos datos.

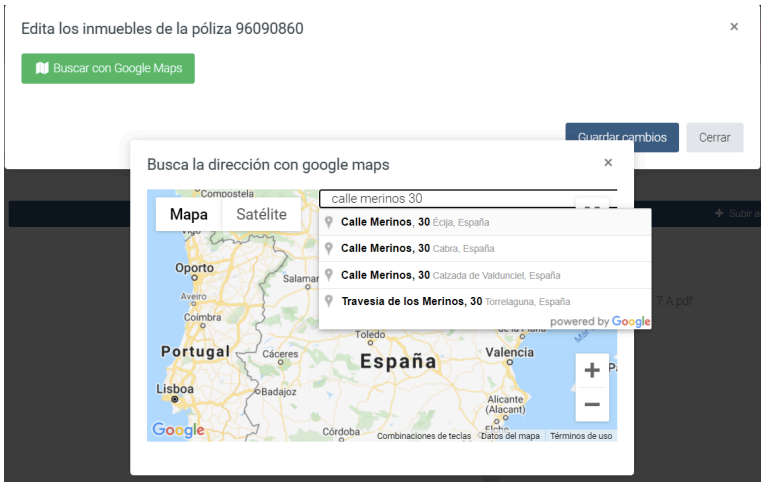



*Ilustración 22 :Introducir dirección para obtener catastro en IMeureka*

En la ilustración podemos ver cómo buscando una calle le aparecen las recomendaciones al usuario, en muchos casos por tanto no es necesario introducir tipo de vía, provincia… Ya se encarga google Maps de devolver las coordenadas.

Quedan por tanto integrados los datos del catastro en la plataforma de IMeureka.


*Ilustración 23 :Datos catastrales calle Merinos 30 en IMeureka*

Si el usuario decide hacer click bajo "referencia catastral", se le llevará a ese inmueble en la web del catastro.

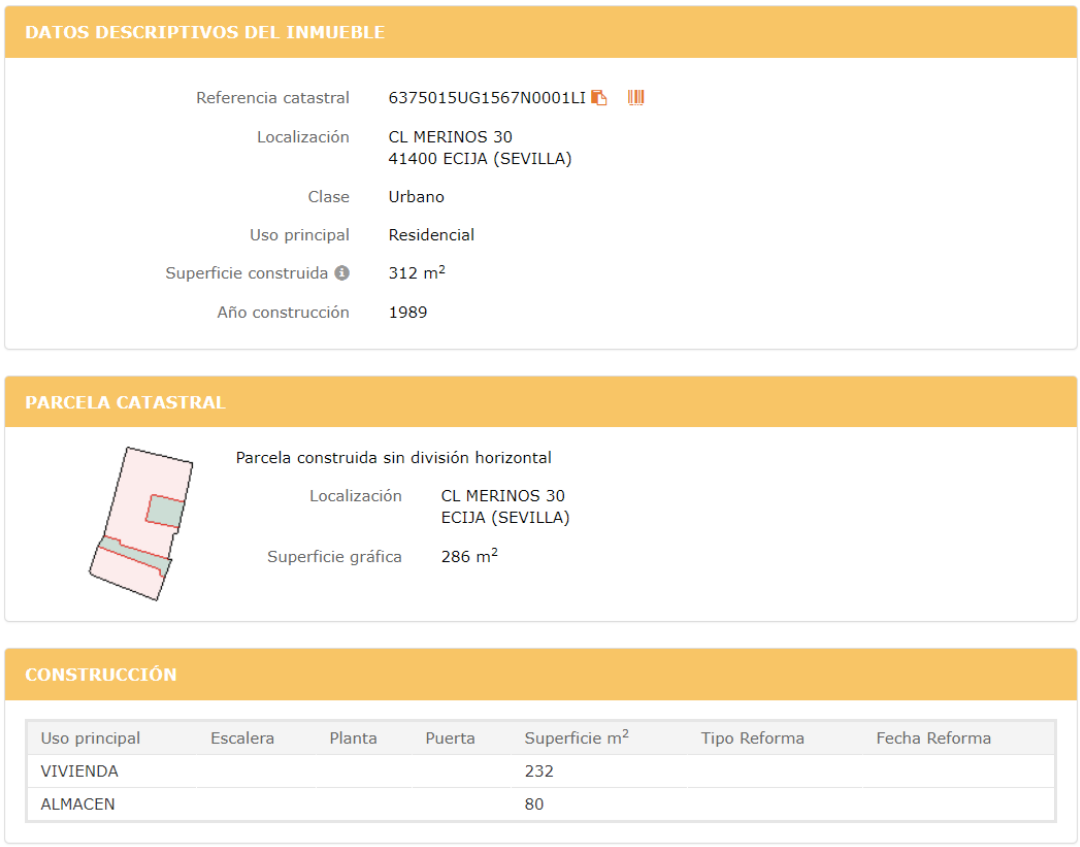



*Ilustración 24 :Datos catastrales calle Merinos 30 en catastro*

En caso de que en la ubicación seleccionada haya más de un inmueble (habitual en edificios y centros comerciales) se le dará la opción de seleccionar uno de ellos para almacenarlo.

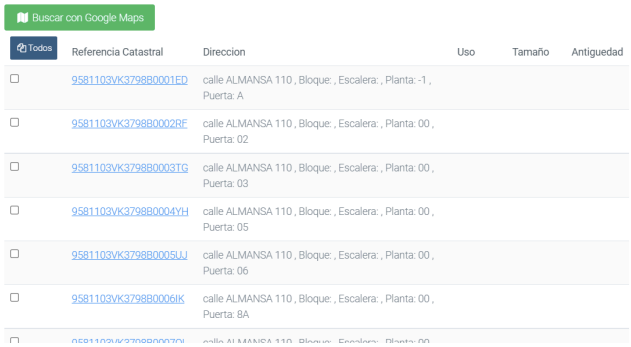



*Ilustración 25: Datos catastrales Almansa 110 en IMeureka*

Esta herramienta ayuda a ahorrar tiempo tanto a corredores como a usuarios, en caso de que una póliza contenga inmuebles, pueden usar este servicio para rellenar sus datos la manera más rápida posible.

### 5.2.2 Herramienta de empresas:

Tras extraer el millón de empresas y almacenarlas en base de datos, se desarrolló un buscador que permitiese a usuarios inexpertos explotar los datos extraídos.

La herramienta de empresas consta de un buscador que les permite filtrar por casi cualquier parámetro:

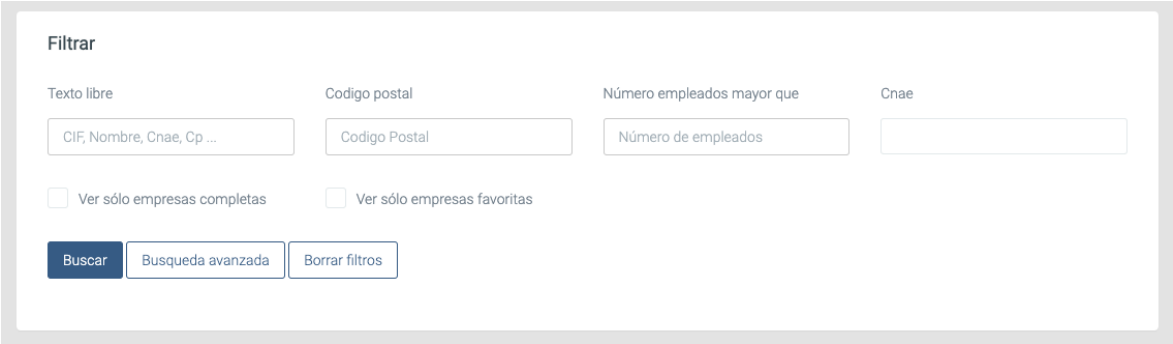



*Ilustración 26: Buscador de empresas en IMeureka*

Internamente, lo que hace el usuario sin saberlo es construir una query MYSQL, y verá los resultados en el listado que se muestra a continuación.


*Ilustración 27: Listado de empresas en IMeureka*

El uso esperado es aprovechar los viajes / tiempo muerto que tengan los comerciales para rellenar.

Imaginemos que el comercial va a vender una póliza de seguros a Sevilla a una empresa de aceite. Probablemente estudie la situación del riesgo, que sea similar para las empresas de la zona.

El CNAE 104 es fabricación de aceites

El comercial rápidamente ve que hay 35 empresas que se dedican a la fabricación de aceites. Pulsando en obtener informe se descarga los datos, que podrá llevar cuando vaya a hacer visitas.

Gracias a esto el comercial descubre empresas a las que les puede hacer ofertas similares a las que ya se ha preparado:

| ID_empresa | Razon_social | CIF | Cnae | Direccion |
|---|---|---|---|---|
| 3044120 | PETROLERA CANAI | B91262238 | 1044. Fabricación ( | AVDA. SAN FRANC |
| 2988228 | MOLINO DE CARM | B91297481 | 1044. Fabricación ( | FINCA SANTA ELEN |
| 3799055 | AROMATICA EURO | B91191767 | 1044. Fabricación ( | C/ Salvador Moren |

| Comunidad | Localidad | Provincia | Cp | Fecha_constitucior | Industria_descripci | Ultima_cuenta_an |
|---|---|---|---|---|---|---|
| Andalucía | Sevilla | Sevilla | 41018 | 14/1/03 | Productos alimenti | 0 |
| Andalucía | Carmona | Sevilla | 41410 | 2/7/03 | Productos alimenti | 0 |
| Andalucía | Estepa | Sevilla | 41560 | 25/1/02 | Productos alimenti | 0 |

| Cargos | Ejercicio | Importe_neto_cifra | Resultado_ejercicie | Empleados | Link | Nombre_comercia |
|---|---|---|---|---|---|---|
| 6 | 0 | 0.00000 | 0 | 0 | No encontrado | No encontrado |
| 3 | 0 | 0.00000 | 0 | 0 | No encontrado | No encontrado |
| 2 | 0 | 0.00000 | 0 | 0 | No encontrado | No encontrado |

| Puntuacion | Numero_resenas | Tlf | Nombre | Cargo | Email Contacto |
|---|---|---|---|---|---|
| No encontrado | No encontrado | No encontrado | BENITEZ LEON SEB | administrador mar | No encontrado |
| No encontrado | No encontrado | No encontrado | CARRILLO AGUILAI | administrador unic | No encontrado |
| No encontrado | No encontrado | No encontrado | PEREZ ENFEDAQUI | administrador solic | No encontrado |

*Ilustración 28: Datos de empresas filtradas por CNAE 1044 y descargadas en excel*

Tiene un botón de favoritas, asociadas a un corredor, para que el corredor pueda volver de un vistazo a las herramientas

Se permite filtrar por cualquier campo que se ha visto, particularmente, por código postal, numero empleados, CNAE, fecha de creación y facturación.

### 5.2.3 Licitaciones:

La idea de esta herramienta es que les sirva a los brókeres asociados a IMeureka a captar más negocio, para ello, constantemente se va a actualizando con las nuevas licitaciones que van surgiendo.

Al entrar en la herramienta, los brókeres disponen de un buscador, con el cual pueden las licitaciones que en principio les interesen:

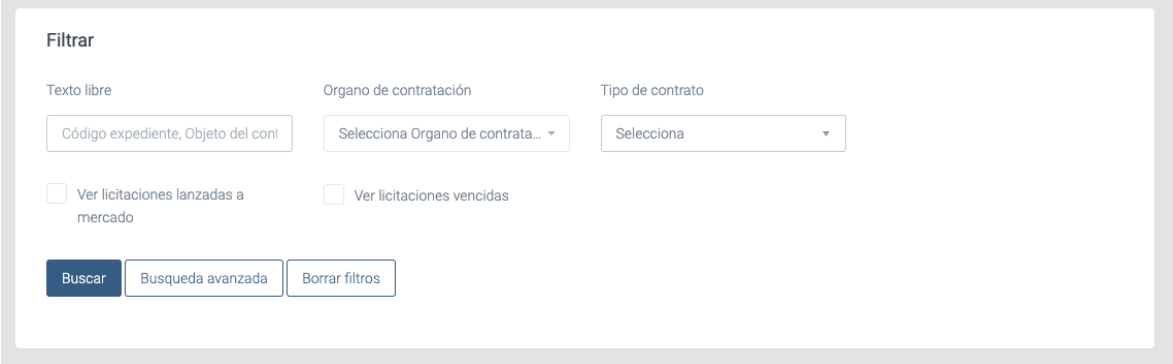



*Ilustración 29: Filtro de licitaciones*

El resultado es mostrado en un listado.


*Ilustración 30: Resultado filtrado licitaciones*

En caso de pulsar sobre "ver" se verán los detalles de la licitación:


VER LICITACIÓN

**Codigo expediente :** PcPG/2021/802449

**Órgano de contratación :** Consorcio Galego de Servizos de Igualdade e Benestar

**Tipo de contrato :** Privado

**Estado licitacion :** Publicada

**Objeto del contrato :** Seguro de responsabilidad civil para los/las participantes de las prácticas formativas no laborales en empresa en itinerarios de inclusión sociolaboral de la red de equipos comarcales de inclusión social gestionados por el Consorcio Gallego de Servicio de Igualdad y Bienestar. Financiado con fondos REACT-UE en el marco del programa operativo FSE GALICIA 2014-2020

**Presupuesto base sin impuestos :** 700,00 €

**Valor estimado contrato :** 0,00 €

**Codigo CPV :** 66516000 Servicios de seguros de responsabilidad civil.

**Lugar de ejecucion :** España - Galicia

**Procedimiento de contratacion :** Abierto simplificado

**Url :** https://contrataciondelestado.es/wps/poc?uri=deeplink%3Adetalle_licitacion&idEvl=kUB3ZkR6w2lvYnTkQN0%2FZA%3D%3D

¡No se encuentran archivos adjuntos!

*Ilustración 31: Detalles de una licitación*

En lanzar oferta se desarrolló una funcionalidad puntera, y es que cuando el corredor pulsa sobre este botón, se le crea un email autogenerado con los datos de la licitación a cotizar (que puede modificar):


*Ilustración 32: Lanzador automático de cotizaciones*

Y un menú seleccionable con las personas a las que puede escribir:

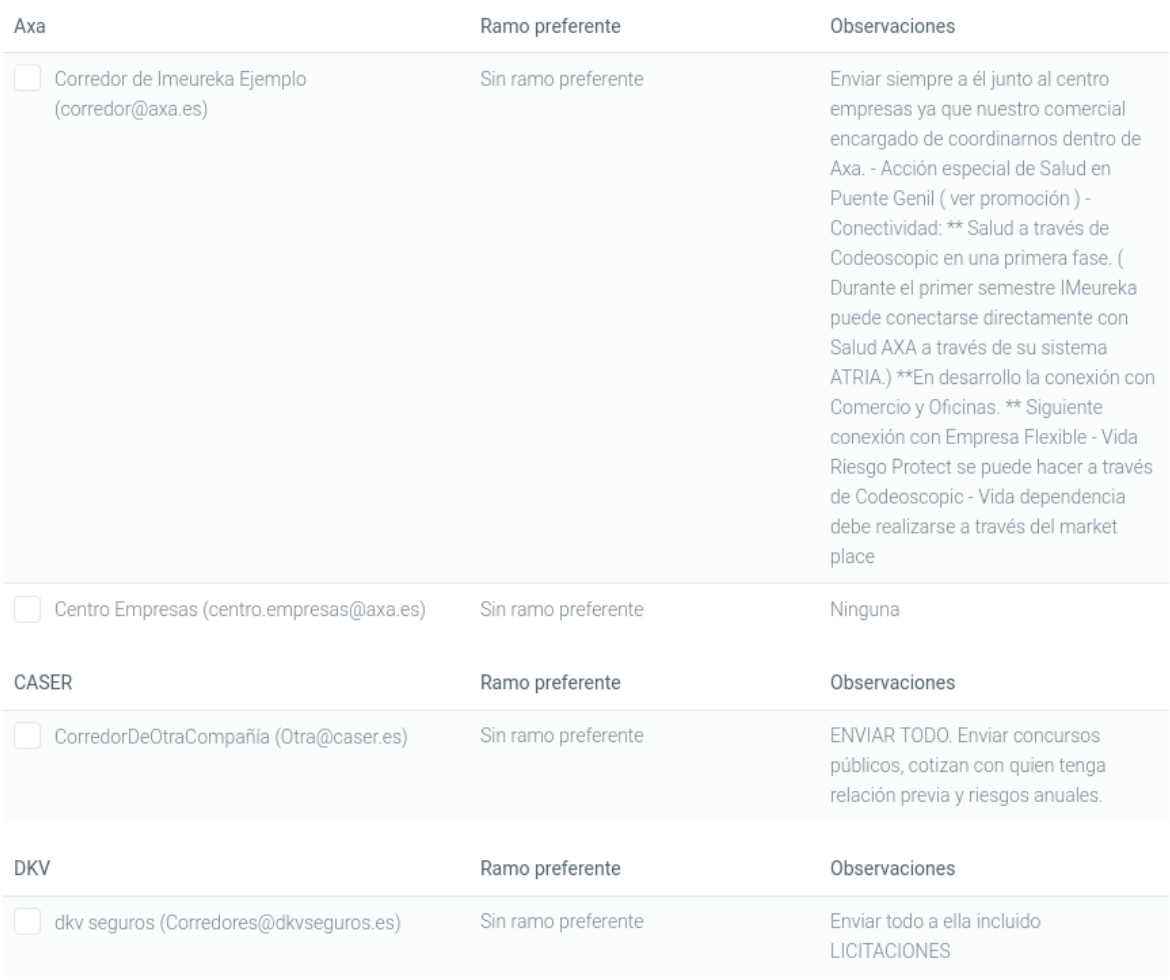



*Ilustración 33: Seleccionar a quién mandarle la contización*

TEU (gestión de multas):

Para el apartado de gestión de multas se empezó buscando la información en el buscador de tablón edictal único. Se comprobó que el tablón no es un archivo de multas, si no que únicamente muestra las multas por pagar, y tras abonarlas desaparecen del mismo. Esto causa que sea muy complicado encontrar alguna multa, ya que, en general, son pagadas en menos de 15 días para acogerse al 50% de descuento que ofrece la administración por pronto pago.

Suponía por tanto un gran esfuerzo de recursos y tiempo para un caso que es muy raro que se de.

Pero más allá de ello, el motivo principal por el que no se creó una API para el tablón edictal es que se concluyó que los datos extraíbles del TEU no eran útiles para nuestro caso de uso, ya que no es información que le interese al corredor, si no a la aseguradora.

En realidad, entregarle esa información a la aseguradora incurre en un perjuicio hacia nosotros mismos, ya que el cliente podrá conseguir una oferta mejor a través de otros corredores que no ofrezcan esa información. Al informar a la aseguradora que está cotizando el riesgo de que el cliente a asegurar tiene multas pendientes es muy probable que, como mínimo, se encarezca la prima. Con lo que se puede prever que el cliente contratará su seguro con otra empresa que le podrá ofrecer mejores condiciones.

# Capítulo 6. ANÁLISIS DE RESULTADOS

El proyecto puede ser dividido en dos partes diferenciadas, la sección acerca de extracción de datos, en la cual se usan tres técnicas distintas de extracción y la sección de visualización de los datos, en la cuál se crean las herramientas de explotación para los datos que posteriormente serán vendidas a corredores bajo un modelo de suscripción.

Desde el punto de vista legal, se ha cumplido a rajatabla con la ley vigente, llegando incluso a parar extracciones planeadas tras realizar análisis legales de lo que se pretendía hacer ; que resultaron desfavorables.

Se han extraído los datos necesarios en el tiempo estimado para ello y se han desarrollado programas que hoy en día son herramientas de uso diario para muchos corredores. Los datos extraídos le son útiles a las corredurías tradicionales para ser más eficientes y aumentar sus oportunidades de negocio.

Se han cumplido los objetivos del TFM ya que la plataforma, con estas nuevas funcionalidades, es ahora mucho más atractiva para corredores.

Las bases de datos tienen actualmente los siguientes tamaños:

Empresas: 966445 registros

Contactos de empresas: 2860478 registros

Licitaciones: 469 registros

Catastro: 62 registros

Y se pueden observar los detalles de los datos extraídos tanto a lo largo del TFM como en los anexos finales.

El sistema de licitaciones lleva meses funcionando sin interrupción, con lo que se demuestra que el sistema desarrollado es sólido.

# Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

IMeureka, al ser una empresa de nueva creación, necesitaba empezar por la extracción de datos, el primer paso dentro del proceso de Big Data.

Para la empresa, una vez concluida esta extracción, se le abren muchas puertas. Una vez que se empiece a usar la herramienta de licitaciones por parte de otros corredores, esos datos podrán ser usador por la correduría de IMeureka para encontrar patrones en la adjudicación de licitaciones. Un ejemplo podría ser cruzar los datos de los clientes de IMeureka (bajo el paraguas RGPD de interés legítimo) con las licitaciones conseguidas para averiguar porqué ganaron la licitación e intentar imitarlos.

Otro trabajo podría ser el estudio de las empresas consultadas por el buscador, para ver en qué tipo de empresas se fijan los corredores que más venden.

Actualmente en IMeureka se está trabajando en la digitalización de documentos de pólizas con CNN de reconocimiento de imágenes y NLP para agilizar el proceso y aportar otra herramienta de tipo suscripción que aporte valor a los corredores.

# Capítulo 8. BIBLIOGRAFÍA

[1] Imeureka.com. 2021. *¿Qué es IMeureka? | Imeureka.com*. [online] Available at: <https://www.imeureka.com/quienes-somos>.
Introducción de la empresa similar al TFM realizado por el mismo autor DISEÑO & BACKEND DE UNA APP CONCEBIDA PARA SIMPLIFICAR LA GESTIÓN DE RIESGOS DE EMPRESAS EN LA COMUNICACIÓN DE SINIESTROS A SU COMPAÑÍA ASEGURADORA. Visible en el repositorio de TFM de la universidad pontifica comillas

[2] Imeureka.com. 2021. *¿Qué es IMeureka? | Imeureka.com*. [online] Available at: <https://www.imeureka.com/quienes-somos>.

[3] Sedecatastro.gob.es. 2021. *Sede Electrónica del Catastro - Inicio*. [online] Available at: <https://www.sedecatastro.gob.es/> [Accessed 7 July 2021].

[4] Contrataciondelestado.es. 2021. *Plataforma de Contratación del Sector Público*. [online] Available at: <https://contrataciondelestado.es/wps/portal/licitaciones> [Accessed 7 July 2021].

[5] dgt, w., 2021. *Consulta si tienes una multa y no te hemos podido localizar (TEU)*. [online] Sede.dgt.gob.es. Available at: <https://sede.dgt.gob.es/es/multas/consulta-multa-tablon-edictal-unico/> [Accessed 7 July 2021].

[6] Infocif.es. 2021. *Información GRATIS de Empresas Españolas*. [online] Available at: <http://www.infocif.es/> [Accessed 7 July 2021].

[7] PowerData, G., 2021. *GDPR: Lo que debes saber sobre el reglamento general de protección de datos*. [online] Powerdata.es. Available at: <https://www.powerdata.es/gdpr-proteccion-datos> [Accessed 7 July 2021].

[8] europapress.es. 2021. *El valor de las insurtech europeas se multiplica por cinco desde 2016, hasta 23.000 millones, según Dealroom*. [online] Available at: <https://www.europapress.es/economia/finanzas-00340/noticia-valor-insurtech-europeas-multiplica-cinco-2016-23000-millones-dealroom-20210627130537.html> [Accessed 7 July 2021].

[9] Selenium.dev. 2021. *SeleniumHQ Browser Automation*. [online] Available at: <https://www.selenium.dev/> [Accessed 7 July 2021].

[10] GitHub. 2021. *php-webdriver/php-webdriver*. [online] Available at: <https://github.com/php-webdriver/php-webdriver> [Accessed 7 July 2021].

[11] Postman.com. 2021. [online] Available at: <https://www.postman.com/> [Accessed 16 June 2021].

[12] Es.wikipedia.org. 2021. JSON - Wikipedia, la enciclopedia libre. [online] Available at: <https://es.wikipedia.org/wiki/JSON#:~:text=JSON%20(acr%C3%B3nimo%20de%20JavaScript%20Object,para%20el%20intercambio%20de%20datos.&text=Una%20de%20las%20supuestas%20ventajas,sint%C3%A1ctico%20(parser)%20para%20%C3%A9l.> [Accessed 16 June 2021].

[13] Git-scm.com. 2021. *Git*. [online] Available at: <https://git-scm.com/> [Accessed 7 July 2021].

[14] Bitbucket. 2021. *Bitbucket | The Git solution for professional teams*. [online] Available at: <https://bitbucket.org/product/> [Accessed 7 July 2021].

[15] Jetbrains.com. 2021. PhpStorm: el IDE rápido e inteligente para programación en PHP de JetBrains. [online] Available at: <https://www.jetbrains.com/es-es/phpstorm/>

[16] Es.wikipedia.org. 2021. XAMPP - Wikipedia, la enciclopedia libre. [online] Available at: <https://es.wikipedia.org/wiki/XAMPP> [Accessed 16 June 2021].

[17] Herranz, A., 2021. *España, el país en el que más multas GDPR se pusieron en el primer trimestre de 2021*. [online] Xataka.com. Available at: <https://www.xataka.com/pro/espana-pais-que-multas-gdpr-se-pusieron-primer-trimestre-2021#:~:text=Linkedin-,Espa%C3%B1a%2C%20el%20pa%C3%ADs%20en%20el%20que%20m%C3%A1s%20multas%20GDPR%20se,el%20primer%20trimestre%20de%202021&text=Durante%20los%20tres%20primeros%20meses,33%2C61%20millones%20de%20euros.> [Accessed 7 July 2021].

[18] Zyte (formerly Scrapinghub) #1 Web Scraping Service. 2021. *GDPR Compliance For Web Scrapers: The Step-by-step Guide*. [online] Available at: <https://www.zyte.com/blog/web-scraping-gdpr-compliance-guide/> [Accessed 7 July 2021].

[19] En.wikipedia.org. 2021. *HiQ Labs v. LinkedIn - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/HiQ_Labs_v._LinkedIn> [Accessed 7 July 2021].

[20] Conceptos Jurídicos. 2021. *Delitos contra la Propiedad Intelectual: concepto y requisitos*. [online] Available at: <https://www.conceptosjuridicos.com/delitos-contra-la-propiedad-intelectual/#:~:text=a%20una%20persona.-,Un%20delito%20contra%20la%20propiedad%20intelectual%20consiste%20en%20reproducir%2C%20plagiar,272%20del%20C%C3%B3digo%20Penal%20espa%C3%B1ol.> [Accessed 7 July 2021].

[21] Análisis del mercado asegurador por Mapfre https://documentacion.fundacionmapfre.org/documentacion/publico/i18n/catalogo_imagenes/grupo.cmd?path=1099983

[22] Es.wikipedia.org. 2021. *Índice de Herfindahl - Wikipedia, la enciclopedia libre*. [online] Available at: <https://es.wikipedia.org/wiki/%C3%8Dndice_de_Herfindahl#:~:text=El%20%C3%8Dndice%20de%20Herfindahl%20o,muy%20concentrado%20y%20poco%20competitivo.> [Accessed 7 July 2021].

[23] Catastro.meh.es. 2021. [online] Available at: <http://www.catastro.meh.es/ws/webservices_catastro.pdf> [Accessed 7 July 2021].

[24] Ovc.catastro.meh.es. 2021. *Callejero de la sede electrónica del catastro. Web Service*. [online] Available at: <https://ovc.catastro.meh.es/ovcservweb/ovcswlocalizacionrc/ovccallejero.asmx> [Accessed 7 July 2021].

[25] Ovc.catastro.meh.es. 2021. *OVCCoordenadas Web Service*. [online] Available at: <https://ovc.catastro.meh.es/ovcservweb/OVCSWLocalizacionRC/OVCCoordenadas.asmx> [Accessed 7 July 2021].

[26] Infoconcurso.com. 2021. *Licitaciones y concursos públicos en España*. [online] Available at: <https://www.infoconcurso.com/?gclid=Cj0KCQjw24qHBhCnARIsAPbdtlIjHTWcNZmptWpZUgjna9FGSiXjZjR-Cn1_LrbKwTkVlR5qJ1DzVWAaAg3WEALw_wcB> [Accessed 7 July 2021].

[27] [4]

[28] Einforma.com. 2021. *Información de Empresas Españolas | eInforma*. [online] Available at: <https://www.einforma.com/> [Accessed 7 July 2021].

[29] Emailsgo.es. 2021. *Listado de Empresas Españolas - Email Empresas - Guia de Base de Datos de Empresas*. [online] Available at: <https://www.emailsgo.es/goo_basesdedatos.htm> [Accessed 7 July 2021].

[30] Email Databases and Email Addresses List. 2021. *1-click Download: Spain email business Database*. [online] Available at: <https://bigdatalowcost.com/best-spain-email-business-database/> [Accessed 7 July 2021].

[31] Bases de datos de empresas y listados de empresas. 2021. *Nos actualizamos!!! Estrenamos la versión 2020 de bases de datos y listados de empresas con email*. [online] Available at: <https://www.centraldecomunicacion.es/tienda/base-datos-empresas/> [Accessed 7 July 2021].

[32] Bases de datos de empresas y listados de empresas. 2021. *Nos actualizamos!!! Estrenamos la versión 2020 de bases de datos y listados de empresas con email*. [online] Available at: <https://www.centraldecomunicacion.es/tienda/base-datos-empresas/> [Accessed 7 July 2021].

[33] datacertia. 2021. *Listados de Empresas optimizados para captación de clientes*. [online] Available at: <https://www.datacertia.com/> [Accessed 7 July 2021].

[34] Databusiness. 2021. *Comprar Bases de Datos de Empresas Españolas 2021*. [online] Available at: <https://databusiness.es/> [Accessed 7 July 2021].

[35] 2021. [online] Available at: <https://www.publiemail.es/bases-datos/bases-de-datos-toda-espana> [Accessed 7 July 2021].

[36] BoldData. 2021. *Company Database Spain – Buy dataset of 5.087.915 Spanish businesses*. [online] Available at: <https://bolddata.nl/en/database/spain/> [Accessed 7 July 2021].

[37] Censo.camara.es. 2021. *Censo Nacional de Empresas*. [online] Available at: <https://censo.camara.es/> [Accessed 7 July 2021].

[39] Infocif.es. 2021. *Infocif*. [online] Available at: <http://www.infocif.es/buscador/#/> [Accessed 7 July 2021].

[40] Free-proxy-list.net. 2021. *Free Proxy List - Just Checked Proxy List*. [online] Available at: <https://free-proxy-list.net/> [Accessed 7 July 2021].

[41] 2021. [online] Available at: <https://developers.google.com/custom-search/v1/overview> [Accessed 7 July 2021].

[42] Sullivan, D., 2021. *Google: Bing Is Cheating, Copying Our Search Results*. [online] Search Engine Land. Available at: <https://searchengineland.com/google-bing-is-cheating-copying-our-search-results-62914> [Accessed 7 July 2021].

[43] [closed], I., 2021. *Is it ok to scrape data from Google results?*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/22657548/is-it-ok-to-scrape-data-from-google-results> [Accessed 7 July 2021].

[44] Es.wikipedia.org. 2021. *MX (registro) - Wikipedia, la enciclopedia libre*. [online] Available at: <https://es.wikipedia.org/wiki/MX_(registro)> [Accessed 7 July 2021].

[45] CodexWorld. 2021. *Verify Email Address and Check if Email is Real using PHP - CodexWorld*. [online] Available at: <https://www.codexworld.com/verify-email-address-check-if-real-exists-domain-php/> [Accessed 7 July 2021].

[46] 2021. [online] Available at: <https://developers.google.com/maps/documentation/geocoding/overview> [Accessed 7 July 2021].

# ANEXO ODS

Este proyecto colabora en el ODS 8 para el trabajo decente y el crecimiento económico. La parte de crecimiento económico mediante el aumento de eficiencias que aumentan los beneficios de todos los usuarios implicados, y el trabajo decente mediante la disminución de trabajo repetitivo que tienen que realizar los intermediadores, aumentando su rentabilidad y su calidad de vida.

También colabora en el apartado 9 con la innovación, ya que es puntera en muchos ámbitos del sector asegurador.

Finalmente, también contribuye con el ODS 15 ya que al digitalizar el sector asegurador se disminuye el uso de papel, protegiendo así los bosques del plantea.

# ANEXO: JSON EJEMPLO EMPRESA

## JSON EMPRESA EJEMPLO (RECORTADO)


```
{
  "IdEmpresa": 6582004,
  "IdFichaEmpresa": 0,
  "RazonSocial": "NESTLE GLOBAL SERVICES SPAIN SL",
  "CIF": "B67038240",
  "FechaGeneracion": "0001-01-01T00:00:00",
  "CnaeInfo": {
    "Cnae3_text": "6920. Actividades de contabilidad, teneduría de libros,
auditoría y asesoría fiscal"
  },
  "Direccion": "CL CLARA CAMPOAMOR Num.2",
  "Comunidad": "Cataluña",
  "Localidad": "Esplugues de Llobregat",
  "Provincia": "Barcelona",
  "Codigo_Postal": "08950",
  "FechaConstitucionOrigen": "27/07/2017",
  "Id_EmpresaIndice": 0,
  "EstadoEspecialEmpresa": 0,
  "IndustriaDescripcion": "Consultoría empresarial y otros",
  "CuentasAnuales": [
    {
      "ID_Cuentas": 734618,
      "Ejercicio": 2018,
      "TipoCuestionario": "Normal",
      "TipoCuentas": 1,
      "TipoCuentaAnual": 1,
      "Original": true,
      "Idioma": "Castellano",
      "FechaDescarga": "2019-12-05T15:36:29.777",
      "Origen": "EXTERNO",
      "FechaInicioEjercicio": "2018-01-01T00:00:00",
      "FechaFinEjercicio": "2018-12-31T00:00:00"
    },
    {
      "ID_Cuentas": 734621,
      "Ejercicio": 2017,
      "TipoCuestionario": "Normal",
      "TipoCuentas": 1,
      "TipoCuentaAnual": 1,
      "Original": false,
      "Idioma": "Castellano",
      "FechaDescarga": "2019-12-05T15:36:52.777",
      "Origen": "EXTERNO",
      "FechaInicioEjercicio": "2018-01-01T00:00:00",
      "FechaFinEjercicio": "2018-12-31T00:00:00"
```

```
    }
  ],
  "TipoCuentasAnuales": [
    {
      "Numero": 1
    }
  ],
  "TipoCuentas": 1,
  "UltimaCuentaAnual": {
    "Ejercicio": 2018,
    "ImporteNetoCifraDeNegocioICIF": 49972000,
    "ResultadoEjercicio": 1108000,
    "SumTotalEmpleados": 477
  },
  "CuentasDisponibles": [
    {
      "IdPartida": 0,
      "Ejercicio": 2017
    }
  ],
  "Cargos": [
    {
      "Id_CargosClasificados": 73667074,
      "IdEmpresa": 6582004,
      "FechaActo": "2019-11-21T00:00:00",
      "FechaIntegracion": "2020-08-10T14:14:08.607",
      "Nombre": "KPMG AUDITORES SL",
      "CargoEspejo": "otro",
      "SubGrupo": 32,
      "Cargo": "AUDIT.CUENT.:",
      "EstadoActivo": 1,
      "NumEmpresasVinculadas_NombreCompleto_N": 6953,
      "NumEmpresasVinculadas_NombreCompleto_C": 1003,
      "NumEmpresasVinculadas_NombreCompleto": 7910
    },
    {
      "Id_CargosClasificados": 73667065,
      "IdEmpresa": 6582004,
      "FechaActo": "2017-08-22T00:00:00",
      "FechaRenuncia": "2018-08-02T00:00:00",
      "FechaIntegracion": "2020-08-10T14:14:08.607",
      "Nombre": "LAURENT DEREUX",
      "CargoEspejo": "otro",
      "SubGrupo": 55,
      "Cargo": "Consejero",
      "EstadoActivo": 0,
      "NumEmpresasVinculadas_NombreCompleto_N": 1,
      "NumEmpresasVinculadas_NombreCompleto_C": 4,
      "NumEmpresasVinculadas_NombreCompleto": 5
    },
    {
      "Id_CargosClasificados": 73667067,
      "IdEmpresa": 6582004,
```

```
    "FechaActo": "2017-08-22T00:00:00",
    "FechaRenuncia": "2020-08-10T00:00:00",
    "FechaIntegracion": "2020-08-10T14:14:08.607",
    "Nombre": "OLALDEFERNANDEZ DE BETOÑO JORGE",
    "CargoEspejo": "otro",
    "SubGrupo": 55,
    "Cargo": "Consejero",
    "EstadoActivo": 0,
    "NumEmpresasVinculadas_NombreCompleto_N": 8,
    "NumEmpresasVinculadas_NombreCompleto_C": 6,
    "NumEmpresasVinculadas_NombreCompleto": 10
  },
  {
    "Id_CargosClasificados": 73667069,
    "IdEmpresa": 6582004,
    "FechaActo": "2017-08-22T00:00:00",
    "FechaRenuncia": "2018-03-16T00:00:00",
    "FechaIntegracion": "2020-08-10T14:14:08.607",
    "Nombre": "TERENCE MALCOLM STACEY",
    "CargoEspejo": "otro",
    "SubGrupo": 55,
    "Cargo": "Consejero",
    "EstadoActivo": 0,
    "NumEmpresasVinculadas_NombreCompleto_C": 1,
    "NumEmpresasVinculadas_NombreCompleto": 1
  },
  {
    "Id_CargosClasificados": 73667073,
    "IdEmpresa": 6582004,
    "FechaActo": "2018-08-02T00:00:00",
    "FechaIntegracion": "2020-08-10T14:14:08.607",
    "Nombre": "JACQUES ALEXANDRE REBER",
    "CargoEspejo": "otro",
    "SubGrupo": 114,
    "Cargo": "Presidente",
    "EstadoActivo": 1,
    "NumEmpresasVinculadas_NombreCompleto_N": 2,
    "NumEmpresasVinculadas_NombreCompleto_C": 1,
    "NumEmpresasVinculadas_NombreCompleto": 3
  },
  {
    "Id_CargosClasificados": 73667068,
    "IdEmpresa": 6582004,
    "FechaActo": "2017-08-31T00:00:00",
    "FechaRenuncia": "2020-08-10T00:00:00",
    "FechaIntegracion": "2020-08-10T14:14:08.607",
    "Nombre": "OLALDE FERNANDEZ DE BETOÑO JORGE",
    "CargoEspejo": "vicepresidente",
    "SubGrupo": 142,
    "Cargo": "Vicepresidente",
    "EstadoActivo": 0,
    "NumEmpresasVinculadas_NombreCompleto_N": 8,
    "NumEmpresasVinculadas_NombreCompleto_C": 6,
```

```
      "NumEmpresasVinculadas_NombreCompleto": 10
    }
  ],
  "CargosFiltrados": [
    {
      "Valor": "otro"
    },
    {
      "Valor": "vicepresidente"
    }
  ],
  "FechaConstitucion": "2017-07-27T00:00:00",
  "InformacionFinanciera": [
    {
      "ID_Partida": 0,
      "TipoCuentas": 1,
      "Codigo": 30000,
      "ValorEnEuros": 3000,
      "Ejercicio": 2017
    },
    {
      "ID_Partida": 0,
      "TipoCuentas": 1,
      "Codigo": 21100,
      "ValorEnEuros": 3000,
      "Ejercicio": 2017
    },
    {
      "ID_Partida": 0,
      "TipoCuentas": 1,
      "Codigo": 21000,
      "ValorEnEuros": 3000,
      "Ejercicio": 2017
    },
    {
      "ID_Partida": 0,
      "TipoCuentas": 1,
      "Codigo": 20000,
      "ValorEnEuros": 3000,
      "Ejercicio": 2017
    },
    {
      "ID_Partida": 0,
      "TipoCuentas": 1,
      "Codigo": 12400,
      "ValorEnEuros": 3000,
      "Ejercicio": 2017
    },
    {
      "ID_Partida": 0,
      "TipoCuentas": 0,
      "Codigo": 1,
      "ValorEnEuros": 49972000,
```

```
      "Ejercicio": 2018
    }
  ]
}
```

# ANEXO DATOS CATASTRO:

| Ref_catastral | Uso | Tamano | Antiguedad | Tipo | Municipio | Provincia |
|---|---|---|---|---|---|---|
| 5555618VK4755F0001 | Almacen-Estacionami | 102 | 1970 | ALMACEN | MADRID | MADRID |
| 5555618VK4755F0002 | Almacen-Estacionami | 75 | 1970 | ALMACEN | MADRID | MADRID |
| 5555618VK4755F0003 | Residencial | 53 | 1970 | VIVIENDA | MADRID | MADRID |
| 5555618VK4755F0004 | Residencial | 81 | 1970 | VIVIENDA | MADRID | MADRID |
| 5555618VK4755F0005 | Residencial | 86 | 1970 | VIVIENDA | MADRID | MADRID |

| Codigo_mun | Codigo_provi | CP | ID_master_t | Kilometro | Bloque | Planta | Escalera | Puerta |
|---|---|---|---|---|---|---|---|---|
| 900 | 28 | 28017 | 23 | | | -1 | | A |
| 900 | 28 | 28017 | 23 | | | -1 | | B |
| 900 | 28 | 28017 | 23 | | | 00 | | A |
| 900 | 28 | 28017 | 23 | | | 00 | | B |
| 900 | 28 | 28017 | 23 | | | 01 | | A |
| | | | | | | | | |

| Via | Numero | Latitude | Longitude | Extra_info |
|---|---|---|---|---|
| GONZALO DE BERCEC | 10 | | | [{"Uso_principal":"ALI |
| GONZALO DE BERCEC | 10 | | | [{"Uso_principal":"ALI |
| GONZALO DE BERCEC | 10 | | | [{"Uso_principal":"VIV |
| GONZALO DE BERCEC | 10 | | | [{"Uso_principal":"VIV |
| GONZALO DE BERCEC | 10 | | | [{"Uso_principal":"VIV |

[{"Uso_principal":"ALMACEN","Escalera":"","Planta":"-1","Puerta":"A","Superficie":"73"},{"Uso_principal":"ALMACEN","Escalera":"","Planta":"00","Puerta":"A1","Superficie":"27"},{"Uso_principal":"ELEMENTOS COMUNES","Escalera":"","Planta":"","Puerta":"","Superficie":"2"}]

# ANEXO: Petición catastro con múltiples inmuebles (con información redundante recortada):

```
<?xml version="1.0" encoding="utf-8"?>
<consulta_dnp xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.catastro.meh.es/">
    <control>
        <cudnp>51</cudnp>
    </control>
    <lrcdnp>
        <rcdnp>
            <rc>
                <pc1>9581103</pc1>
                <pc2>VK3798B</pc2>
                <car>0001</car>
                <cc1>E</cc1>
                <cc2>D</cc2>
            </rc>
            <dt>
                <loine>
                    <cp>28</cp>
                    <cm>79</cm>
                </loine>
                <cmc>900</cmc>
                <np>MADRID</np>
                <nm>MADRID</nm>
                <locs>
                    <lous>
                        <lourb>
                            <dir>
                                <cv>225</cv>
                                <tv>CL</tv>
                                <nv>ALMANSA</nv>
                                <pnp>110</pnp>
                            </dir>
                            <loint>
                                <pt>-1</pt>
                                <pu>A</pu>
                            </loint>
                            <dp>28040</dp>
                            <dm>9</dm>
                        </lourb>
                    </lous>
                </locs>
            </dt>
        </rcdnp>
        <rcdnp>
            <rc>
```

```
            <pc1>9581103</pc1>
            <pc2>VK3798B</pc2>
            <car>0002</car>
            <cc1>R</cc1>
            <cc2>F</cc2>
        </rc>
        <dt>
            <loine>
                <cp>28</cp>
                <cm>79</cm>
            </loine>
            <cmc>900</cmc>
            <np>MADRID</np>
            <nm>MADRID</nm>
            <locs>
                <lous>
                    <lourb>
                        <dir>
                            <cv>225</cv>
                            <tv>CL</tv>
                            <nv>ALMANSA</nv>
                            <pnp>110</pnp>
                        </dir>
                        <loint>
                            <pt>00</pt>
                            <pu>02</pu>
                        </loint>
                        <dp>28040</dp>
                        <dm>9</dm>
                    </lourb>
                </lous>
            </locs>
        </dt>
    </rcdnp>
    <rcdnp>
        <rc>
            <pc1>9581103</pc1>
            <pc2>VK3798B</pc2>
            <car>0051</car>
            <cc1>G</cc1>
            <cc2>W</cc2>
        </rc>
        <dt>
            <loine>
                <cp>28</cp>
                <cm>79</cm>
            </loine>
            <cmc>900</cmc>
            <np>MADRID</np>
            <nm>MADRID</nm>
            <locs>
                <lous>
                    <lourb>
```

```
                                <dir>
                                    <cv>225</cv>
                                    <tv>CL</tv>
                                    <nv>ALMANSA</nv>
                                    <pnp>110</pnp>
                                </dir>
                                <loint>
                                    <pt>00</pt>
                                    <pu>8C</pu>
                                </loint>
                                <dp>28040</dp>
                                <dm>9</dm>
                            </lourb>
                        </lous>
                    </locs>
                </dt>
            </rcdnp>
        </lrcdnp>
</consulta_dnp>
```

# Anexo licitaciones:

Ejemplo datos almacenados de licitaciones

| ID_licitacion | Codigo_expediente | ID_organo_c | ID_master_t | Estado_licita | Objeto_del_contrato | Presupuesto_ |
|---|---|---|---|---|---|---|
| 1 | 2021/C003/000003 | 205 | 2 | 1 | la contratación de la c | 552590,89 |
| 2 | Exp.Núm. FURV2021- | 165 | 2 | 1 | L'objecte del contract | 74520 |
| 3 | CC2021/25 | 198 | 2 | 1 | Servicios de seguro de | 1150000 |
| 4 | DNA 27/2021 | 199 | 2 | 1 | Póliza de Vida y Accid | 1300000 |
| 5 | Exp C21/08 | 202 | 2 | 1 | Contratación del progr | 1424000 |

| Valor_estim | Codigo_CPV | Lugar_de_ejecucion | Procedimiento_de_co | Fecha_fin_presentaci | ID_producto |
|---|---|---|---|---|---|
| 2762954,45 | 66510000-Servicios de | España - Lugo | Abierto | 2021-03-31 14:00:01 | 1;25;17;5;6;7;9;23 |
| 163944 | 66512200 Servicios de | España - Cataluña | Abierto | 2021-03-08 11:00:01 | |
| 1430000 | 66516000-Servicios de | España - Asturias | Abierto | 2021-03-12 23:59:01 | 25;1;5;6;7;9;23 |
| 1300000 | 66511000-Servicios de | España - Madrid | Abierto | 2021-03-31 23:59:01 | 13;10;11;12 |
| 3560000 | 66516000-Servicios de | España - Córdoba | Abierto | 2021-03-31 14:00:01 | |

| Url | ID_broker | Enviado | Fec_intro |
|---|---|---|---|
| https://contratacionde | 0 | | 2021-03-01 11:00:18 |
| https://contratacionde | 0 | | 2021-03-01 11:00:58 |
| https://contratacionde | 0 | | 2021-03-01 11:01:53 |
| https://contratacionde | 0 | | 2021-03-01 11:02:36 |
| https://contratacionde | 0 | | 2021-03-01 11:03:10 |