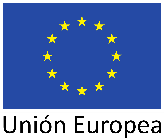




MANUAL DEL USUARIO



Programa Operativo Fondo Europeo de Desarrollo Regional Aragón 2014-2020

*Construyendo Europa desde Aragón*

Contenido

[1. Fichero/s de configuración 4](#_Toc59467159)

[1.1 Ficheros auxiliares 4](#_Toc59467160)

[1.1.1 regex.xml 4](#_Toc59467161)

[1.1.2 auxiliar.xml 4](#_Toc59467162)

[1.1.3 stopwords.txt 4](#_Toc59467163)

[1.2 Ficheros de configuración de boletines 4](#_Toc59467164)

[1.2.1 Etiquetas generales 4](#_Toc59467165)

[1.2.2 Etiquetas específicas de los boletines aragoneses 5](#_Toc59467166)

[1.2.3 Etiquetas específicas del BOE 7](#_Toc59467167)

[2. Pasos necesarios para incluir un nuevo boletín 8](#_Toc59467168)

[2.1 Modificaciones triviales 8](#_Toc59467169)

[2.1.1 Fichero de configuración 8](#_Toc59467170)

[2.2 Modificaciones no triviales 8](#_Toc59467171)

Control de versiones

|  |  |  |  |
| --- | --- | --- | --- |
| Versión | Fecha | Autor | Descripción |
| 1.0 |  | Hiberus | Primera versión (incompleta) del documento |

# Fichero/s de configuración

Los ficheros de configuración son archivos de formato XML que se utilizan en diferentes puntos del código recuperando su información para darle una **mayor variabilidad a los resultados**. De esta forma, añadiendo, modificando o eliminando información de estos es posible cambiar el resultado esperado de algunas partes del programa. A continuación, se van a explicar el uso de los diferentes ficheros existentes, así como la información que contienen y cómo alterarla.

## Ficheros auxiliares

Primero se van a explicar los ficheros auxiliares brevemente. Además de los ficheros de configuración propiamente dichos que se van a explicar en las siguientes secciones, se han incorporado varios ficheros, también en formato XML, que actúan en el código modificando la información que este va a utilizar.

### regex.xml

Este fichero contiene la información relativa a las **expresiones regulares**. Sacándolas del propio texto a este fichero hace más sencilla su modificación en caso de requerirlo. Las etiquetas superiores son <reglas> y <términos\_tipo>. En la primera, y dividiéndose por artículos de apertura y artículos de cierre, aparecen las diferentes expresiones regulares utilizadas para tratar de extraer información de los artículos con este método. En la segunda, y divididas en <libre\_e\_interna>, <libre> e <interna>, se encuentran los diferentes términos observados que refieren al tipo de la convocatoria. Es muy ventajoso que estén ubicados aquí, ya que basta con crear una nueva etiqueta <termino> en cualquiera de los tres tipos para que el código lo tenga en cuenta para su clasificación.

### auxiliar.xml

Este fichero funciona, como su nombre indica, a modo de **fichero auxiliar del código**, evitando reutilizar información como la <correspondencia\_meses>. Incluye información cuya modificación puede resultar interesante para el usuario como el número mínimo de campos requeridos para almacenar la información del artículo en la base de datos (<num\_min\_campos>), o el formato por defecto elegido para obtener el texto de los distintos boletines (<formatos\_por\_defecto>). Además, se incluyen una serie de etiquetas que contienen cadenas de caracteres (o *strings*) para diferentes puntos del código, como identificación de ofertas o identificación de tablas que contienen puestos de trabajo.

### stopwords.txt

No es un fichero de configuración ni auxiliar, si no que es un archivo de formato TXT que contiene una lista de *stopwords*, o **palabras sin significado semántico**. Esta lista se utiliza para la limpieza de los textos en algunas ocasiones, evitando, por ejemplo, que un puesto termine por una preposición (algo que puede ocurrir en ocasiones, si el modelo NER detecta texto posterior al del puesto). Este fichero contiene palabras a razón de una por línea, por lo que es muy sencillo añadir nuevas palabras.

## Ficheros de configuración de boletines

En esta sección se va a detallar el formato de los ficheros de configuración de los boletines, así como la explicación de las diferentes etiquetas en estos aparecidas.

### Etiquetas generales

A continuación, se va a enumerar y explicar las etiquetas que aparecen en los cinco ficheros de configuración tratados:

* **<puntos\_corte>**: En esta etiqueta se incluyen los diferentes puntos de corte utilizados para extraer el texto de los ficheros pdf. Es necesario hacer el recorte de cada página para evitar obtener el encabezado en cada página o los números que aparecen en los laterales y en el pie de las páginas. Estos puntos de corte se han calculado con varias observaciones, investigando con la herramienta [pdfplumber](https://pypi.org/project/pdfplumber/) hasta qué puntos llega el texto y hasta qué puntos lo que no interesa extraer. Los puntos de corte son **<x0>** (distancia del lado izquierdo del carácter hasta el lado izquierdo de la página), **<x1>** (distancia del lado derecho del carácter hasta el lado izquierdo de la página, **<top>** (distancia de la parte superior del carácter hasta el lado superior de la página) y **<bottom>** (distancia de la parte inferior del carácter hasta el lado superior de la página). Además, se incluye cada una de estas etiquetas con el sufijo ‘\_fecha’, con la intención de permitir obtener la fecha de cada artículo por este medio.
* **<etiquetas\_xml>**: El texto que estas etiquetas indican refiere a la ruta utilizada en el código para obtener los atributos indicados de los artículos en formato XML. Las **<auxiliares>** sirven para obtener campos requeridos para realizar distintas operaciones. Sus subetiquetas varían en función del boletín tratado, requiriendo de diferente información auxiliar si se está tratando el BOE o los boletines aragoneses (para otros boletines, como no se sabe si requerirán de algunas de estas operaciones, no se puede dar por hecho que van a requerir de este tipo de etiquetas). Por otro lado, las **<a\_guardar>** son las de los metadatos que se van a guardar en el fichero de información, para su posterior almacenamiento en la base de datos. En estas, el nombre de la propia etiqueta aquí indicado es el nombre que se le va a atribuir en dicho fichero de información.
* **<charsets**>: En esta etiqueta se incluye los diferentes charsets utilizados a lo largo del código para los diferentes boletines. Como la ingesta del BOE y de los aragoneses es diferente, sus subetiquetas varían.

### Etiquetas específicas de los boletines aragoneses

Además de las etiquetas troncales, hay algunas específicas para los diferentes boletines, necesarias por el modo de obtener los documentos. Las requeridas específicamente por los cuatro boletines aragoneses son:

* **<parámetros\_url>:** Dado que la ingesta de los boletines aragoneses requiere de parametrización de la URL, es necesario incluir los valores de los diferentes parámetros tratados en sus ficheros de configuración. Para los cuatro boletines únicamente es necesario indicar el parámetro **<secc-c>**. Además, como el BOA cuenta con secciones y subsecciones en su fichero de configuración aparece también la etiqueta **<info\_seccion>**, con **<seccion>** y **<subsección>** indicados en su interior. Esta etiqueta puede aparecer múltiples veces, una por cada sección de la que se quieran obtener artículos. Los valores de estas etiquetas han de concordar con los utilizados para los parámetros (si se indica una sección que no existe, podría fallar. Por ejemplo, los provinciales no cuentan con secciones ni subsecciones por lo que no se han de indicar).
* **<etiquetas\_xml\_sumario>:** En esta etiqueta se encuentran las subetiquetas **<raíz>** y **<registro>**. La primera contiene la ruta a la raíz del XML Sumarizado, mientras que la segunda contiene la raíz a cualquier registro del mismo.

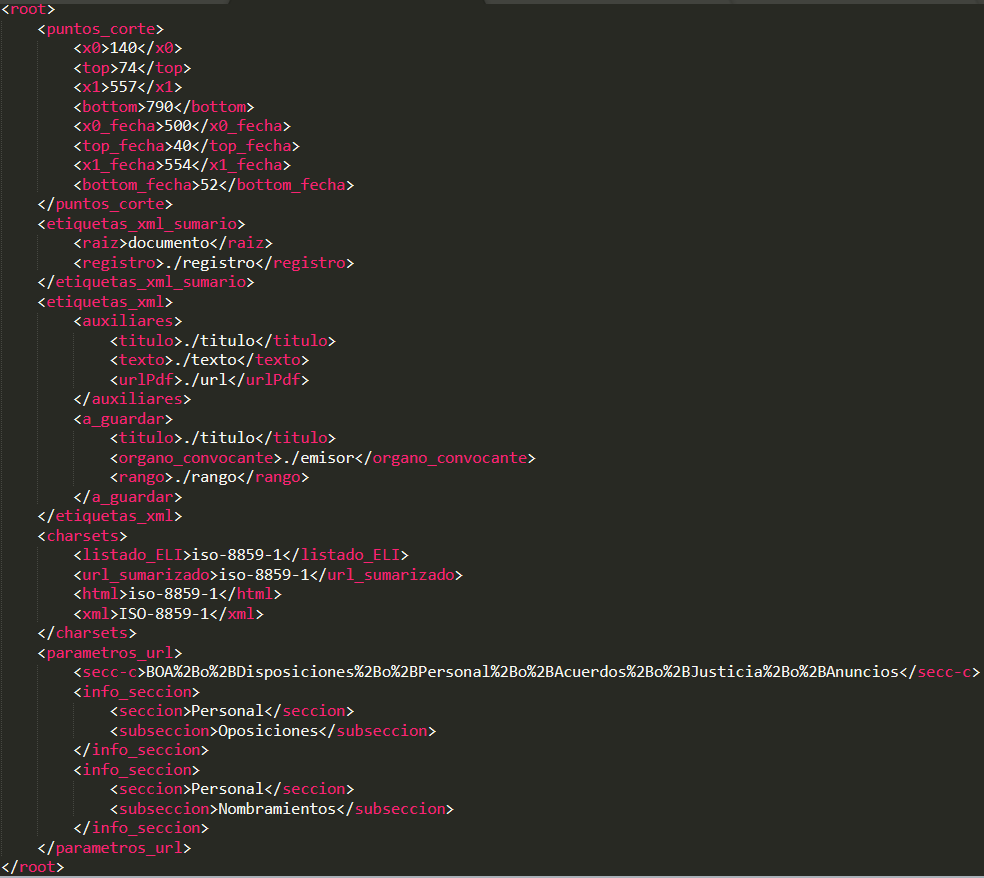


Ilustración 1. Formato del BOA

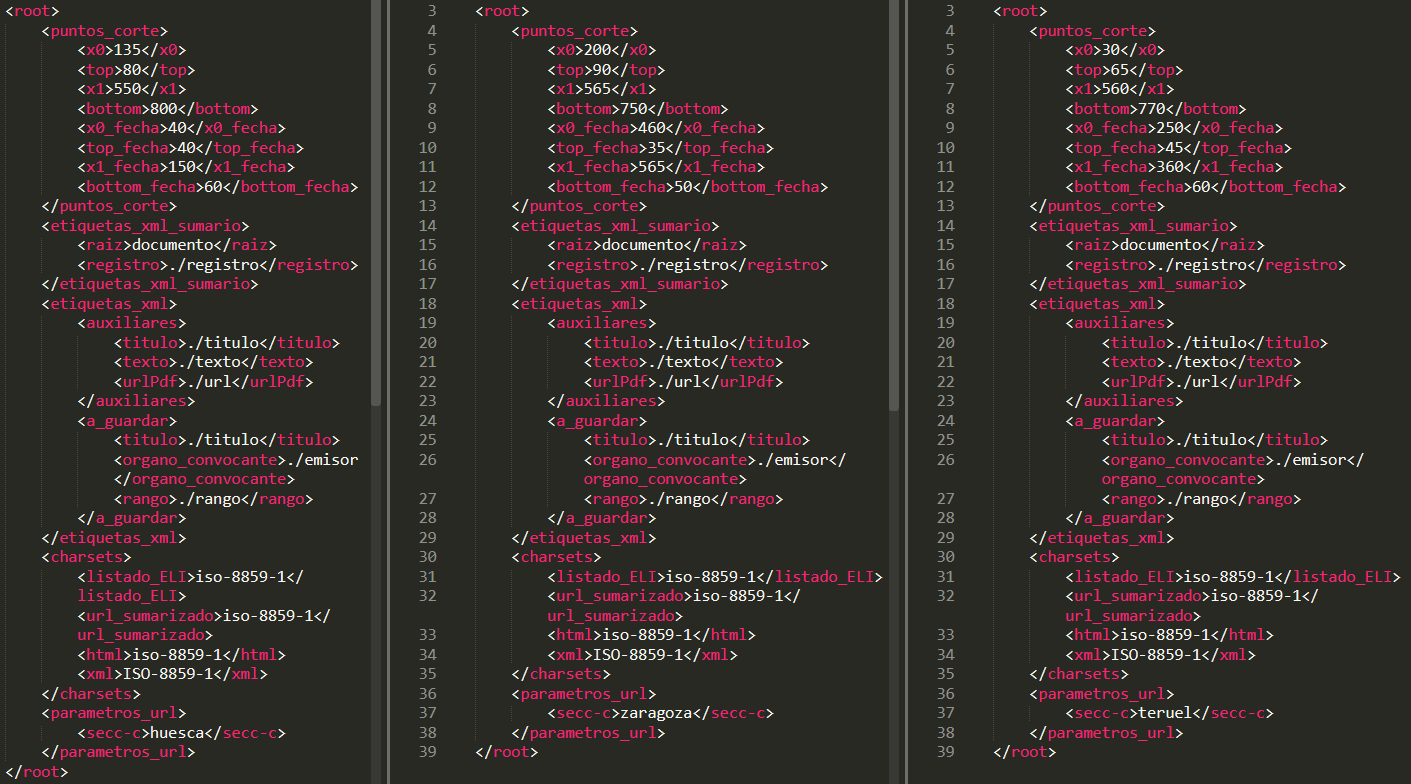


Ilustración . Ficheros de configuración de BOPH, BOPZ y BOPT

### Etiquetas específicas del BOE

El BOE también cuenta con etiquetas específicas por su modo de obtener los documentos:

* **<prefijo\_url>:** Indica el prefijo de las url para obtener los diferentes artículos. A este prefijo se le concatena el trozo de url indicado en cada ítem del XML Sumario, al que se llega con <url\_formatos>.
* **<prefijo\_url\_sumario>:** Indica el prefijo de las url para obtener los XML Sumarios. A este prefijo se le concatena el día del que obtener dicho XML.
* **<secciones\_xml>:** Contiene subetiquetas que cuentan con la ruta del XML Sumario a cada ítem de la sección indicada por la subetiqueta. De esta forma, y al igual que con <info\_seccion> en el BOA, es posible recuperar artículos de varias secciones.
* **<url\_formatos>:** Indica los formatos que se van a recuperar de cada artículo (han de ser formatos válidos), y su texto indica el nombre de las etiquetas que contienen, en el XML Sumario del BOE, el sufijo del enlace que lleva a dicho formato del artículo.

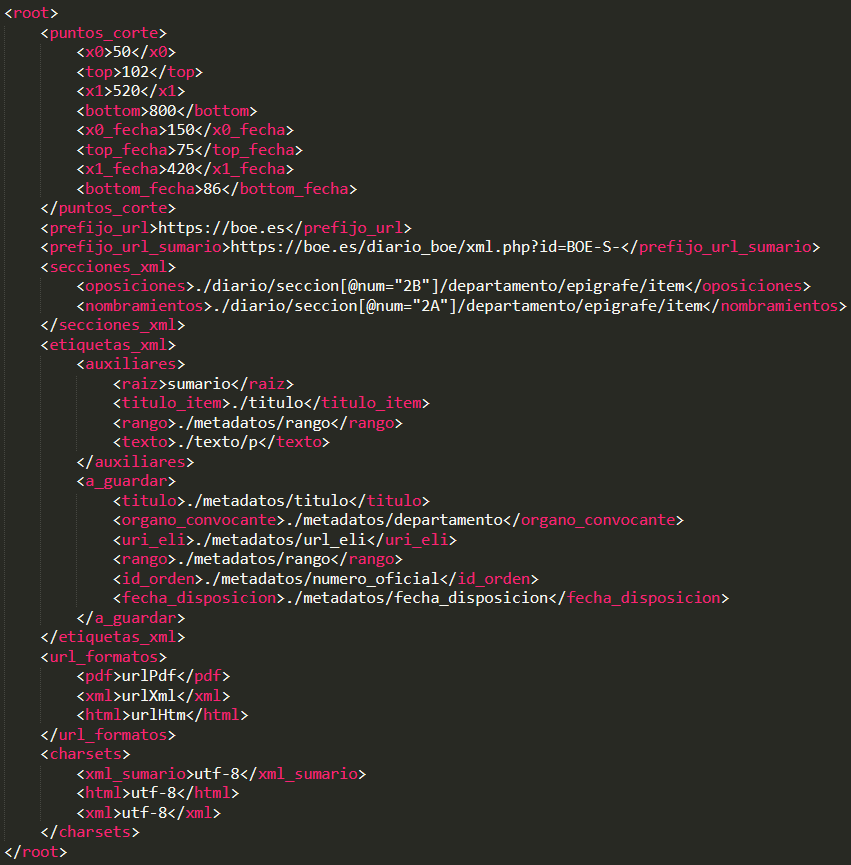
****

Ilustración . Fichero de configuración del BOE

# Pasos necesarios para incluir un nuevo boletín

En esta sección se van a explicar los pasos necesarios para la inclusión de un nuevo boletín no tratado.

## Modificaciones triviales

Algunas de las modificaciones a realizar son triviales, es decir, **no requieren de personal técnico** para su realización, ya que trata de crear ficheros de configuración y añadir etiquetas en dicho XML. Las modificaciones de los ficheros de configuración han de cambiarse tanto en el servidor de frontend como en el de backend, para que se mantengan iguales.

### Fichero de configuración

Es necesario crear un fichero de configuración, que esté guardado en el directorio ‘ficheros\_configuracion’ junto a los demás, que tenga formato XML y que tenga como nombre ‘fuente\_conf.xml’, siendo ‘fuente’ el nombre del boletín. A continuación, se van a enumerar las etiquetas necesarias para la ejecución de la ingesta sin error. Estas etiquetas han de estar dentro de una etiqueta raíz llamada **<root>**:

* **<prefijo\_url\_sumario>:** El prefijo de la URL del sumario, como ocurre en el BOE.
* **<etiquetas\_xml>:** Como subetiqueta, solo aparece <a\_guardar>. <auxiliares> no aparece en un principio porque no se puede presuponer que sea necesario realizar operaciones extra, ni cuáles serían. Respecto a **<a\_guardar>**:
  + Como subetiquetas han de estar o **<texto>** (con la ruta al texto completo de cada registro) o **<titulo>** (con la ruta al título de cada registro), o ambos. En caso contrario, no se podrá extraer información del artículo por lo que será descartado.
  + Si como subetiquetas aparecen etiquetas cuyos nombres contengan **‘htm’**, **‘pdf’** o **‘xml’**, la información de estas ha de contener la ruta, dentro del XML Sumario, a los enlaces de los respectivos formatos. De esta forma, se conseguirá recuperar cada formato indicado y guardar estos enlaces en el fichero de información. Funciona de forma similar a <url\_formatos>, explicado en la Sección 1.2.3., solo que en este caso los enlaces habrán de aparecer enteros.
  + Como se ha comentado en la Sección 1.2.1, las etiquetas que aparezcan en esta subetiqueta serán guardadas en el mismo.
* **<etiquetas\_xml\_sumario>:** Contiene las etiquetas del XML Sumario obtenido. Como mínimo ha de tener la etiqueta **<registro>**, que indique la ruta a cada registro a extraer del sumario y la etiqueta **<raíz>**, que indique la ruta a la raíz del sumario (la raíz es la etiqueta dentro de la cual se hallan todas las demás etiquetas del fichero XML).

También es posible incluir la etiqueta **<secciones\_xml>** para indicar las secciones del XML a obtener de la misma forma que en el BOE. Si no lo hay, recuperará de todas.

Otra modificación trivial es, en el fichero de configuración ‘auxiliar.xml’, añadir una subetiqueta en <formatos\_por\_defecto> siendo el tag el nombre de la nueva fuente y siendo el texto el formato preferido.

## Modificaciones no triviales

Se ha tratado de generalizar todo cuanto se ha podido la llamada ingesta extra. Sin embargo, se han tomado algunas licencias necesarias como dar por hecho que los nuevos boletines se obtendrán de una forma similar a lo ofrecido por el BOE (con prefijo de URL, concatenando el día y obteniendo un XML Sumario), ya que no es posible generalizar tanto ese primer paso de recuperación de documentos. Se ha optado por la forma del BOE por ser una forma más general y menos dependiente de tantos parámetros como la ingesta del BOA. Por ello, para conseguir que funcione es probable que, como primer paso, se modifique el código de la ingesta para adaptarlo fielmente al nuevo boletín, estudiado y analizado.

Otra modificación no trivial sería el obtener los puntos de corte del artículo en formato pdf para poder obtener su texto limpio. Para esto, es necesario analizar un corpus no despreciable de documentos para determinar la estructura del texto general para, luego, utilizar la herramienta [pdfplumber](https://pypi.org/project/pdfplumber/) para encontrar esos puntos de corte para recoger el texto sin ruido.

Una vez se han ingestado y convertido a txt los artículos, se almacenan en el mismo directorio que los de otros boletines, por lo que no es necesario hacer operaciones especiales para los pasos siguientes.