Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Grado en Ingeniería Informática (10II)

Trabajo Fin de Grado

Extracción y Normalización de Convenios Colectivos

Autor: Jaime Bautista Salinero
Tutor(a): Víctor Rodríguez Doncel

Madrid, 01 - 20

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Grado en Ingeniería Informática (10II)

Título: Extracción y Normalización de Convenios Colectivos

01 - 20

Autor: Jaime Bautista Salinero
Tutor: Víctor Rodríguez Doncel
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

«Aquí va el resumen del TFG. Extensión máxima 2 páginas.»

Abstract

«Abstract of the Final Degree Project. Maximum length: 2 pages.»

Tabla de contenidos

Capítulo 1

Introducción

1.1. Contexto

En el primer trimestre del 2019 es calculaban entre 45 y 50 millones de habitantes en España[1], con una tasa de actividad del 58 % y con un total de 19 millones de personas afiliadas a la Seguridad Social de los que 14 están en regimén general[2]. De estos 14 millones de empleados, todos y cada uno de ellos tiene un convenio colectivo asociado a la labor que desempeña.

Estos convenios colectivos, regulan los aspectos de la relación laboral entre las empresas y sus empleados, marcando los derechos y deberes de unos y de otros[3]. Dichos convenios colectivos se pueden encontrar en los Boletines Oficiales de las 58 Autoridades Laborales existentes, desde la Estatal hasta las Provinciales pasando por la Autonómicas.

El acceso a estos boletines se puede realizar en la mayoría de casos a través de los portales web que las propias Autoridades mantienen, no obstante, el acceso a los mismos o la búsqueda de información en ocasiones resulta en una ardua tarea. Por ello, el Ministerio de Trabajo e Inmigración estableció en 2007 que los ciudadanos tuviesen a partir de 2010 el derecho de realizar por médios electrónicos las mismas gestiones que se pueden llevar a cabo de forma presencial[4], contemplando el registro y depósito de todos los convenios colectivos en una base de datos centralizada.

Para atajar esta problemática, desde este ministerio se creo el portal denominado REGCON[1], Registro de Convenios y Acuerdos Colectivos, donde las distintas Autoridades Laborales deben introducir toda la información relativa a cualquier creación, o modificación relacionado con los convenios colectivos, calendarios laborales o revisiones salariales entre otras, con su fecha de validez y enlace a la documentación en cuestión.

Este portal facilita en gran medida la búsqueda de los distintos convenios colectivos, en base a tipos de trámites, vigencia, naturaleza a Autoridades Laborales. No obstante, la extracción de información concreta y la relación que pueden tener unos convenios colectivos con otros no se solventa desde este portal, no siendo más que un directorio de todos los convenios existentes dentro de nuestras fronteras.

[1] https://expinterweb.empleo.gob.es/regcon/

1.2. Motivación

En la sociedad actual, donde el acceso y la necesidad de información se hace más notorio debido al crecimiento de las nuevas tecnologías y uso de los smartphones en los últimos años, existe una necesidad de poder consultar de manera rápida y eficaz cierta información sobre los derechos que nos conciernen como trabajadores así como estar al tanto de las actualizaciones que se ejercen sobre los mismos.

A pesar de la existencia de un estándar europeo de identificación y descripción de la normativa publicada en los diarios y las bases de datos oficiales, que pemite acceder a la legislación en un formato normalizado[5] y la creciente publicación de datos por parte de las Administraciones en los portales de datos abiertos, existe una gran carencia en la implementación de este estándar en los convenios colectivos.

Es por ello que existe una necesidad, que conlleva el estudio y la indexación de los convenios colectivos y el acesso a través de los distintos dispositivos de los que los trabajadores disponen, a cierta información en el momento que la requieran para poder ejercer sus derechos, así como para estar al tanto de los deberes que les acomete.

1.3. Objetivo

Mediante este trabajo, se prentende estudiar la accesibilidad a los convenios colectivos por parte de las distintas Autoridades Laborales, así como de los distintos formatos e idiomas en que se publican.

Se establecerá una herramienta capaz de buscar los convenios indexados desde el portal REGCON y se procederá a su extracción, normalizado e indexado, para posteriormente permitir el acceso a los mismos a través de un motor de búsqueda que permita cualquier término. Se hará una propuesta de presentación web, así como de traducción automática de los textos a otros idiomas.

Más importante que proceder a la codificación del sistema por completo, se cree elemental el diseño, tanto de la arquitectura física del sistema como del software, permitiendo que en el futuro sea un sistema abierto en el que se permitan aportaciones externas o integraciones con otros sistemas sin que sea un bloque monolítico.

En este proyecto se usarán diversas tecnologías, por un lado para que sirva de aprendizaje y por otro para que sea un sistema con posibilidad de cambio y adaptabilidad a nuevos entornos y tecnologías y se pueda estudiar la interacción entre los mismos, y descartar, si fuese necesario ciertos desarollos.

1.4. Estructura del documento

A continuación se detallan de manera superficial el contenido de los distintos capítulos, permitiendo al lector dirigirse a la parte que consideré más relevante en su lectura:

- En el capítulo 2, se presentarán las técnicas que se emplearán para llevar a cabo el proyecto, estudiando distintas tecnologías presentes en la actualidad.
- En el capítulo 3, se realizará un análisis de los requerimientos del sistema y se detallará su diseño en la mayor medida posible, justificando las decisiones

tomadas durante el mismo.

- En el capítulo 4, nos adentraremos en la implementación del sistema, lenguajes, librerías y técnicas de desarollo empleadas, así como la publicación del sistema al exterior y la solución obtenida.
- En el capítulo 5, se detallará la forma en la que se ha gestionado el proyecto, con las tareas realizadas, estimaciones y tiempo real empleado, así como el presupuesto empleado si procediese.
- En el capítulo 6, se presentarán los futuros trabajos posibles partiendo del estado del sistema en el momento de finalización del proyecto.
- Finalmente, en el capítulo 7, se realizará un ejercicio de restrospectiva y análisis del trabajo realizado.
- Al final del documento podremos encontrar los diversos anexos referenciados a lo largo del texto.

Capítulo 2

Estado del Arte

2.1. Web Spidering y Harvesting

El Web Spidering o Araña web, es un programa informático creado para inspeccionar el contenido de las páginas para analizar o extraer información, descargarla en forma de 'instantánea', o indexarla. Este programa tendrá como input una lista de URLs, llamadas semillas, por las que comenzará a rastrear, realizando las acciones para las que ha sido programada.

A medida que la araña navega a través de las páginas o portales web va almacenando las nuevas URL identificadas y procesando su contenido en base a distintas políticas[6]. Entre estas políticas encontramos:

Política de selección Selecciona que páginas o contenido descargar.

Política de re-visita Mediante la que se establece cuando se debe de acceder de nuevo a las páginas para detectar cambios.

Política de cortesía A través de la que se indica los medidas adoptadas para evitar la caida de los servidores por sobrecarga u otras causas, así como de la infraestructura de red.

Política de paralelización Que indica como deben actuar los distintos procesos de la misma araña para maximar la velocidad de descarga y no repetir el mismo trabajo.

Otro aspecto a tener en cuenta a la hora de crear una araña web es la seguridad que los dueños pueden aplicar a sus páginas web mediante el fichero robots.txt, permitiendo solo ciertos ciertos User-Agents, propiedades de una petición HTTP mediante la que se identifican navegadores y arañas. No obstante, al ser una simple cadena de caracteres, no resulta complicada la 'suplantación' de identidad de cualquier navegador, aunque esto no rige las normas establecidas de cortesía para arañas web.

En cuanto al Web harvesting se trata de la extracción de cierto contenido de páginas web, a una base de datos central para su posterior consulta o análisis. Por norma general esta extracción se realiza a empleando arañas web. Este contenido puede ser procesado, parseado o copiado y usado para técnicas de data mining.

2.2. Bases de datos documentales

El modelo relacional de una base de datos fue inventado en una época en la que el manejo de datos estaba orientado principalmente a aplicaciones administrativas. No obstante, en el panorama actual nos encontramos con diversos tipos de datos susceptibles de almacenar; datos semi-setructurados y desestructurados, datos continuos, de sensores o gráficas, por dar unos ejemplos[7].

Este modelo relacional fue propuesto para gestionar y almacenar datos estructurados, mientras que otras soluciones, como las bases de datos NoSQL surgidas recientemente, estan pensadas para almacenar otros tipos de datos no estructurados como documentos o flujos de datos.

En la actualidad, y dado que diferentes aplicaciones tienen distintos requisitos, el manejo de datos se ha vuelto cada vez más complejo, separando la persistencia de las mismas en base a los distintos interes que existen, pudiendo combinar las distintas soluciones.

En cuanto a la forma de almacenar de las distintas bases de datos, las bases de datos relacionales, almacena su información en tablas, compuestas de columnas o atributos, y a su vez en tuplas o registros. Los atributos para las tuplas de una tabla comparten tipo de datos y longitud máxima entre otras características. En las bases de datos orientadas a documentos, compuestas de colecciones donde se almacenan los documentos, compuestos por distintos atributos sin necesidad de compartir sus características entre unos documentos y otros, aportando un mayor grado de flexibilidad con respecto a las relacionales.

Entre las base de datos relacionales más usadas actualmente se encuentran Oracle, MySQL, Microsoft SQL Server y PostgreSQL. Con respecto a las bases de datos orientadas a documentos tenemos MongoDB, Couchbase y CouchDB.

2.3. Motores de Búsqueda

Como complemento a modelos de persistencia se pueden emplear motores de búsqueda, que indexan la información introducida de forma que al realizar una búsqueda, dicho motor devuelve una lista de aciertos o hits con unos pesos según la relevancia de la información con los términos de búsqueda. Esto significa que sirve como interfaz o punto de partida a la hora de consultar elementos de interés.

Existen distintos tipos de sintáxis a la hora de realizar la consulta al motor de búsqueda. Podemos encontrar motóres que acepten el lenguaje natural, otros que precisen una sintaxis común, e incluso construir peticiones en base a documentos o imágenes.

Para que un motor de búsqueda funcione según nuestros requerimientos será imprescindible comprender como funciona, tanto el indexado como lo recuperación de resultados, acomodando las distintas características y elementos a nuestro propósito.

Actualmente los dos motores de búsqueda más conocidos son ElasticSearch, que además provee de un sistema de persistencia documental, y Solr, de la Apache Software Foundation. Estos dos motores están desarrollados en Java y basados en el proyecto Apache Lucene, con capacidad de indexado y búsqueda de texto completo.

2.4. Arquitectura REST

La arquitectura REST es un estilo de software que define unos estándares para la creación de servicios Web, empleando operaciones sin estado. Estas operaciones permiten que el sistema que se encuentra detrás de la resolución de estas pueda crecer y escalar, sin que ello requiera un rediseño de las mismass.

El protocolo más usado con esta arquitectura es HTTP, empleando las distintas operaciones de este protocolo. Estas operaciones permitirán interactuar con los servidores, desde la adquisición de información a la introducción o envío de datos así como el borrado, si el diseño lo permite. Un servicio Web que siga esta arquitectura con todas sus indicaciones es lo que se denomina RESTful.

Actualmente, existen multitud de librerias que se pueden emplear con la mayoría de los lenguajes de programación debido a la extensión de su uso en sistemas de todo el mundo.

2.5. Linked Data

Mediante el Linked Data se establece un 'lenguaje' a través del cual se pueden comunicar los programas informáticos entre sí. Se trata de establecer e identificar los distintos elementos que componen los datos, asignándoles un atributo previamente identificado.

El objetivo es poder interconectar los distintos datos entre sí, datos que hasta el momento solo se esperaba ser leidos por humanos, pero que las necesidades de nuevas tecnologías, ya sea de análisis o consulta de información, requiere poder extraer el máximo número de referencias sobre un dato, y con esto toda la información disponible del mismo.

Se trata de estructurar datos no estructurados, como pueden ser documentos que podemos encontrar en la web, identificando un conjunto de metadatos sobre los mismos.

Para el caso que nos ocupa, a la hora de diseñar nuestro repositorio de datos en base a estándares o normas basadas en la web semántica, nos permitirá devolver resultados a búsquedas más precisas, acotando, en ocasiones, el texto donde se encuetra la información más relevante a dicha consulta. Será posible también la asociación de las identidades identificadas en nuestros datos con otros conjuntos de información por medio de links, facilitando la búsqueda a los consumidores del servicio.

Por otro lado, es una forma de adelantarse a un futuro cada vez más cercano, como puede ser la automatización de bots, o programas automáticos, que intenten responder mediante el uso de lenguaje natural a consultas realizadas por el mismo medio, y para los que la respuesta deberá ser lo más acertada posible.

A la hora de diseñar este estándar existen distintos formatos de serialización, RDF, Turtle, JSON-LD, para el que deberemos analizar la mejor opción de modo que se acople al resto de la infraestructura del sistema. Deberemos decidir también si empleamos vocabularios ya creados para esta colección, como puede ser el vocabulario ELI o el vocabulario usado por otros proyectos como Lynx, o definiremos uno propio y acoplado a nuestra causística.

En cuanto al manejo de estos documentos existen frameworks que facilitan su publicación en un formato adecuado, como Apache Jena o RDF4J, ambos desarrollados en Java. No obstante, no es necesario el empleo de uno de estos frameworks para obtener un formato valido.

Capítulo 3

Análisis y Diseño del Sistema

3.1. Análisis de REGCON

Una de las primeras tareas realizadas ha sido analizar el portal web donde se recogen todos los convenios colectivos nacionales, llamado REGCON como diminutivo de Registro de Convenios Colectivos.

Los aspectos analizados en este portal han sido principalmente la manera en la que se pueden obtener los convenios colectivos con la máxima información posible y sobre esto, la posibilidad de automatizar este proceso para ejecutarlo de manera periódica. Este análisis ha sido exhaustivo, definiendo los pasos necesarios para su posterior codificación, desde las peticiones HTTP hasta el formato del fichero de resultados y los enlaces obtenidos de estos resultados.

A partir del fichero de resultados de buúsqueda que podemos obtener desde el portal, se han examinado los enlaces de todas y cada una de las Autoridades Laborales, identificando, antes de proceder a la codificación, los distintos errores que puedan surgir como pueden ser enlaces que no apuntan al recurso indicado o errores en la introducción. De esta manera se tendrán en cuenta todas estas causísticas a la hora del diseño e implementación del software.

En total se han analizado 58 Autoridades Laborales, para los que se ha rellenado los siguientes campos:

- Autoridad Laboral
- Ámbito Territorial
- Enlaces a páginas oficiales (Página de boletín, portal de datos abiertos)
- Idiomas
- Formato de ficheros
- Formatos de peticiones HTTP
- Comentarios adicionales

Debido a la extensión de este análisis, los resultados se recogerán en el ANEXO 1.

En cuanto a los resultados desde la búsqueda de los convenios colectivos por Autoridad Laboral en el portal de REGCON, hasta la descarga de los resultados e identificación de los campos descargados se muestran a continuación.

3.1.1. Procedimento de búsqueda de convenios

El primer paso a realizar será una petición HTTP de tipo POST, en la que se pasa a la url de búsqueda un parámetro

Al realizar una búsqueda en el portal REGCON para una autoridad laboral obtendremos una lista de trámites que podremos exportar a un Excel (insertar imágen búsqueda). Analizando el Excel y abriendo el fichero con un editor de texto y codificación ISO-8859-1, podemos ver claramente que tiene una estructura en XML, lo que nos permitirá extraer de forma sistemática toda la información contenida en la búsqueda realizada, ya que contiene los siguientes campos:

- Código del Acuerdo
- Denominación
- Tipo de Trámite
- Autoridad Laboral
- Fecha de Inscripción
- Vigencia Desde
- Vigencia Hasta
- URL Boletín

Sabiendo que podemos extraer la información de una consulta, el siguiente paso es analizar los pasos necesarios para poder automatizar las consultas de todas las Autoridades Laborales y extracción de su información. Analizando las peticiones HTTP mediante el Monitor de Red del navegador, en este caso Firefox, nos percatamos de que se trata de una petición empleando el método POST con un cuerdo de tipo formdata y que contiene, entre otros, un parámetro llamado 'autoridadLaboral' con valor numérico.

Por tanto el primer paso a realizar será la siguiente petición HTTP (Ver id de autoridades laborales en el anexo):

```
curl -X POST \
  https://expinterweb.empleo.gob.es/regcon/pub/consultaPublicaEstatal \
  -d autoridadLaboral=1

POST  https://expinterweb.empleo.gob.es/regcon/pub/consultaPublicaEstatal
Headers:
    Content-Type: application/x-www-form-urlencoded
Body:
    autoridadLaboral=1
```

3.1.2. Limpieza de los resultados

Tras obtener el fichero de resultados de manera automática deberemos limpiarlo y organizarlo de manera que su manejo sea mas sencillo. Por ello se han eliminado

todas las entidades XML, las líneas que no contienen URL de descarga y otras líneas que no contienen información relevante por medio de expresiones regulares.

El documento de resultados final contendrá una fila por cada acuerdo con sus valores separados por ';', similar a un document CSV. El resultado será el siguiente:

```
1 Codigo del Acuerdo; Denominacion; Tipo de Tramite; Autoridad Laboral; Fecha de Inscripcion;
      Vigencia Desde; Vigencia Hasta; URL Boletin;
```

3.1.3. Código CNAE

El CNAE[8][9] o Clasificación Nacional de Actividades Económicas es un código que, como su propio nombre indica, clasifica y agrupa las unidades productoras según su actividad, implementado para permitir la elaboración de estadísticas. Actualmente, se emplea la revisión CNAE-2009.

La estructura de este código, según la jerarquía de sus niveles, es la siguiente:

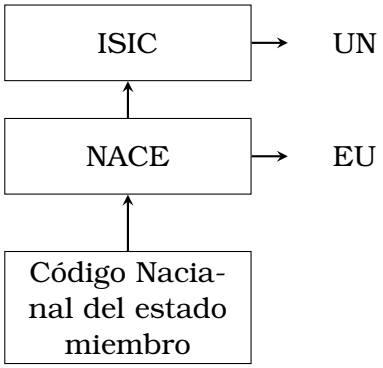
Nombre	Nº categorías	Identificación
Sección (Actividad)	21	código alfabético de una cifra
División	88	código alfanumérico de dos cifras
Grupo	272	código alfanumérico de tres cifras
Clase	629	código alfanumérico de cuatro cifras

Este código está basado en el NACE Rev. 2[10][11] (Nomenclatura estadística de actividades económicas de la Comunidad Europea), commpuesto de 21 secciones (A-U), 88 divisiones (01-99), 272 grupos (01.1-99.0), 615 clases (01.11-99.00). Se encuentra en línea con el ISIC Rev. 4, considerandose su versión europea. Existen actividades económicas en el NACE Rev. 2 que no aparecen en algunos de los estados miembros de la UE.o

Según se indica en la descripción de esta clasificación, se permite que los estados miembro usen una versión nacional basada en esta nomenclatura, y que encajen en el framework estructural y jerárquico de la misma[12], garantizando la perfecta coherencia de los códigos de los distintos paises miembros y el empleado por la UE.

Como mencionabamos anteriormente, el NACE Rev. 2 está basado en el ISIC Rev. 4 [13][14], Estándar industrial internacional de todas las actividades económicas, desarrollado por Naciones Unidas.

Es por ello, que la estructura de códigos de actividades económicas quedaría de la siguiente manera:



En el caso del CNAE coindicen todas las categorías, tanto en código como en título y contenido con las del NACE Rev. 2, exceptuando 14 clases que se han desagregado para cubrir las necesidades de la realidad nacional, siendo su correspodencia la siguiente[15]:

NACE Rev. 2	CNAE-2009
10.20	10.21; 10.22
10.41	10.43; 10.44
10.51	10.53; 10.54
35.11	35.15; 35.16; 35.17; 35.18; 35.19
41.20	41.21; 41.22
59.11	59.15; 59.16
59.13	59.17; 59.18
85.42	85.43; 85.44
87.30	87.31; 87.32
88.10	88.11; 88.12
91.01	91.05; 91.06

Hay que destacar que una unidad productiva puede llevar a cabo una o más actividades económicas, permitiendo la asignación de una o más categorías de la CNAE-2009.

El resto de valores de los distintos países de la UE se basan igualmente en la anterior nomenclatura europea para clasificar sus actividades económicas, de forma que sea sencillo las comparaciones. Como ejemplo de las nomenclaturas de algunos paises encontramos:

- Francia (NAF)[16]: Nomenclatura de Actividades Francesas, 5 caracteres.
- Italia (ATECO)[17]: Traducción italiana del NACE Rev. 2 adaptado a las características del sistema económico italiano. Adicionalmente contiene categoría y subcategoría (5 y 6 cifras respectivamente).
- Alemania (WZ)[18]: Además de la estructura correspondiente con el NACE Rev. 2, añade 839 subclases.
- Países Bajos (Dutch SBI 2008)[19]: los primeros 4 dígitos corresponden con los del NACE Rev. 2.

En la página del Eurostat[20], y más concretamente en la base de datos RAMON[21] (Gestión y Referencias de Nomenclaturas), podemos encontrar las metodologías estadísticas nacionales y accesos a los distintos institutos nacionales de estadística de todos los paises miembros.[22]

La lista de los códigos CNAE-2009 se puede obtener de la página del INE[8] que contiene toda la información del CNAE-2009 y la tabla con la correspondencia con el NACE Rev. 2 en cualquier idioma hablado en la UE se puede ver y descargar de la base de datos RAMON, eligiendo el idioma que nos interesa.

Se puede encontrar una correspondencia entre el código CNAE-2009 y NACE Rev. 2, con los títulos en idiomas español e inglés en el enlace:

- Excel: `https://upm365-my.sharepoint.com/:x:/g/personal/j_bautistas_alumnos_upm_es/EYZ1ZBfhKB9CvaYy4kiA9EUBdLOREd6vfsZ66US_FurZ1g?e=HU1zCE`

- CSV: `https://upm365-my.sharepoint.com/:x:/g/personal/j_bautistas_alumnos_upm_es/EWQbEmusyF5CnqiQvowATYYBDd1peIEZRPKfxM-Q2uu5Lw?e=ytAtJU`

- XML: `https://upm365-my.sharepoint.com/:u:/g/personal/j_bautistas_alumnos_upm_es/EdzZK-8CpfxLvGcD65Pw1BcBIeKPouupQL1tzuvKSEjGMQ?e=575Bl1`

3.1.3.1. Extracción del código CNAE de REGCON

La extracción del código CNAE-2009 correspondiente a los convenios colectivos de la página del REGCON[23] resulta más compleja que el procedimiento anterior, ya que existe una gran dependencia de las peticiones HTTP que hay que realizar con datos almacenados en las cookies del navegador, así como peticiones cuyo cuerpo es de tipo binario.

3.2. Descripción del Sistema

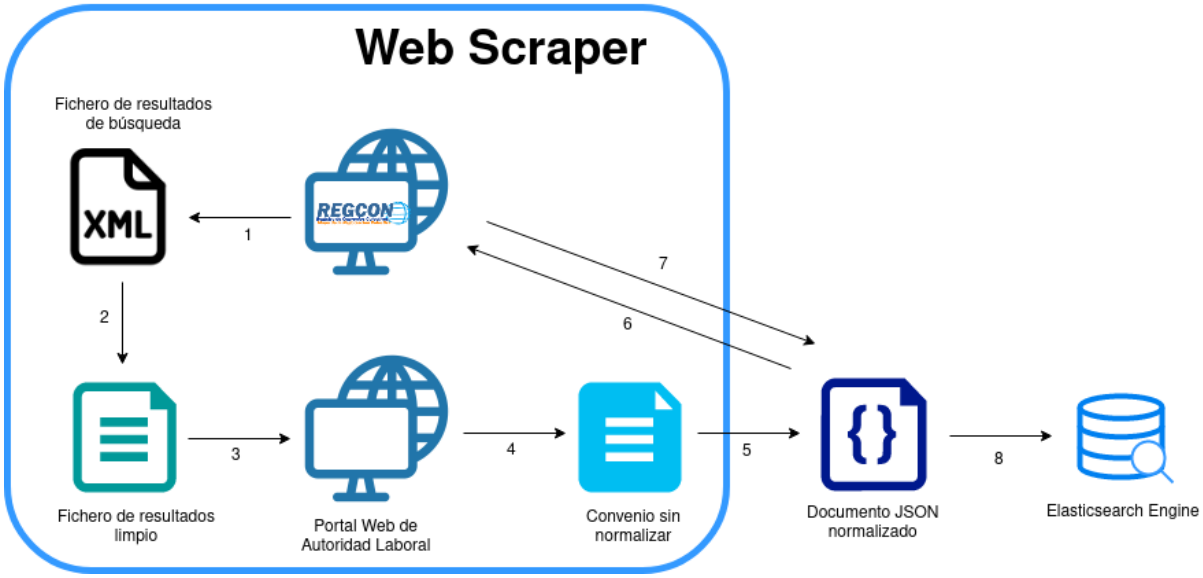
El sistema desempañará diversas funciones que permitirán cumplir los objetivos esperados del sistema. De alguna manera el elemento protagonista será el Documento, en este caso un documento de carácter laboral y más concretamente un convenio colectivo.

Tras una búsqueda en el portal de convenios colectivos llegaremos a un documento que puede ser de distintos formatos (XML, HTML, JSON, RDF, TURTLE o PDF), que se descargará y del que se extraerán los metadatos y el contenido identificando, en la medida de lo posible, las distintas estructuras que componen el texto, como pueden ser las distintas secciones o artículos, así como títulos, párrafos y tablas.

La información extraida se normalizará, creando un documento JSON que posteriormente se intoducirá en un servidor de búsqueda de Elasticsearch, que a su vez será el sistema de persistencia de dichos documentos.

En una etapa posterior, a este servidor de búsqueda se le podrá aplicar por encima una interfaz que permita la búsqueda en texto completo y la visualización de los resultados. Otra opción, podría ser la creación de una API entre la capa de búsqueda y de visualización, otorgando al operador del sistema más control sobre los permisos y acceso público y generando la posibilidad de explotar está información por otros medios como aplicaciones móviles, bots de mensajería instantánea o integraciones con otros sistemas.

A continuación, se muestra el proceso que sigue un documento por el sistema, indicando a su vez los pasos intermedios que se realizarán:



1. Descarga del fichero de resultados correspondiente a una Autoridad Laboral.
2. Limpieza del fichero de resultados, eliminando los elementos que no aporten valor o contengan errores. Se eliminan entidades de XML dando al fichero un formato más sencillo de manejar.
3. Lectura línea a línea del fichero resultados limpio, para comenzar la generación de metadatos y el acceso al documento en red que contiene el convenio.
4. Descarga del convenio sin normalizar del portal oficial mediante el análisis y la identificación de las entidades HTML en caso de que no se presente como elemento único a través de un modelo de datos específico como RDF.
5. Extracción de Metadatos y normalización del documento y formato en JSON-LD, identificando las distintas estructuras que componen el texto.
6. Ingestión por parte de Elasticsearch de los documentos normalizados.

Este es el diseño a alto nivel donde se muestran las funciones que tiene que realizar el sistema tras el análisis llevado a cabo en el apartado anterior sobre el funcionamiento del portal de convenios colectivos REGCON y el formato de los resultados del fichero de resultados para cada Autoridad Laboral.

3.3. Diseño de la Solución

3.3.1. Modelo de datos del documento

Debido al caracter semiestructurado de la información que vamos a almacenar, ya que existen elementos dinámicos en número, se ha creido conveniente emplear un formato de documento JSON, compatible además con bases de datos documentales y que posee un tipo de medio (MIME) oficial para internet, application/json, pudiendo facilitar posteriormente la creación de un front end.

Ya que existen iniciativas similares a las que este sistema intenta solucionar, como el proyecto Lynx[24], dirigido desde el Ontology Engineering Group[25] de la Universidad Politécnica de Madrid, se ha decidido usar el modelo de datos definido por

ellos[26], por lo que el formato final será JSON-LD[27], de forma que se permita la interoperabilidad con el desarrollo de dicho proyecto.

No obstante, existe un metadato no contemplado en dicho modelo de datos, el código CNAE que identifica las actividades económicas a las que pertenece. Es por ello que lo incluiremos en dicho modelo empleando la propiedad RDF ei2a:CNAE[28] definida por la iniciativa Aragón Open Data[29] al no haberse encontrado definido en la ontología ELI[30] ni en la página de vocabularios de la UE[31].

La lista de metadatos definida por Lynx es la siguiente[32]:

Grupo	Propiedad	Uso	Propiedad RDF
Elementos básicos	id	Identificador Lynx del documento	dct:identifier
	text	Texto del documento	rdf:value
	parts	Partes del documento	eli:has_part
General	type	Tipo del documento	dct:type
	rank	Subtipo del documento	eli:type_document
	language	Idioma del documento	dct:language
	jurisdiction	Jurisdicción empleando ISO	eli:jurisdiction
	wasDerivedFrom	URL original de donde fue extraido	prov-o:wasDerivedFrom
	title	Título del documento	dct:title
	hasAuthority	Autoridad que emite el documento	lkg:hasAuthority
	nick	Nombres alternativos del documento	foaf:nick
	version	Consolidado, borrador o boletín	eli:version
	subject	Temas o palabras clave del documento	dct:subject
Identificadores	id_local	Identificador local (p. ej. BOE-A-2020-1234)	eli:id_local
	identifier	Identificador oficial	dct:identifier
	CNAE	Código Nacional de Actividades Económicas	ei2a:CNAE
Fechas	first_date_entry_in_force	Fecha en la que entra en acción	eli:first_date_entry_in_force
	date_no_longer_in_force	Fecha en la que expira	eli:date_no_longer_in_force
	version_date	Fecha de publicación	eli:version_date
Mapeos	hasEli	Identificador oficial	lkg:hasEli
	hasPDF	Link a la versión PDF	lkg:hasPDF
	hasDbpedia	Link a la versión equivalente en dbpedia	lkg:hasDbpedia
	hasWikipedia	Link a la versión equivalente en wikipedia	lkg:hasWikipedia
	sameAs	Versión equivalente	owl:sameAs
	seeAlso	Documento relacionados	rdf:seeAlso
Interno	creator	Creadores del documento en Lynx	dct:creator
	created	Fecha en la que se creo	dct:created

Dado que en este sistema limitamos la extracción de textos a los convenios colectivos, el valor para el metadato de nombre 'type' será siempre 'labor law'.

La estructura en JSON-LD del documento con datos que sirvan de ejemplo se puede encontrar a continuación:

```
{
    "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
    "@id": "BOE-A-2020-1234",
    "@type": "http://lynx-project.eu/def/lkg/LynxDocument",
    "text": "Texto de ejemplo con diferentes partes. Parte 1: Titulo 1. Parte 2: Titulo 2.",
    "metadata": {
        "identifier": [
            "123456789"
        ],
        "creator": [
            "Victor R.",
            "Jaime B."
        ],
        "first\_date\_entry\_in\_force": [
            "20200101"
        ],
        "subject": [
```

```
            "tema de ejemplo",
            "palabra clave",
        ],
        "created": [
            "2020-01-01T08:00:15"
        ],
        "jurisdiction": [
            "es"
        ],
        "date\_no\_longer\_in\_force": [
            "20190101"
        ],
        "version\_date": [
            "20200101"
        ],
        "language": [
            "es"
        ],
        "wasDerivedFrom": [
            "https://direcciondeextraccion.es"
        ],
        "id\_local": [
            "987654321"
        ],
        "type": [
            "labor law"
        ],
        "title": [
            "Convenio Colectivo de Ejemplo"
        ],
        "version": [
            "dof"
        ],
        "seeAlso": [
            "https://documentoenotroidiomaodialecto.es"
        ],
        "nick": [
            "RESOLUCION TSF/1234/2020, de 01 de enero, por la que se dispone la inscripcion \
                ldots (codigo num. 12345678987654)."
        ],
        "CNAE": [
            {
                "1234": "Servicio de extraccion de metadatos",
                "4321": "Fabricacion de codigo maquina"
            }
        ],
        "rank": [
            "res"
        ],
        "hasPDF": [
            "https://documentoenpdf.es"
        ],
        "hasAuthority": [
            "Autoridad Laboral"
        ],
        "sameAs": [
            "https://documentooficialenotroformato.es"
        ]
    },
    "parts": [
        {
            "@id": "http://lkg.lynx-project.eu/res/123456789-es-2020-01-01T08:00:15/#offset\_0
                \_40",
            "offset\_ini": 0,
            "offset\_end": 40,
            "title": ""
        },
        {
            "@id": "http://lkg.lynx-project.eu/res/123456789-es-2020-01-01T08:00:15/#offset\_4
                1\_59",
```

```
85              "offset\_ini": 41,
86              "offset\_end": 59,
87              "title": "Parte 1: Titulo 1",
88              "parent": {
89                  "@id": "http://lkg.lynx-project.eu/res/123456789-es-2020-01-01T08:00:15/#
                        offset\_0\_40"
90              }
91          },
92          {
93          "@id": "http://lkg.lynx-project.eu/res/123456789-es-2020-01-01T08:00:15/#offset\_60\_7
                7",
94              "offset\_ini": 60,
95              "offset\_end": 77,
96              "title": "Parte 2: Titulo 2",
97              "parent": {
98                  "@id": "http://lkg.lynx-project.eu/res/123456789-es-2020-01-01T08:00:15/#
                        offset\_41\_59"
99              }
100         }
101     ],
102 }
```

3.3.2. Diagrama de Clases

Habiendo realizado el análisis del portal, de los documentos que deseamos extraer y la definición del documento normalizado podemos proceder a diseñar las distintas clases que usaremos para desempeñar las funciones del sistema.

Se ha tomado la decisión de emplear 6 paquetes diferentes. El paquete padre lo hemos denominado *es.upm.fi.caspider* y contendrá la clase **MainClass**, que será la entrada a la ejecución del sistema. Dentro de este paquete encontramos el resto:

es.upm.fi.caspider

- *documents*
    - *constants*
- *laborauthority*
- *search*
- *util*

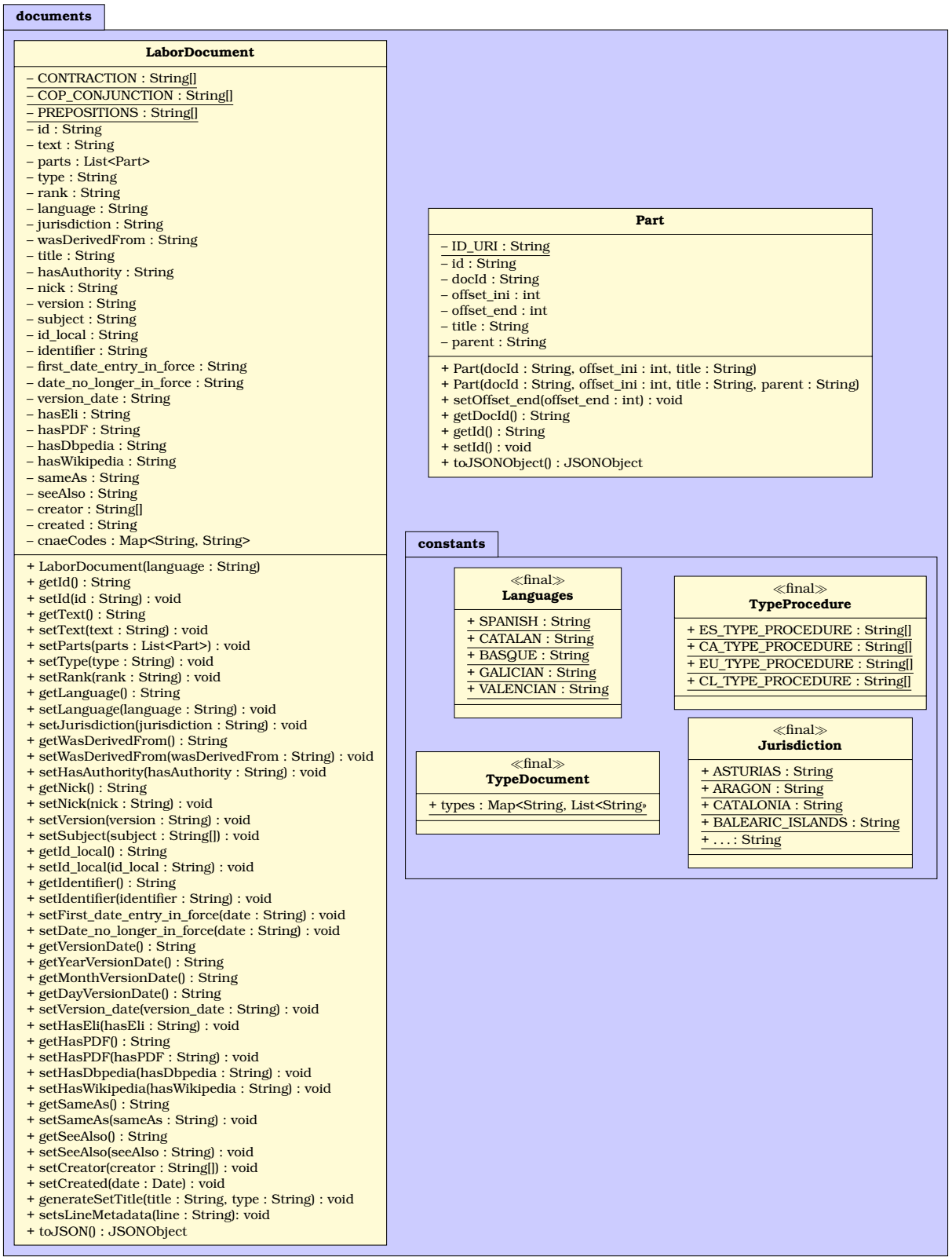
3.3.2.1. Paquete *documents*

El paquete *documents* contendrá la funcionalidad correspondiente de un convenio colectivo y las distintas partes que lo componen y el paquete *constants* las clases que representan una propiedad o metadato del documento y que sus valores están limitados por el vocabulario empleado por el modelo de datos.

La clase **LaborDocument** representa la estructura de un convenio colectivo ya normalizado, con métodos que asistan en la asignación de las propiedades y la exportación su correspondiente objeto JSON.

En cuanto a la clase **Part**, corresponderá a la parte de un documento, que ayudará a la creación y referencia de las partes que sean hijas de otras partes. Igual que la clase **LaborDocument**, contendrá su popia exportación a un objeto JSON.

Referenta al paquete *constants*, encontraremos los valores para las clases **Language**, **Jurisdiction**, **TypeProcedure** (traducciones entre los distintos tipos encontrados en REGCON) y **TypeDocument** (valores para el atributo 'rank' basado en los valores literales en los distintos idiomas y dialectos).

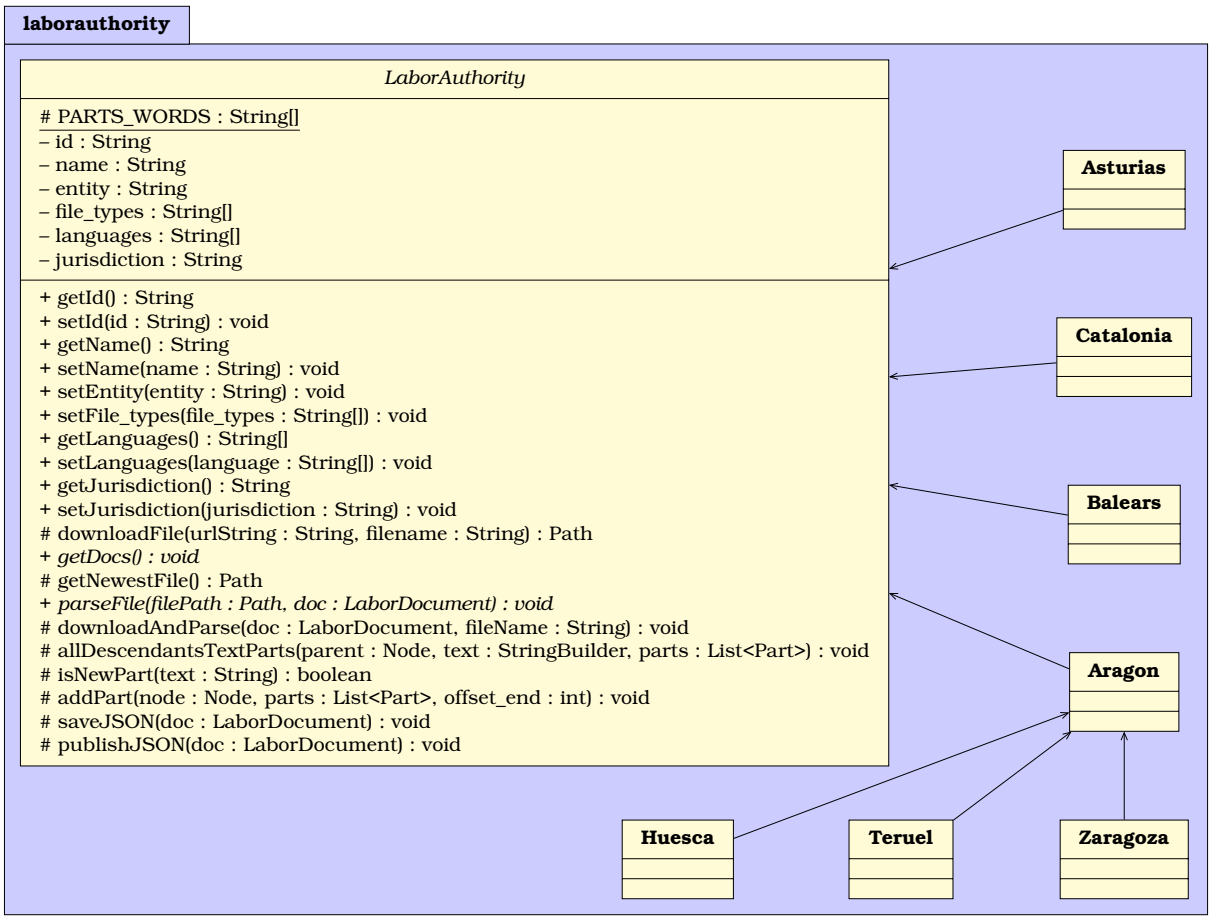


3.3.2.2. Paquete *laborauthority*

Este paquete contiene todas las clases que pertenezcan o sean una Autoridad Laboral, partiendo desde una clase padre abstracta y cada una de las clases hijas que la heredan.

Dentro podemos encontrar la clase **LaborAuthority**, que será la definición de una autoridad laboral y contiene los métodos necesarios para acceder a sus portales web donde publican los boletines y descargar y parsear los correspondientes convenios colectivos. A su vez, define métodos que deberán ser implementados por sus clases hijas debido a la particularidad de cada autoridad.

En el diagrama se muestra un ejemplo donde vemos que las clases **Asturias**, **Catalonia**, **Balears**, **Aragon** heredan de la clase **LaborAuthority** y que **Aragon**, **Huesca** y **Teruel** heredan de Aragón. No obstante, está última causística en la que las clases correspondientes a las Autoridades Laborales provinciales heredan de la clase de la Autoridad Laboral autonómica no es común, pero en este caso los portales serán el mismo y solo cambiarán pequeños detalles.

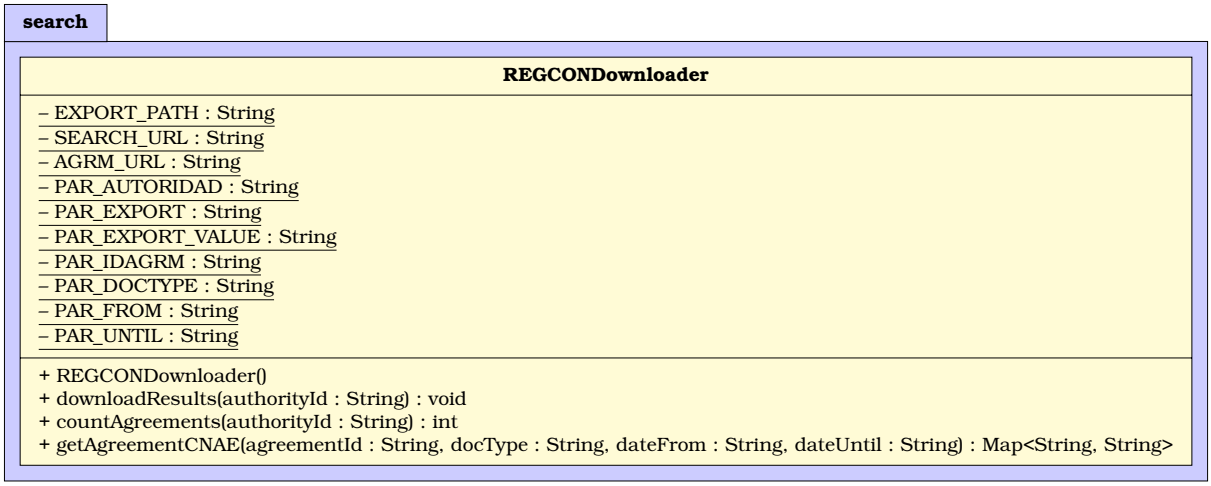


3.3.2.3. Paquete *search*

En el paquete denominado *search* encontraremos las clases necesarias para descargar los resultados de búsqueda de los convenios colectivos y sus correspondientes links. Actualmente, solo se ha identificado la necesidad de existencia de una clase,

denominada **REGCONDownloader**, pero en el futuro podría surgir la necesidad de crear nuevas clases.

Dado el id de una Autoridad Laboral, que podremos obtener al realizar una búsqueda en el portal REGCON, la clase **REGCONDownloader** buscará y descargará el fichero de resultados de dicho Autoridad Laboral y que contendrá los metadatos indicados en el apartado 3.1.2.

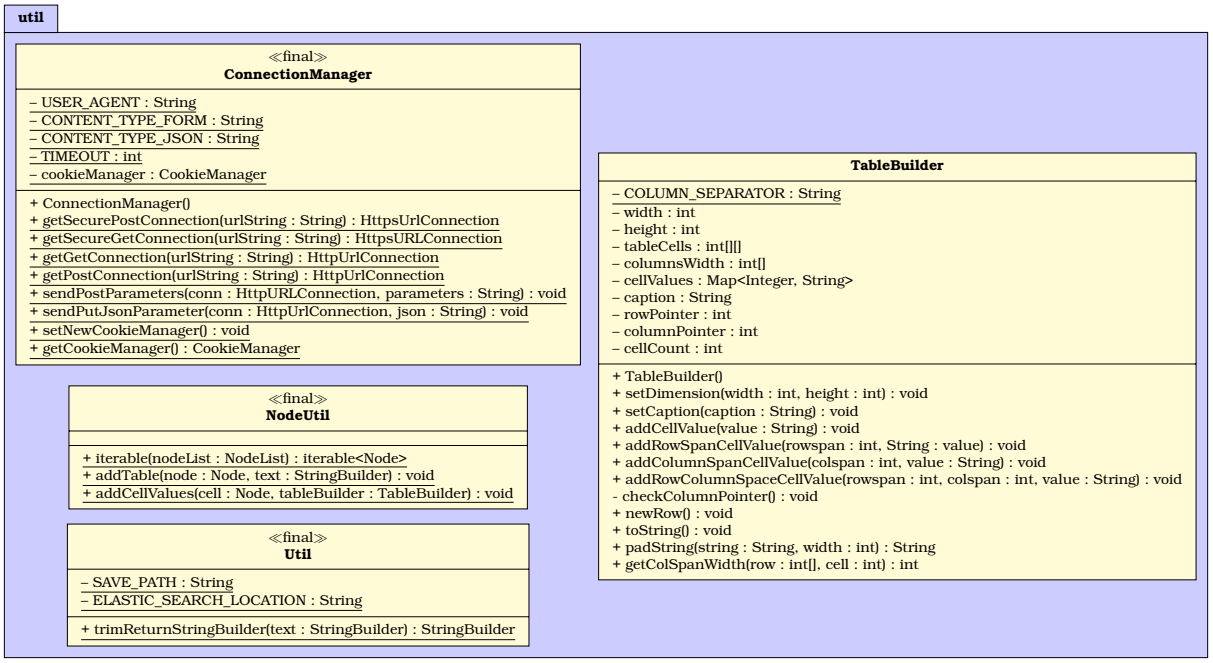


3.3.2.4. Paquete *util*

En cuanto al paquete *Util*, encapsulará todas las clases comúnes que se pueden ser usadas por el resto de clases del sistema. En un comienzo solo se identificó la necesidad de la clase **ConnectionManager**, que debido al amplio número de conexiones que se realizan, a través de está clase se reduce la cantidad de código duplicado al inicializar las conexiones.

Posteriormente, y a medida que avanzaba el desarrollo, se han encontrado otras necesidades, como la clase **NodeUtil**, que contiene un método para iterar sobre un objeto NodeList y otros dos que hacen uso de la clase **TableBuilder**, creada para dar formato de tabla al texto en plano a partir del parseo de una tabla en HTML.

Por último, la clase **Util** contendrá diferentes métodos que no casan en ningún otro sitio y no resulta conveniente crearles una clase específica.


util
≪final≫
ConnectionManager
– USER_AGENT : String
– CONTENT_TYPE_FORM : String
– CONTENT_TYPE_JSON : String
– TIMEOUT : int
– cookieManager : CookieManager
+ ConnectionManager()
+ getSecurePostConnection(urlString : String) : HttpsUrlConnection
+ getSecureGetConnection(urlString : String) : HttpsURLConnection
+ getGetConnection(urlString : String) : HttpUrlConnection
+ getPostConnection(urlString : String) : HttpUrlConnection
+ sendPostParameters(conn : HttpURLConnection, parameters : String) : void
+ sendPutJsonParameter(conn : HttpUrlConnection, json : String) : void
+ setNewCookieManager() : void
+ getCookieManager() : CookieManager
≪final≫
NodeUtil
+ iterable(nodeList : NodeList) : iterable<Node>
+ addTable(node : Node, text : StringBuilder) : void
+ addCellValues(cell : Node, tableBuilder : TableBuilder) : void
≪final≫
Util
– SAVE_PATH : String
– ELASTIC_SEARCH_LOCATION : String
+ trimReturnStringBuilder(text : StringBuilder) : StringBuilder
TableBuilder
– COLUMN_SEPARATOR : String
– width : int
– height : int
– tableCells : int[][]
– columnsWidth : int[]
– cellValues : Map<Integer, String>
– caption : String
– rowPointer : int
– columnPointer : int
– cellCount : int
+ TableBuilder()
+ setDimension(width : int, height : int) : void
+ setCaption(caption : String) : void
+ addCellValue(value : String) : void
+ addRowSpanCellValue(rowspan : int, String : value) : void
+ addColumnSpanCellValue(colspan : int, value : String) : void
+ addRowColumnSpaceCellValue(rowspan : int, colspan : int, value : String) : void
- checkColumnPointer() : void
+ newRow() : void
+ toString() : void
+ padString(string : String, width : int) : String
+ getColSpanWidth(row : int[], cell : int) : int

Capítulo 4

Implementación

4.1. Inicialización, Entorno y Librerías

Dento del sistema o proyecto, actualmente podemos encontrar dos elementos, la araña web que realiza el 'scraping' de los convenios colectivos y su parseado y el motor de búsqueda que permite indexar y recuperar dichos convenios colectivos.

En este apartado detallaremos sus entornos de desarrollo y su previsión si fuese a ser puesto en producción y como se realiza la comunicación entre ellos.

Para llevar el sistema a producción se ha decidido por desplegarlo en un servidor con el sistema operativo CentOS 8, debido su estabilidad y su alto apoyo por parte de la comunidad al tener mucha similitud con Red Hat. A partir de este sistema operativo, se lanzarán las distintas aplicaciones en entornos aislados las unas de las otras, de forma que si la seguridad de una aplicación sea comprometida se pueda limitar el daño ocasionado al no compartir recursos y comunicarse unicamente por las vías indicadas en la configuración.

Asimismo, se facilita la escalabilidad de las aplicaciones como su disponibilidad, pudiendose configurar elasticsearch como un cluster de nodos, y pudiendo ejecutar la araña web en distintos nodos en función de las autoridades laborales a las que se deese acceder, dividiendo de esa manera la carga entre contenedores docker y si fuese necesario dividiendo los contenedores docker en distintos nodos físicos.

A la hora de realizar los análisis de los distintos portales mencionados y sus peticiones HTTP se han usado tanto el navegador Firefox y sus herramientas de desarrollador como el software Postman.

A continuación, se comentarán las características más concretas de cada uno de los elementos.

4.1.1. Araña web

Para el desarrollo del sistema se ha empleado la versión 8 del lenguaje Java, debido a su alta compatibilidad entre sistema operativos y con la idea de que, si el código se abriese publicamente, cualquiera pudiese añadir funcionalidad. Para la gestión del proyecto y sus dependencias se ha usado Maven, facilitando en gran medida las fases de compilado, construscción del paquete y despliegue.

En cuanto al entorno de desarrollo se ha usado en un principio un IDE, en este caso IntelliJ IDEA, posteriormente, y con objeto de acercar el desarrollo a un entorno más similar al que se ejecutaría en producción, se tomo la decisión de llevar el desarrollo a un sistema Linux. No obstante, por el momento no se ha probado su ejecución en el ecosistema docker.

Para el control de versiones se ha usado git y como repositorio del codigo fuente GitLab.

Las librerias empleadas para este desarollo han sido diversas pero a su vez se han intentado limitar para mantener bajo control cualquier tipo de incompatibilidades y/o errores. Más concretamente se han usado las siguientes librerias:

JSON in Java[33] Librería que implementa codificadores y decodificadores de JSON, así como la capacidad de conversión desde otros formatos de texto semi-estructurados o estructurados. Facilita en gran medida la creación de objetos JSON a partir de los atributos de una clase que posteriormente se pueden convertir en ficheros o cadenas de caracteres.

Apache Commons Text[34] Proporciona funcionalidades y algoritmos extras para tratar con cadenas de caractéres. En este caso se usa para eliminar caractéres o entidades especiales usados en HTML, o en caso de que proceda, convertirlos a lenguaje natural.

JSoup Java HTML Parser[35] Se trata de una librería para trabajar con documentos HTML a través de una API que permite extraer, buscar y manupilar nodos similar a la funcionalidad de un DOM.

HtmlUnit[36] Navegador sin interfaz de usuario. Su uso es debido a la complejidad de extraer los códigos CNAE del portal REGCON por medio de peticiones HTTP, es por ello que mediante esta librería simulamos la sequencia de selección de botones necesaria para acceder a la página donde se muestra dicha información.

A medida que avance el sistema no se espera que las tres primeras librerías sean eliminadas, lo que no ocurre con la tercera, que el objetivo es prescindir de ella una vez se alcance la funcionalidad necesaria por medio de las librerías anteriores y las que inlcuye Java por defecto.

4.1.2. Motor de Búsqueda

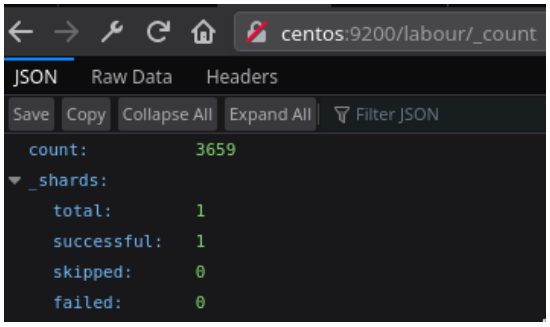
En cuanto al motor de búsqueda se ha empleado Elasticsearch, ya que además de aportar funcionalidades de búsqueda sobre texto completo sirve como sistema de persistencia de documentos.

En este caso sí se ha desplegado en un entorno docker, descargando la imagen del contenedor desde el directorio oficial de imágenes proporcionado por Docker[37]. En el momento de la instalación se instaló la última versión disponible, la 7.4.2.

Tras el despliegue se ha creado un índice denominado 'labour' donde se han introducido todos los textos normalizados tras la extracción. El número total de documentos se puede ver ejecutando en un navegador el siguiente comando:

```
1 <url>:9200/labour/\_count
```

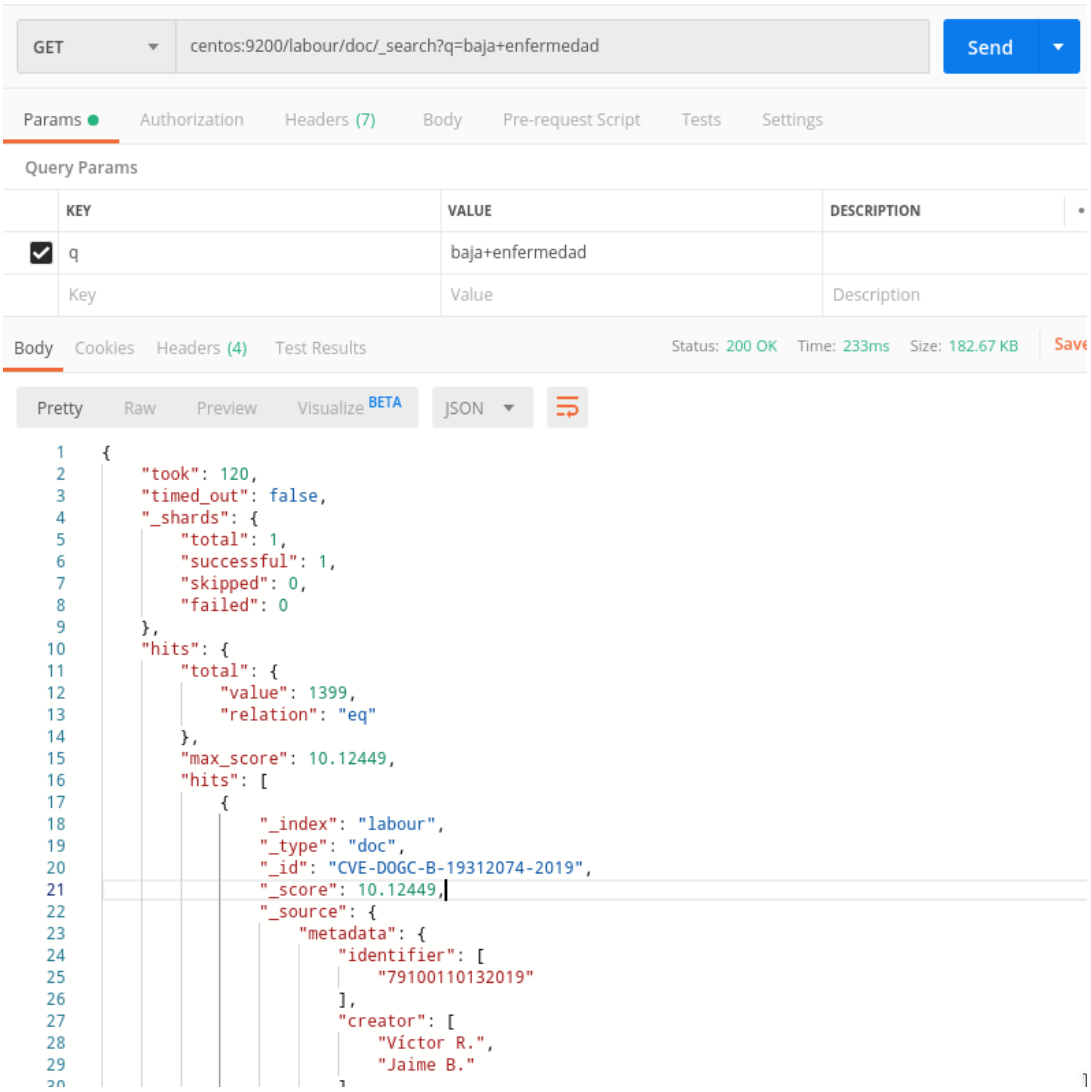
Y obtendremos los siguientes resultados:



A partir de aquí y gracias a la API REST que aporta elasticsearch podemos obtener todos los documentos o realizar búsquedas en función del texto. A continuación mostramos dos peticiones HTTP, la primera sirve para obtener todos los elementos introducidos en la base de datos y la segunda para realizar una consulta en función de a una frase o texto:

```
1 GET <url>:9200/labour/doc/\_search
2 GET <url>:9200/labour/doc/\_search?q=baja+enfermedad
```

En esta segunda consulta estamos búscando los textos que contenga 'baja enfermedad' haciendo referencia a la baja por enfermedad, con objeto de poder extraer información relevante en caso de que fuese necesario. Al realizar esta consulta en nuestro conjunto de datos, se obtienen 1399 aciertos:



A las consultas se les puede añadir mucha más complejidad, filtrando por campos, usando carácteres comodín, empleando expresiones regulares y muchas otras opciones. Se puede encontrar más información en la documentación dedicada de elasticsearch[38].

4.2. Limpieza y parseado

A parte de eliminar las entidades HTML o XML de un convenio colectivo descargado en caso de que proceda, se han elaborado detecciones de capítulos y secciones, de manera que se puedan dividir las distintas secciones y rellenar adecuadamente el elemento 'parts' definido en el documento normalizado.

Previamente a la eliminación de las entidades, se ha tratado de identificar en función del tipo de la entidad y de sus propiedades si se podían tratar de elementos separadores. Logicamente esto dependerá en gran medida del diseño de cada boletín. Posteriormente, y en caso de haber encontrado los elementos, se ha confirmado si el texto contenía alguna de las siguientes cadenas, y si lo contenía se han identificado como separadores de sección.

Las cadenas que se han empleado para identificarlas son:

```
{"Capítulo", "Capítol", "CAPÍTOL", "CAPITOL", "CAPÍTULO", "CAPITULO", "Artículo",
"Article", "ARTICLE", "ARTICULO", "ARTÍCULO", "Art.", "ART.", "Anexo", "Annex",
"ANEXO", "ANNEX", "Sección"};
```

Igualmente, en base a las entidades se han identificado los distintos tipos de metadatados. Cabe destacar que por cada Autoridad Laboral la manera de extraer los metadatos a partir de las entidades será diferente, por lo que la tarea resulta costosa en tiempo.

Otra transformación realizada al texto final ha consistido en dar formato de tabla, sobre texto plano, a una tabla escrita con entidades HTML. Para ello se ha creado una clase específica, **TableBuilder**, y que introducirá espacios y saltos de línea según corresponda para no perder información a la hora de visualizar el texto.

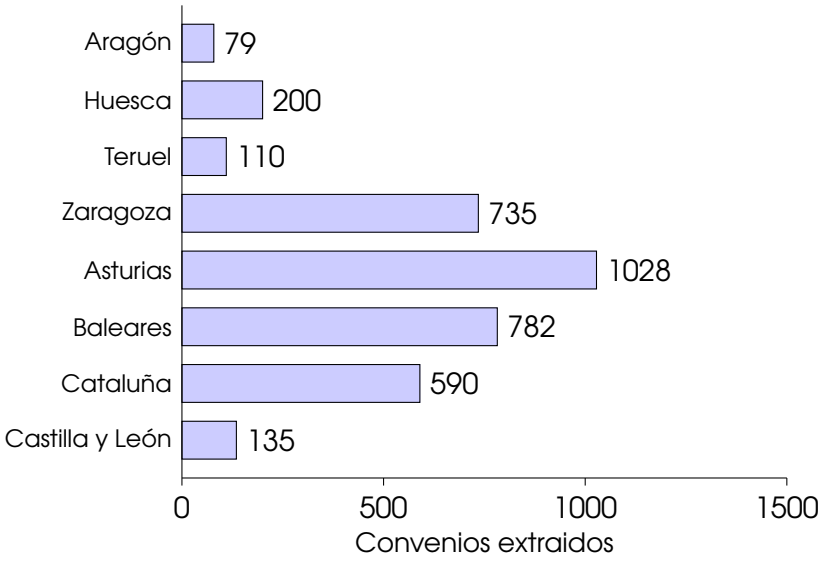
Por lo que una tabla ent HTML se verá en texto plano de la siguiente manera:

Subgrupo	Sueldo base		C. grupo		Total	
	Mensual	Anual	Mensual	Anual	Mensual	Anual
A1	1.957,20	27.400,80	248,85	2.986,20	2.206,05	30.387,00
A2	1.816,01	25.424,14	248,85	2.986,20	2.064,86	28.410,34
A3	1.698,37	23.777,18	248,85	2.986,20	1.947,22	26.763,38
B1	1.644,89	23.028,46	221,94	2.663,28	1.866,83	25.691,74
C1	1.417,16	19.840,24	158,01	1.896,12	1.575,17	21.736,36
C2	1.290,34	18.064,76	158,01	1.896,12	1.448,35	19.960,88
D1	1.193,31	16.706,34	129,85	1.558,20	1.323,16	18.264,54
D2	1.153,25	16.145,50	129,85	1.558,20	1.283,10	17.703,70
E	1.102,97	15.441,58	126,98	1.523,76	1.229,95	16.965,34

```
Subgrupo   Sueldo base   C. grupo   Total
           Mensual    Anual       Mensual    Anual      Mensual    Anual
A1         1.957,20   27.400,80   248,85     2.986,20   2.206,05   30.387,00
A2         1.816,01   25.424,14   248,85     2.986,20   2.064,86   28.410,34
A3         1.698,37   23.777,18   248,85     2.986,20   1.947,22   26.763,38
B1         1.644,89   23.028,46   221,94     2.663,28   1.866,83   25.691,74
C1         1.417,16   19.840,24   158,01     1.896,12   1.575,17   21.736,36
C2         1.290,34   18.064,76   158,01     1.896,12   1.448,35   19.960,88
D1         1.193,31   16.706,34   129,85     1.558,20   1.323,16   18.264,54
D2         1.153,25   16.145,50   129,85     1.558,20   1.283,10   17.703,70
E          1.102,97   15.441,58   126,98     1.523,76   1.229,95   16.965,34
```

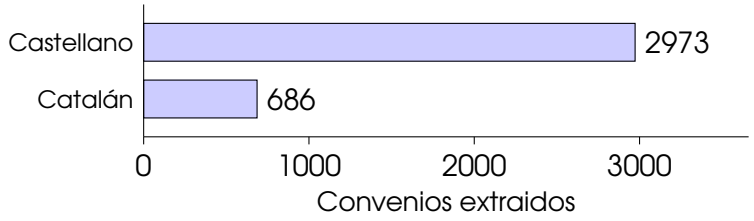
4.3. Resultados Obtenidos

Tras la ejecución del sistema se descargado y normalizado 3659 convenios colectivos de 8 Autoridades Laborales de un total de 58. En concreto, de las siguientes Autoridades Laborales:



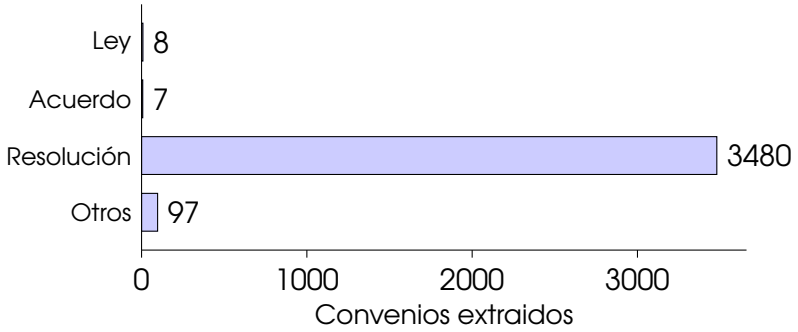
4.3.1. Convenios extraídos por idioma

En este apartado cuantificaremos los convenios colectivos extraidos por idioma, que en este caso solo serán en español y en catalán.



4.3.2. Convenios extraídos por tipo de documento

En este apartado cuantificaremos los convenios colectivos extraidos por tipo de documento. Correspondiente al atributo 'rank' y con valores del vocabulario eli:type_document.



Como observamos el resultado no suma los 3659, por lo que habrá convenios de los que no se haya podido extraer el tipo de documento.

Capítulo 5

Gestión del Proyecto

5.1. Ciclo de vida

Para el desarrollo de este software se ha seguido un ciclo de vida que se ha considerado el más adecuado dadas las circunstancias y el problema planteado. En este ciclo, se han tratado los aspectos de definición del problema y análisis del estado del arte, posteriormente de realizar un análisis de la fuente de información. Tras esto se ha procedido al diseño del modelo de datos y del software para finalizar con su desarrollo y realizar las primeras pruebas de funcionamiento.

A continuación, se detallan las distintas fases con el proceso que conlleva cada una:

- **Definición y análisis del estado del arte**: se lleva a cabo el estudio inicial del problema, generando las primeros vestigios de las soluciones que se podrían plantear para las necesidades identificadas. Se investigan las distintas tecnologías que podrían ser necesarias en el desarrolo del proyecto y se realiza una búsqueda sobre otros proyectos similares.

- **Análisis de la fuente de información y adquisición de datos**: se realiza una búsqueda inicial de las distintas fuentes existentes y una vez identificadas un exhaustivo análisis que nos permita extraer la información de las mismas de la manera más óptima posible, tomando gran interés en la calidad de los datos.

- **Diseño del modelo de datos y software**: tras haber identificado las estructuras de las que extraeremos los datos, el siguiente paso el diseñar el modelo en los que los queremos convertir, y posteriormente realizar el diseño del software en base al modelo de datos, facilitando el proceso.

- **Desarrollo**: codificación de la aplicación que hemos diseñado en el apartado anterior siguiendo el diseño y modificando si fuese necesario. Una vez haya algo materializado se procederá a la ejecución de las primeras pruebas que puedan sacar a la luz los problemas que existan hasta el momento.

- **Documentación**: será necesario documentar la aplicación, en especial si se tiene pensado permitir desarrollos externos. Se explicarán todas las etapas del sistema y se mostrarán los resultados obtenidos.

5.2. Planificación

5.2.1. Lista de Tareas

Aquí mostramos las lista de tareas llevada a cabo, así como las que finalmente no han entrado en como parte de esta aplicación, mostradas en rojo, pero que sí estaban planificadas.

Tarea 1 Análisis del portal web donde se recogen todos los convenios colectivos.

- Documentación de la información disponible en el portal.
- Pruebas de extracción y Web Scraping.

Tarea 2 Estudio, uno a uno, del acceso a los convenios colectivos de cada Autoridad Laboral a partir de la información obtenida desde el portal REGCON.

- Identificación de las peticiones HTTP necesarias.
- Documentación de las observaciones obtenidas.

Tarea 3 Valoración de distintas tecnologías de almacenamiento para el texto estructurado.

- Información de distintos formatos de almacenamiento de texto (estructurado, semi-estructurado, no estructurado).
- Soportes de almacenamiento para el formato seleccionado.
- Técnicas y tecnologías de recuperación, indexado y búsqueda.

Tarea 4 Arquitectura del sistema (extracción, almacenaje, búsqueda y visualización).

- Diseño de las políticas de la araña web.
- Diseño de la estructura de los documentos.
- Configuración del indexado y búsqueda de texto en el motor de búsqueda.
- Diseño de la arquitectura REST y las operaciones permitidas.

Tarea 5 Implementación y desarrollo del sistema.

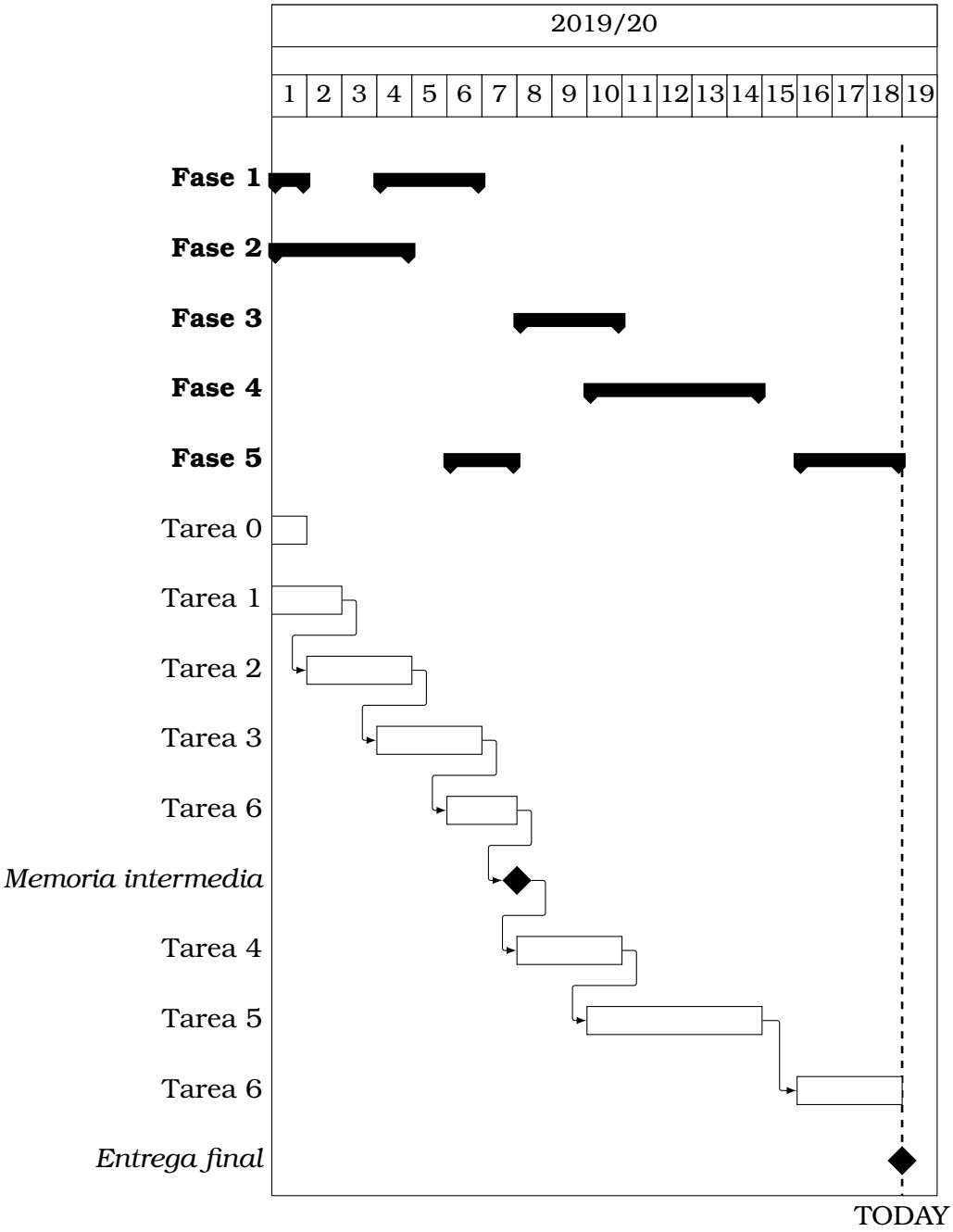
- Codificación araña web y conexión con las bases de datos.
- Implementación de la API REST que permita el acceso a la información.
- Desarrollo del front end haciendo uso de las operaciones declaradas en la API.
- Consolidación de los distintos módulos y arreglos finales.

Tarea 6 Escritura de la memoria y presentación.

5.2.2. Diagrama de Gantt

La división de trabajo se ha hecho por semanas, de las que se han calculado 18 desde la entrega del plan de trabajo, 27 de septiembre de 2019, hasta la entrega de la memoria final del día 14 de enero de 2020. Para la semana número 15, que corresponde con el periodo vacacional de Navidad y fin de año, no se han planeado tareas a

desarrollar, concentrando el mayor trabajo en las semanas previas y la consolidación del material documental en las dos semanas restantes. De esta forma se dividirán las 324 horas correspondientes al desarrollo del trabajo, realizado una media de 20h por



semana.

A pesar de haber realizado una estimación al comienzo y que ha resultado fiel durante las primeras semanas, este ha derivado debido a motivos personales y la complejidad de la tarea 5, más concretamente en la extracción de los códigos CNAE, que ha llevado bastante más tiempo del esperado. No obstante, se han dedicado todas las horas planeadas a cada tarea, aunque estaestass hayan visto reducido su ámbito en ciertas ocasiones.

Capítulo 6

Trabajos Futuros

En el proyecto no se ha conseguido alcanzar todos los objetivos propuestos al principio, principalmente por haber encontrado complicaciones en otras tareas del mismo. Además, se han identificado otras funcionalidades, mejoras y solución de problemas que sería interesante añadir si este proyecto siguiese adelante.

En cuanto a las funcionalidades propuestas, estas deberán aportar valor al usuario, de tal forma que le sea mucho más sencilla la consulta de la información, que esta se presente de una forma más clara o simplemente que aporta algo que complemente dicha información obtenida. Deberemos tener cuidado y valorar adecuadamente las características que, a piriori no aportan nada al usuario, pero que indirectamente y a través de la implementación de estas aumentan las posibilidades.

Aunque las funcionalidades podrían llegar a ser infinitas, hemos identificado las siguientes como interesantes a la hora de complementar el sistema:

- Interfaz de usuario sencilla y que permita la búsqueda de información sobre texto completo, así como filtros sobre los distintos metadatos extraidos
- Traducción automática de los convenios colectivos, a través de una plataforma como DeepL[39].
- Definición de una API REST entre elasticsearch y la interfaz de usuario que permita la creación de otro tipo de aplicaciones multiplataforma. Además, se podrán crear endpoints orientados unicamente al estudio y análisis de la información.
- Sistema de avisos y/o suscripción mediante correo electrónico para nuevos convenios colectivos según su número de acuerdo, empresa, o propiedades similares.
- Ejecución del sistema como un demonio, descargando e incluyendo en la base de datos cada día nuevas actualizaciones. EL funcionamiento sería periódico e incremental.

Aunque no aporten un valor directo al usuario, existen otros desarrollos que podrían ser convenientes:

- Establecer las distintas políticas de selección, re-visita, cortesía y paralelización, respetando unas mínimas normas de cordialidad en el uso y el acceso a las

distintas páginas web y portales a los que se acceden.

- Crear una base de datos relacional más orientada al ámbito operacional, que lleve registro de todos los convenios descargados, parseadoso que hayan presentado algún problema.

En cuanto a los arreglos y mejoras de algún procedimiento que se cree conveniente llevar a cabo, encontramos:

- Prescindir de la librería HtmlUnit generando las correspondientes consultas empleando las liberías por defecto de la distribución de java escogida.
- Añadir mensajes de registro, estado y depuración para poder trazar el comportamiento del sistema de una manera más eficaz.
- Añadir y/o mejorar el tratamiento de errores existente, notificando al administrador del sistema el instante, el motivo y el estado de la aplicación cuando surjan.
- Extraer todo la información de configuración a un archivo externo, que permita la edición de las URLs del servidor de elastic, modo de depuración, publicación o descarga de los convenios, etc.

Capítulo 7

Conclusiones

Falta mucho trabajo por hacer por parte de las administraciones para que esto pueda llegar a ser un producto sostenible. Existen portales altamente dependientes de cookies, con paso de parámetros a por medio de consultas SQL codificadas en base64.

En otras ocasiones, los convenios colectivos no están separados del boletín que se publicó ese día, dificultando su acceso, lectura y extracción.

Una simple API REST facilitaría las cosas sobremanera y funcionalmente no se necesarían más de dos endponts: búsqueda, que devuelve todos y búsqueda por número de convenio.

Bibliografía

[1] INE. *Población residente en España*. URL: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176951&menu=ultiDatos&idp=1254735572981.

[2] Migraciones y Seguridad Social Ministerio de Trabajo. *Situación de afiliados en alta por regímenes y autonomías*. URL: http://www.seg-social.es/wps/wcm/connect/wss/0057519c-1e10-4e31-aafe-d286da979dbd/01-Med-C-P-Reg+01-2019.pdf?MOD=AJPERES&CONVERT_TO=linktext&ContentCache=NONE&CACHE=NONE&CACHEID=ROOTWORKSPACE.Z18_9H5AH880M8TN80QOV0H20V0000-0057519c-1e10-4e31-aafe-d286da979dbd-mSGuSKd.

[3] Ministerio de Empleo y Seguridad Social. *Ley del Estatuto de los Trabajadores*. URL: https://www.boe.es/eli/es/rdlg/2015/10/23/2/con.

[4] Ministerio de Trabajo e Inmigración. *Real Decreto Real Decreto 713/2010, de 28 de mayo, sobre registro y depósito de convenios y acuerdos colectivos de trabajo.* URL: https://www.boe.es/eli/es/rd/2010/05/28/713/con.

[5] Comisión Europea. *Identificador Europeo de Legislación*. URL: https://eur-lex.europa.eu/eli-register/about.html?locale=es.

[6] Carlos Castillo. "Effective Web Crawling". Tesis doct. University of Chile, nov. de 2004. URL: http://chato.cl/research/crawling_thesis.

[7] Paolo Atzeni y col. *The relational model is dead, SQL is dead, and I don't feel so good myself*. Inf. téc. SIGMOD Record, 2013.

[8] Instituto Nacional de Estadística. *Clasificación Nacional de Actividades Económicas. CNAE*. URL: http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614.

[9] Julián Vida Barea. *CNAE*. URL: https://www.cnae.com.es/.

[10] Eurostat. *Nace Rev. 2 Metadata*. URL: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC.

[11] Eurostat. *Nace Rev. 2*. URL: https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF.

[12] Eurostat. *Nace Rev. 2. General description*. URL: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GEN_DESC_VIEW_NOHDR&StrNom=NACE_REV2&StrLanguageCode=EN.

[13] Eurostat. *Glossary:International standard industrial classification of all economic activities (ISIC)*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:International_standard_industrial_classification_of_all_economic_activities_(ISIC).

[14] Eurostat. *ISIC Metadata*. URL: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=ISIC_4&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC.

[15] Instituto Nacional de Estadística. *Introducción a la CNAE-2009*. URL: http://www.ine.es/daco/daco42/clasificaciones/cnae09/int_cnae_2009.pdf.

[16] Insee. Institut national de la statistique et des études économiques. *NAF. Nomenclature d'activités française*. URL: https://www.insee.fr/fr/information/2406147.

[17] Istat. Istituto Nazionale di Statistica. *Classificazione delle attività economiche Ateco 2007*. URL: https://www.istat.it/it/archivio/17888.

[18] Statistiche Ämter des Bundes und der Länder. *Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008)*. URL: https://www.klassifikationsserver.de/klassService/index.jsp?variant=wz2008.

[19] Statistics Netherlands (CBS). *Standard Industrial Classifications (Dutch SBI 2008, NACE and ISIC)*. URL: https://www.cbs.nl/en-gb/our-services/methods/classifications/activiteiten/standard-industrial-classifications--dutch-sbi-2008-nace-and-isic--#id=the-structure-of-sbi-2008-version-2018-0.

[20] Eurostat. *Eurostat*. URL: https://ec.europa.eu/eurostat.

[21] Eurostat. *RAMON - Reference And Management Of Nomenclatures*. URL: https://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC.

[22] Eurostat. *Structural business statistics*. URL: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NAT_SBS&StrLanguageCode=EN&IntPcKey=25915925&StrLayoutCode=HIERARCHIC.

[23] Migraciones y Seguridad Social Ministerio de Trabajo. *REGCON. Registro de acuerdos colectivos*. URL: https://expinterweb.empleo.gob.es/regcon/.

[24] Lynx Consortium. *Lynx. Legal Knowledge Graph for Multilingual Compliance Services*. URL: http://lynx-project.eu/.

[25] Ontology Engineering Group. *Ontology Engineering Group*. URL: http://www.oeg-upm.net/.

[26] Lynx Consortium. *Lynx D2.4 Data Management Plan*. URL: https://zenodo.org/record/3236320.

[27] W3C JSON-LD Working Group. *JSON for Linking Data*. URL: https://json-ld.org/.

[28] Gobierno de Aragón. *Propiedad RDF CNAE EI2A*. URL: http://opendata.aragon.es/def/ei2a#CNAE.

[29] Gobierno de Aragón. *Estructura de Información Interoperable de Aragón EI2A*. URL: https://opendata.aragon.es/def/ei2a/.

[30] Comisión Europea. *European Legislation Identifier dataset*. URL: https://op.europa.eu/en/web/eu-vocabularies/model/-/resource/dataset/eli.

[31] Comisión Europea. *EU Vocabularies*. URL: https://op.europa.eu/en/web/eu-vocabularies.

[32] Lynx Consortium. *Lynx Documents with metadata*. URL: http://lynx-project.eu/data2/data-models.

[33] *JSON in Java. Maven repsository*. URL: https://mvnrepository.com/artifact/org.json/json.

[34] Apache. *Apache Commons Text. Maven repository*. URL: https://mvnrepository.com/artifact/org.apache.commons/commons-text.

[35] *JSoup Java HTML Parser. Maven repository*. URL: https://mvnrepository.com/artifact/org.jsoup/jsoup.

[36] Sourceforge. *HtmlUnit. Maven repository*. URL: https://mvnrepository.com/artifact/net.sourceforge.htmlunit/htmlunit.

[37] *Elasticsearch docker image*. URL: https://hub.docker.com/_/elasticsearch.

[38] *Elasticsearch query string query*. URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html.

[39] *DeepL Translator*. URL: https://www.deepl.com/translator.

Anexo

En este anexo se recogen todos los detalles del análisis realizado de las diferentes Autoridades Laborales a partir del fichero de resultados descargado desde el REGCON.

.1. Estatal

Autoridad Laboral Estatal

Ámbito Territorial Estatal

Enlace(s) a las página(s) oficiale(s) `https://boe.es`

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, PDF

Comentarios adicionales Los links del fichero de resultados redirigen a la versión PDF. La versión en XML se puede obtener empleando el link `https://boe.es/diario_boe/xm` donde [CVE] es igual al valor encontrado en el enlace al PDF. Para obtener el HTML similar al anterior, `https://www.boe.es/buscar/doc.php?id=[CVE]`

.2. Asturias

Autoridad Laboral Asturias

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) `https://sede.asturias.es/portal/site/Asturias/menuitem.048b5a85ccf2cf40a9be6aff100000f7/?vgnextoid=c0c756a575acd010VgnVCM100000bb030a0aRCRD&i18n.http.lang=es&calendarioPqB true`

`https://www.asturias.es/portal/site/webasturias/menuitem.7b2ff7592927f53 ?vgnextoid=2cab7cd61f918510VgnVCM100000ce212b0aRCRD&i18n.http. lang=es`

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, PDF

Comentarios adicionales Los links del fichero de resultados pueden estar en formato PDF o HTML. Para obtener la versión XML usar el enlace `https://sede.asturias.es/bopa` reemplazando correctamente cada parámetro. Una alternativa es recuperar el id

del anuncio que aparace en el enlace PDF como el nombre del fichero o en el HTML como el parámetro refArticulo.

.3. Aragón

Autoridad Laboral Aragón

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) http://www.boa.aragon.es/#/

http://www.boa.aragon.es/#/opendatabuscador

https://opendata.aragon.es/

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, JSON, PDF

Comentarios adicionales Las URLs del fichero de restultados redirige a la versión PDF. Las versiones XML y JSON se podrán obtener a través de la dirección http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD=VERLST&BASE=BZHT&DOCS=1-100&SEC=OPENDATABOA[XML_or_JSON]&OUTPUTMODE=XML&SORT=-PUBL&SEPARADOR=&%40PUBL-GE=&%40PUBL-LE=&NUMB=&RANG=&TITU-C=&FDIS-C=&TITU=&ORGA-C=&TEXT-C=[Agreement_code]&SECC-C=BOA%2Bo%2BDisposiciones%2Bo%2BPersonal%2Bo%2BAcuerdos%2Bo%2BJusticia%2Bo%2BAnuncios&SECC=&SUBS-C=&MATE-C= En el primer parámetro deberemos seleccionar el formato, XML o JSON, y en el segunda el código de acuerdo. Posteriormente en la página devuelta podremos casar el link del fichero de resultado con el de dicha página.

.4. Huesca

Autoridad Laboral Huesca

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://bop.dphuesca.es/index.php/mod.menus/mem.detalle/idmenu.50003/chk.8b6c14323b2b2646096ff665f91d80d6.html

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, JSON, PDF

Comentarios adicionales Su funcionamiento es igual que el de Aragón, ya que es el portal de este último el que recoge los convenios de sus autoridades provinciales.

.5. Teruel

Autoridad Laboral Teruel

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://236ws.dpteruel.es/DPT/bopt.nsf

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, JSON, PDF

Comentarios adicionales Su funcionamiento es igual que el de Aragón, ya que es el portal de este último el que recoge los convenios de sus autoridades provinciales.

.6. Zaragoza

Autoridad Laboral Zaragoza

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://bop.dpz.es/BOPZ/`

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, JSON, PDF

Comentarios adicionales Su funcionamiento es igual que el de Aragón, ya que es el portal de este último el que recoge los convenios de sus autoridades provinciales.

.7. Andalucía

Autoridad Laboral Andalucía

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) `https://www.juntadeandalucia.es/eboja`

`https://www.juntadeandalucia.es/datosabiertos/portal.html`

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, JSON, PDF

Comentarios adicionales No hay una correpondencia obvia entre los enlaces a la versión PDF y a la versión HTML. La versión HTML se puede obtener a través del link `https://www.juntadeandalucia.es/eboja/buscador/search.do?eboja=on&q=[agreement_code]&startDate=[dd_start]%2F[MM_start]%2F[yyyy_start]&endDate=[dd_end]%2F[mm_end]%2F[yyyy_end]&type=&section=&organisation=&ordenacion=&sentido_ordenacion=descendente` En el primer parámetro deberemos seleccionar el formato, XML o JSON, y en el segunda el código de acuerdo. Posteriormente en la página devuelta podremos casar el link del fichero de resultado con el de dicha página. La versión PDF será del boletín completo del día y no separado por anuncios.

.8. Almería

Autoridad Laboral Almería

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://www.bop.almeria.es/`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales No hay otra versión disponible.

.9. Cádiz

Autoridad Laboral Cádiz

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://www.bopcadiz.es/#start`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales No hay otra versión disponible. Los anuncios no se muestran por separado, si no el boletín al completo, por ello los enlaces hacen referencia a la página en la que se encuentrán.

.10. Córdoba

Autoridad Laboral Córdoba

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://bop.dipucordoba.es/`

Idioma(s) Castellano

Formato(s) de fichero HTML, PDF

Comentarios adicionales Los enlaces hacen referencia a la versión HTML. Los PDF se pueden obtener buscando el link en el documento HTML.

.11. Huelva

Autoridad Laboral Huelva

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://sede.diphuelva.es/servicios/bop`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces hacen referencia al boletín completo, ya que los anuncios no se muestran por separado.

.12. Jaén

Autoridad Laboral Jaén

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://bop.dipujaen.es/`

Idioma(s) Castellano

Formato(s) de fichero HTML, PDF

Comentarios adicionales Los enlaces son a la versión en PDF. No hay correlación entre las URLs de las versiones HTML y PDF. Pada obtener el HTML se puede usar el enlace `https://bop.dipujaen.es/resultados/` seguido de la query `AND E.textoEdictoTXT LIKE ('%agreement_code]%') AND B.fechaBoletin between '[yyyy-MM-dd_publication]' and '[yyyy-MM-dd_publication]'` codificado en base64.

.13. Málaga

Autoridad Laboral Málaga

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://www.bopmalaga.es/`

Idioma(s) Castellano

Formato(s) de fichero HTML, PDF

Comentarios adicionales Los links referencian a las versiones HTML. Los que contienen el parametro 'marcar' o que no contienen parámetro 'control' y 'busqueda' estarán rotos. Para obtener el PDF se puede usar el siguiente enlace: `https://www.bopmalaga.es/descarga.php?archivo=[yyyyMMdd_pubication]-[n_edict]-[yyyy]-00.pdf` reemplazando los parametros correspondientemente, donde n_edict es el número que aparece junto al año de publicación en los links del fichero de resultados.

.14. Sevilla

Autoridad Laboral Sevilla

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://www.dipusevilla.es/bop/`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces redirigen al boleín completo ya que los anuncios no están publicados por separado. Los links que contienen la semiruta 'export' o que pertenezcan al dominio 'cem.junta-andalucia.es' estarán rotos. Para buscar en el portal los enlaces a los boletines que presentan un link roto se puede usar el enlace: `https://www.dipusevilla.es/system/modules/com.saga.sagasuite.theme.diputacion.sevilla.corporativo/handlers/search-bop-term.jsp?p=1&term=[agreement_code]&y=[yyyy_publication]&m=`

.15. Granada

Autoridad Laboral Granada

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://bop2.dipgra.es:8880/opencms/

Idioma(s)

Formato(s) de fichero

Comentarios adicionales Página devuelve el error 404 en el momento de su análisis.

.16. Baleares

Autoridad Laboral Baleares

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) http://www.caib.es/eboibfront/?lang=es

http://www.caib.cat/eboibfront/?lang=ca

Idioma(s)

Castellano

Catalán

Formato(s) de fichero RDF, HTML, PDF

Comentarios adicionales La mayoría de enlaces redirigen a la versión HTML, con alguna excepción de enlaces que terminan en '.pdf'. Para obtener la versión RDF a partir del enlace del HTML será suficiente con eliminar todos los parámetros que este presente y concatenar la ruta '/rdf'. Los PDF pueden ser extraidos a partir del documento HTML. Para obtener la versión en catalán es necesario con cambiar en la ruta '/en/' por '/ca/'.

.17. Canarias

Autoridad Laboral Canarias

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) http://www.gobiernodecanarias.org/boc/

Idioma(s) Castellano

Formato(s) de fichero HTML, PDF

Comentarios adicionales Para obtener los PDF a partir de los enlaces HTML se puede emplear, sustituyendo los parámetros, la siguiente URL: sede.gobcan.es/boc/boc-a-[yyyy]-[bulletin_number-xxx]-[n_edict].pdf, donde n_edict es el numero que está junto al año en el enlace del fichero de resultados.

.18. Las Palmas

Autoridad Laboral Las Palmas

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://www.boplaspalmas.net/nbop2/index.php`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Todos los links redirigen al índice del boletín del día en que se publicó. Se puede obtener el enlace al boletín completo de dicho día con el enlace: `http://www.boplaspalmas.net/nbop2/sumario.php?codigopub=1&fecha_mas_reciente=[publication_date_yyyy-MM-dd]`

.19. Santa Cruz de Tenerife

Autoridad Laboral Santa Cruz de Tenerife

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://www.bopsantacruzdetenerife.org/`

Idioma(s) Castellano

Formato(s) de fichero HTML, PDF

Comentarios adicionales La mayoría de links hacen referencia a las versión HTML. De los que redirigen al PDF, se puede obtener el HTML eliminando el nombre del fichero y el semipath '/descargar'. Para obtenerel PDF del HTML es necesario añadir el path '/descargar/' seguido del nombre de dominio y '/Bop-[n_bulletin]-[year_yy].pdf' al final, donde n_bulletin es el número junto al mes en el link HTML. Los links que contienen 'descargar' al final están rotos y debe ser eliminado y tratarlo como cualquier otro link. Los que contienen 'cgi-bin' están rotos y se debe borrar la cadena '/cgi-bin/bop/listadoPDF.py?' para trabajar con ellos.

.20. Cantabria

Autoridad Laboral Cantabria

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) `https://boc.cantabria.es/boces/`

Idioma(s) Castellano

Formato(s) de fichero XML, PDF

Comentarios adicionales Todos los links redirigen a la versión PDF del anuncia, ya que el XML será para el boletín completa. La verión XML se puede obtener de `https://boc.cantabria.es/boces/boletines.do?boletinBean.fecBolString=[dd/MM/yyyy_publication]&boletinBean.numBolString&boletinBean.tipoBol=`

0&cve&boton=Buscar, y en el caso de los links que tienen mal formato: https://boc.cantabria.es/boces/busquedaAnuncios.do?anuncioBean.filtroFecha=1&anuncioBean.tipoTexto=1&anuncioBean.entrad=[agreement_code]&anuncioBean.tipoBusqueda=todasPalabras&anuncioBean.fecDesdeString=[dd/MM/yyyy_publication]&anuncioBean.fecHastaString=[dd/MM/yyyy_publication]&anuncioBean busqAct=false&idAdmin=-1&idEntidad=-1&organizacionText&unidadText&anuncioBean.idSeccion=-1&anuncioBean.idSubseccion=-1&anuncioBean.idTipAnu=-1&boton=Buscar

.21. Castilla-La Mancha

Autoridad Laboral Castilla-La Mancha

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) https://docm.castillalamancha.es/portaldocm/

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Todos los links van al PDF del anuncio. Los links rotos o con mal formato pueden ser buscados por el número del acuerdo.

.22. Albacete

Autoridad Laboral Albacete

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://www.dipualba.es/WebBop/

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los links que contienen la cadena '#page=[xx]' redirigen al boletín completo, no estando el anuncio concreto por separado.

.23. Cuenca

Autoridad Laboral Cuenca

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s)

- https://www.dipucuenca.es/boletin-oficial-de-l
- https://www.dipucuenca.es/open-data

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Para los links que no son PDFs o están rotos, la única posibilidad es obtener el boletín completo en el que el anuncio fue publicado en

el portal de datos abiertos. Los links que contienen la cadena 'WAR' o 'index.asp' no redirigen a ningún lado.

.24. Ciudad Real

Autoridad Laboral Ciudad Real

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://bop.sede.dipucr.es/`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los links hacen referencia al sumario del boletín del día específico. Los anuncios se pueden buscar por su título en: `http://bop.sede.dipucr.es/buscador?submitted=1&tipo_busqueda=-1&texto=[title_with_+_for_spaces]&fbopdesde=[dd/MM/yyyy_publication]&fbophasta=[dd/mm/yyyy_publication]&clasificacion=-1&entidad=-1`

.25. Guadalajara

Autoridad Laboral Guadalajara

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://boletin.dguadalajara.es/`

Idioma(s) Castellano

Formato(s) de fichero HTML, PDF

Comentarios adicionales Algunos links hacen referencia a la version PDF y otros al HTML. No parece que haya correspondencia entre las URLs de las versiones PDF y las de HTML.

.26. Toledo

Autoridad Laboral Toledo

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://bop.diputoledo.es/webEbop/inicio.jsp`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces redirigen al sumario del boletín al que pertenecen. Se puede realizar una búsqueda en: `https://bop.diputoledo.es/webEbop/ebop.jsp?search=true&start&otrosCriterios&numResults&contents=[agreement_number]&publisher_type_facet&publisher&publication_date=[dd/MM/yyyy_publication]&publication_date_to=[dd/MM/yyyy_publication]`

`&insert_number&bop_number&announcement_type_facet&subject&sort=score+desc`

.27. Castilla y León

Autoridad Laboral Castilla y León

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) `http://bocyl.jcyl.es/`

Idioma(s) Castellano

Formato(s) de fichero XML, HTML, PDF

Comentarios adicionales Todos los links hacen referencia al anuncio en PDF. Se puede obtener el HTML o XML sustituyendo en la ruta 'pdf' por 'html' o 'xml' respectivamente.

.28. Ávila

Autoridad Laboral Ávila

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://www.diputacionavila.es/boletin-oficial/`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales La mayoría de links hacen referencia al boletín completo y no al anuncio.

.29. Burgos

Autoridad Laboral Burgos

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://bopbur.diputaciondeburgos.es/`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales La mayoría de links hacen referencia al sumario del boletín que contiene los links a los distintos anuncios.

.30. León

Autoridad Laboral León

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://www.dipuleon.es/bopSearchAction/Ciudadanos/Boletin_Oficial_Provincia/

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los links van al inicio del portal web. Algunos anuncios se pueden descargar por separado.

.31. Palencia

Autoridad Laboral Palencia

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://www.diputaciondepalencia.es/servicios/boletin-oficial-provincia

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces van al boletín completo ya que los anuncios no se encuentran por separado.

.32. Salamanca

Autoridad Laboral Salamanca

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://sede.diputaciondesalamanca.gob.es/BOP/

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces van al boletín completo. Los anuncios se pueden encontrar por separado pero la búsqueda no funciona, volviendolos inaccesibles.

.33. Segovia

Autoridad Laboral Segovia

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://www.dipsegovia.es/bop

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces van al boletín completo ya que los anuncios no se encuentran por separado.

.34. Soria

Autoridad Laboral Soria

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://bop.dipsoria.es/

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales La mayoría de los enlaces van a los anuncios a excepción de los más antiguos que van al boletín completo.

.35. Valladolid

Autoridad Laboral Valladolid

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://bop.sede.diputaciondevalladolid.es/

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los links se refieren a los anuncios. Para los links rotos, los anuncios se pueden buscar en: https://bop.sede.diputaciondevalladolid.es/buscar/?qbop=[agreement_number]&qdonde=texto&rpalabra=todas&orden=fecha&torden=desc&pag=buscar&command=Buscar&qcapitulo=Todos&qorganismo=&fbdd=[d_publication]&fbdm=[M_publication]&fbda=[yyyy_publication]&fbhd=[d_publication]&fbhm=[M_publication]&fbha=[yyyy_publication]&qanio=&qnumero=&ps=10

.36. Zamora

Autoridad Laboral Zamora

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://www.diputaciondezamora.es/index.asp?MP=8&MS=14&MN=2

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los links se refieren a los anuncios con alguna excepción.

.37. Cataluña

Autoridad Laboral Cataluña

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) https://dogc.gencat.cat/es/index.html?newLang=es_ES&language=es_ES

https://dogc.gencat.cat/ca/index.html?newLang=ca_ES&language=ca_ES

Idioma(s)

Castellano

Catalán

Formato(s) de fichero RDF, TURTLE, XML, HTML, PDF

Comentarios adicionales Los enlaces son a los anuncios, la mayoría en catalán. Los enlaces que no terminan en 'pdf' son a la version HTML donde se pueden extraer los links del resto de versiones. Para obtener la versión en castellano deberemos añadir o cambiar el parámetro 'language=[ca_or_es]_ES&newLang=[ca_or_es]_ES. Los links que contienen la cadena 'PdfProviderServlet' van a la versión PDF. Todos los HTML se pueden obtener con el parámetro 'documentId' y la URL: https://dogc.gencat.cat/es/pdogc_canals_interns/pdogc_resultats_fitxa/index.html?documentId=[doc_id]&language=[ca_or_es]_ES&action=fitxa&newLang=[ca_or_es]_ES

.38. Barcelona

Autoridad Laboral Barcelona

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://bop.diba.cat

Idioma(s)

Castellano

Catalán

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios, algunos en castellano, otros en catalán y algunos on ambos idiomas.

.39. Girona

Autoridad Laboral Girona

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://www.ddgi.cat/bop/faces/consultaF/index.html

Idioma(s) Catalán

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios.

.40. Lleida

Autoridad Laboral Lleida

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `http://bop.diputaciolleida.cat/faces/consultaF/index.html`

Idioma(s) Catalán

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios. Los links pertenecientes al dominio 'http://documentacio.diputaciolleida.cat/' están rotos y su búsqueda por medio del código de convenio o título no devuelve resultados.

.41. Tarragona

Autoridad Laboral Tarragona

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://www.dipta.cat/ebop`

Idioma(s) Catalán

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios.

.42. Extremadura

Autoridad Laboral Extremadura

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) `http://doe.gobex.es/`

`http://doe.gobex.es/pt/index.php`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios.

.43. Badajoz

Autoridad Laboral Badajoz

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://www.dip-badajoz.es/bop/`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios, almacenados en el portal de extremadura.

.44. Cáceres

Autoridad Laboral Cáceres

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) `https://bop.dip-caceres.es/bop/index.html`

Idioma(s) Castellano

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces son a los anuncios, almacenados en el portal de extremadura. Algún está roto y no se puede buscar.

.45. Galicia

Autoridad Laboral Galicia

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) `https://www.xunta.gal/diario-oficial-galicia/portalPublicoHome.do?lang=es`

`https://www.xunta.gal/diario-oficial-galicia/portalPublicoHome.do?lang=gl`

`https://www.xunta.gal/diario-oficial-galicia/portalPublicoHome.do?lang=pt`

Idioma(s) Castellano

Gallego

Portugués (traducción automática)

Formato(s) de fichero HTML, PDF

Comentarios adicionales La mayoría de los links son a la versión HTML en gallego. Para obtener el resto de idiomas modificar en la URL los dos caracteres antes de '.html' por 'es' para castellano, 'gl' para gallego y 'pl' par a portugues. Para obtener la versión PDF cambiar la extensión de 'html' a 'pdf'. Se pueden buscar los links que presenten problemas con la URL: `https://www.xunta.gal/diario-oficial-galicia/buscarAnunciosPublico.do?method=listado&texto=[agreement_number_or_title]&soloTitulo=false&fraseExacta=false&fechaPubDesde[dd_publication]%2F[MM_publcation]%2F[yyyy_publication]&fechaPubHasta=[dd_publication]%2F[MM_publcation]%2F[yyyy_publication]&criterioOrdenacion=ORDENACION_FECHA`

.46. A Coruña

Autoridad Laboral A Coruña

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://bop.dicoruna.es/bopportal/

http://bop.dicoruna.es/bopportal/cambioIdioma.do

Idioma(s) Castellano (traducción automática)

Gallego

Formato(s) de fichero HTML, PDF

Comentarios adicionales La mayoría de los enlaces van a los anuncios en PDF o HTML y se pueden alterar cambiando la extensión de 'html' a 'pdf' o viceversa. Los links que contienen '#page=xx' solo están disponibles en PDF como parte del boletín completo.

.47. Lugo

Autoridad Laboral Lugo

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) http://www.deputacionlugo.gal/boletin-oficial-

http://bop.dicoruna.es/bopportal/cambioIdioma.do

Idioma(s) Gallego

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces van al boletín completo o al PDF conteniendo varios anuncios. Los links indican en que página comienza el anuncio. La búsqueda no funciona para los enlaces que se encuentren rotos.

.48. Orense

Autoridad Laboral Orense

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://bop.depourense.es/portal/

https://bop.depourense.es/portal/cambioIdioma.do

Idioma(s) Mezcla castellano y gallego

Formato(s) de fichero PDF

Comentarios adicionales Los enlaces van al boletín completo al no encontrarse los anuncios por separado.

.49. Pontevedra

Autoridad Laboral Pontevedra

Ámbito Territorial Provincial

Enlace(s) a las página(s) oficiale(s) https://boppo.depo.gal/

https://boppo.depo.gal/consulta-do-boppo

Idioma(s) Castellano (traducción automática)

Gallego

Formato(s) de fichero HTML, PDF

Comentarios adicionales La mayoría de los enlaces va a la versión HTML. Los que tienen la terminación '.pdf' solo están disponibles en este formato y en castellano o gallego indiferentemente.

.50. La Rioja

Autoridad Laboral La Rioja

Ámbito Territorial Autonómico

Enlace(s) a las página(s) oficiale(s) https://iqadi.larioja.org/bor-portada

Idioma(s) Castellano

Formato(s) de fichero RDF, HTML, PDF

Comentarios adicionales Los links son a la versión HTML o PDF. Los enlaces HTML se pueden obtener partiendo de los de PDF cambiando 'PDF' por 'HTML' en el enlace. El RDF se puede obtener de https://ckan.larioja.org/dataset/anu-[announcement_number].rdf, donde el 'announcement_number' son los últimos 6 dígitos del parámetro 'reference' de los links de las otras versiones.