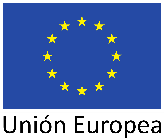




VALIDACIÓN DEL ANÁLISIS EXPLORATORIO



Programa Operativo Fondo Europeo de Desarrollo Regional Aragón 2014-2020

*Construyendo Europa desde Aragón*

Contenido

[1. Estudio de las fuentes de los datos 4](#_Toc55293323)

[1.1 Apuntes generales 4](#_Toc55293324)

[1.2 Formatos abarcados y estructura 4](#_Toc55293325)

[1.2.1 BOE 5](#_Toc55293326)

[1.2.2 BOA y provinciales 9](#_Toc55293327)

[2. Ingesta de los artículos 16](#_Toc55293328)

[2.1 Portal del Boletín Oficial de Aragón 16](#_Toc55293329)

[2.1.1 Investigación y obtención manual de los artículos 16](#_Toc55293330)

[2.1.2 Obtención automática de los artículos 19](#_Toc55293331)

[2.2 Portal del Boletín Oficial del Estado 21](#_Toc55293332)

[3. Corpus utilizado para el análisis 23](#_Toc55293333)

[4. Estudio de los campos 23](#_Toc55293334)

[4.1 Campos básicos 23](#_Toc55293335)

[4.1.1 Fuente de los datos 23](#_Toc55293336)

[4.1.2 Fecha de publicación 23](#_Toc55293337)

[4.1.3 URI ELI de la norma 24](#_Toc55293338)

[4.1.4 Título 26](#_Toc55293339)

[4.1.5 Puesto ofertado 26](#_Toc55293340)

[4.1.6 Fecha de inicio de presentación de solicitudes 33](#_Toc55293341)

[4.1.7 Fecha de fin de presentación de solicitudes 34](#_Toc55293342)

[4.1.8 Enlace a la convocatoria de la fuente oficial 34](#_Toc55293343)

[4.1.9 Órgano convocante 34](#_Toc55293344)

[4.1.10 Datos de contacto 35](#_Toc55293345)

[4.2 Campos necesarios para las vistas 35](#_Toc55293346)

[4.2.1 Estado de la convocatoria 35](#_Toc55293347)

[4.2.2 Identificador de la convocatoria 39](#_Toc55293348)

[4.2.3 Fecha de planteamiento/firma de la convocatoria 39](#_Toc55293349)

[4.2.4 Enlace al artículo de cierre 40](#_Toc55293350)

[4.3 Otros campos mencionados 40](#_Toc55293351)

[4.3.1 Tipo de la convocatoria 40](#_Toc55293352)

[4.3.2 Número de plazas ofertadas 41](#_Toc55293353)

[4.3.3 Cuerpo del puesto ofertado 42](#_Toc55293354)

[4.3.4 Escala del puesto ofertado 42](#_Toc55293355)

[4.3.5 Grupo del puesto ofertado 42](#_Toc55293356)

[4.3.6 Plazo de presentación de solicitudes 43](#_Toc55293357)

[4.3.7 Subescala del puesto ofertado 44](#_Toc55293358)

[4.3.8 Rango del artículo 44](#_Toc55293359)

[4.3.9 ID de la orden del artículo 45](#_Toc55293360)

Control de versiones

|  |  |  |  |
| --- | --- | --- | --- |
| Versión | Fecha | Autor | Descripción |
| 1.0 | 03/11/2020 | Hiberus | Primera versión del documento |

# Estudio de las fuentes de los datos

El presente documento tiene el objetivo de comentar y detallar los aspectos relativos al análisis exploratorio, con la finalidad última de dar a conocer la situación de los documentos que se van a tratar y de la información que de los mismos se va a extraer, para obtener una validación sobre estos aspectos.

En esta sección se va a detallar el estudio realizado de las diferentes fuentes de datos: cuáles son, qué contienen, casos específicos, etc.

## Apuntes generales

Se considera más correcto almacenar y tratar los documentos de los diferentes artículos que conforman cada boletín, en detrimento de los propios boletines diarios, también llamados sumarios. De esta forma, los enlaces a las convocatorias y los propios ficheros almacenados irán dirigidos realmente a la propia convocatoria, y no al boletín del día en los que aparecen, dándole más comodidades al usuario final.

Se ha comprobado que las ofertas de empleo que aparecen en los boletines provinciales no aparecen en el BOA, evitando así posibles duplicados y haciendo necesario tratar los cinco boletines. Tampoco se han observado casos de ofertas del BOA que aparezcan en el BOE.

Para obtener convocatorias de ofertas de empleo, en el BOE y en el BOA es posible filtrar los artículos en función de la sección a la que pertenezcan:

* En el BOE, existe la sección “II. Autoridades y personal”. Además, en el XML Sumario (no así en el buscador avanzado manual), esta sección se divide en las secciones “ II. Autoridades y personal. - A. Nombramientos, situaciones e incidencias” y “II. Autoridades y personal. - B. Oposiciones y concursos”, en las que aparecen los nombramientos y las aperturas de convocatorias, respectivamente.
* En el BOA, la sección que nos interesa es también la sección “II. Autoridades y Personal”, que se divide en subsecciones equivalentes a las del BOE: “II.a. Nombramientos, situaciones e incidencias” y “II.b. Oposiciones y concursos”. En la sección II.b. también se han observado publicaciones de listas de admitidos.
* En los boletines provinciales no existen secciones, por lo que no hay un filtrado de artículos previo.

De esta forma, los artículos de estas secciones son o de aperturas de convocatoria o de cierres de convocatoria. Esta división por secciones nos sirve como primer filtrado, ya que la información de sus artículos interesa para nuestro propósito a diferencia de otros tipos de artículos, como leyes o presupuestos.

## Formatos abarcados y estructura

En esta sección se van a explicar y mostrar los formatos tratados y la estructura de los mismos, para cada uno de los tipos de boletines.

Como se indica en el pliego (punto 2.2.d.2), las fuentes a implementar van a ser cinco: el Boletín Oficial del Estado, el Boletín Oficial de Aragón, el Boletín Oficial de la Provincia de Huesca, el Boletín Oficial de la Provincia de Zaragoza y el Boletín Oficial de la Provincia de Teruel. De ahora en adelante, van a ser llamados BOE, BOA, BOPH, BOPZ y BOPT, respectivamente.

En cuanto a los formatos, los que se van a tratar son los mencionados en el primer ítem del punto 2.1.a del pliego: PDF, XML y HTML.

En los PDFs, se va a mencionar cómo se cortan las páginas para obtener únicamente el texto del artículo, sin otros datos que molesten, sin entrar en detalle de los puntos de corte numéricos por los que se ha cortado.

### BOE

Los artículos del BOE en formato PDF siguen la forma de la siguiente imagen:

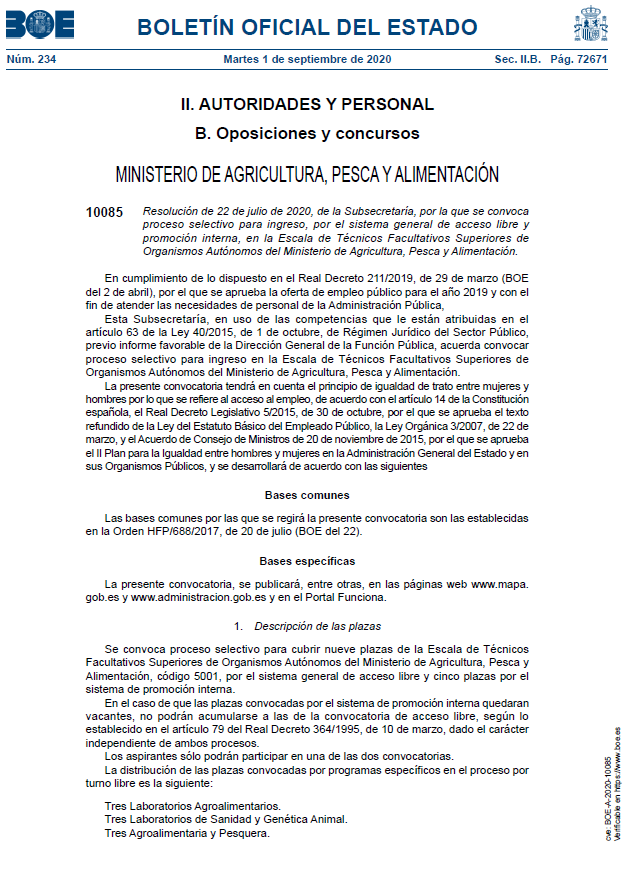


Ilustración 1. Página de PDF del BOE

Como se puede observar, tiene información que se podría precisar como metadatos en el encabezado de cada página. En la primera página cuenta con varios títulos sobre las secciones y el órgano convocante, así como un número identificativo y el resumen del artículo, que sería nuestro campo “título”. En la esquina inferior derecha de cada página tenemos un identificador del artículo y la página web del Boletín Oficial del Estado. Para obtener el texto, se corta por debajo del encabezado, por encima del número de página y a la izquierda del identificador.

En cuanto a la estructura del propio texto, hay una varianza significativa. Las aperturas habitualmente dividen sus bases en diferentes secciones, pudiendo haber anexos que son mucho más variables. Estos pueden abarcar todo el temario de la oposición, tener tablas con información detallada de los puestos (horizontales o verticales, como se comentará en el apartado 4.1.5), información del tribunal, plantillas de solicitudes a cumplimentar…

En cuanto al formato XML, los artículos del BOE tienen la siguiente forma:



Ilustración 2. XML del BOE

Tienen una etiqueta de metadatos donde se guardan estos de una forma bastante detallada, en caso de existir. Se pueden extraer campos que aquí aparecen completos, como el título, o que requieren de pocas modificaciones, como la fecha de publicación. A continuación, en la etiqueta <texto>, aparecen los diferentes párrafos, identificados por la etiqueta <p>.

Además, en caso de tener URI ELI, este XML cuenta además con las etiquetas <url\_eli>, <metadata-eli> y subetiquetas de la segunda. La que interesa especialmente es <url\_eli>:



Ilustración 3. Atributos del XML del BOE sobre la URI ELI

En cuanto al HTML, los artículos del BOE suelen ser como la siguiente imagen:

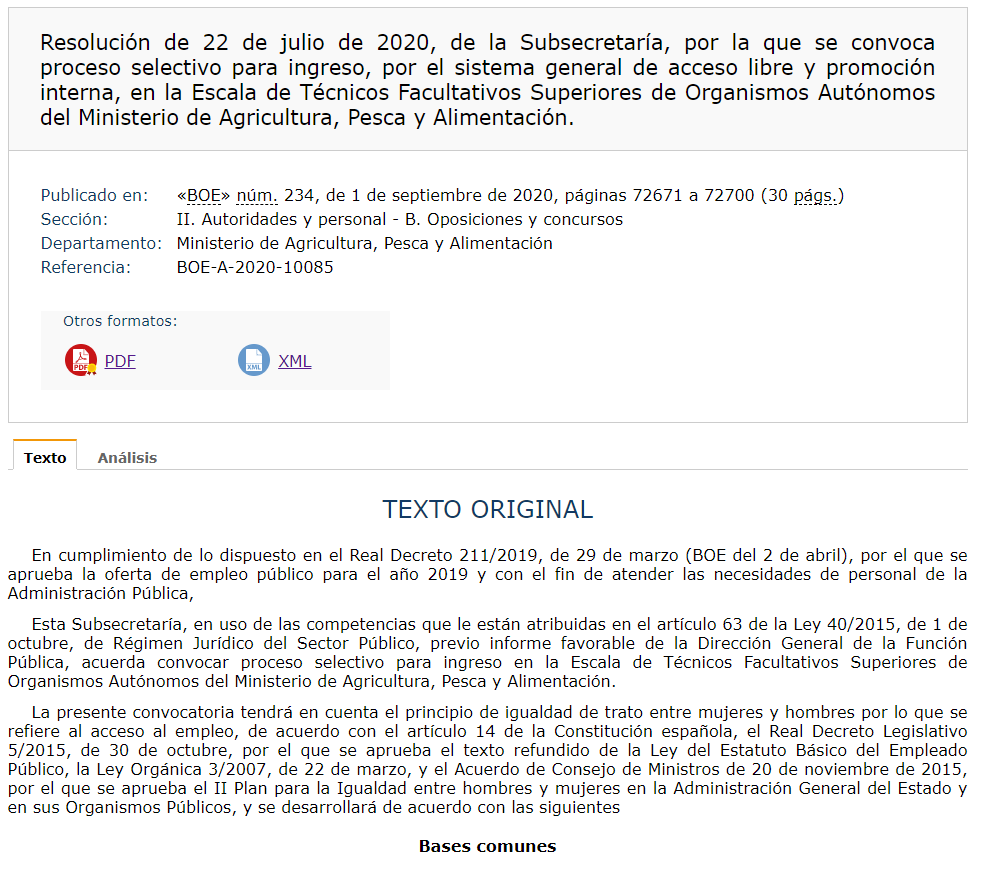


Ilustración 4. HTML del BOE

Respecto a la Ilustración 4, en caso de disponer de URI ELI, aparece bajo “Referencia:” con la etiqueta “Permalink ELI:”.

### BOA y provinciales

En el único formato en los que se diferencian estos cuatro tipos es en el PDF. El XML y el HTML, al ser originados en el mismo portal, tienen formatos iguales. Primero se muestra la estructura del PDF del BOA:

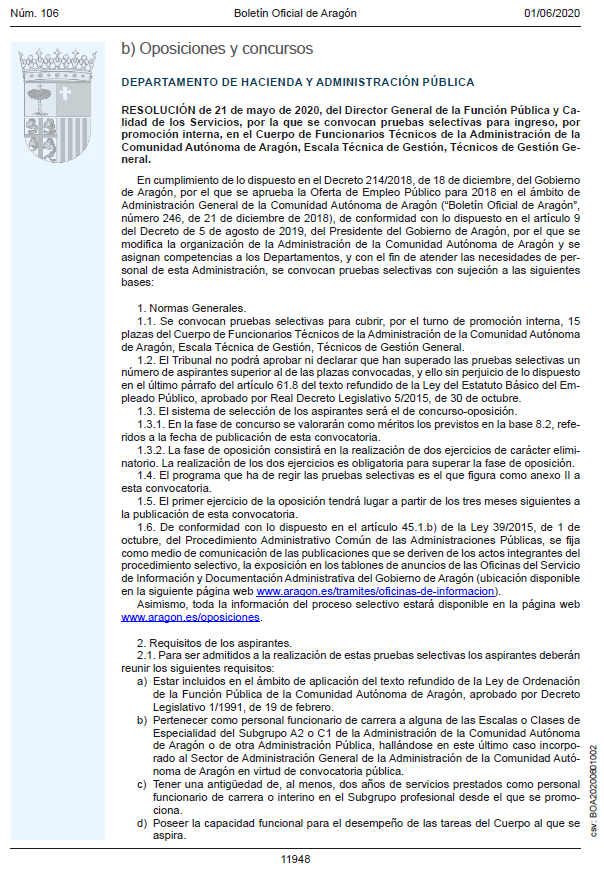


Ilustración 5. PDF del BOA

En el encabezado, únicamente el número, el texto “Boletín Oficial de Aragón” y la fecha de publicación. Abajo a la derecha siempre un número para el csv. En cuanto a la primera página, aparecen entre 0 y 3 títulos para indicar secciones y/o órgano convocante, y debajo aparece en negrita el resumen del artículo. Para obtener el texto, se corta por debajo del encabezado, por encima del pie, a la derecha de la banda azul y a la izquierda del número de csv.

Entre sí, la estructura de los textos de los BOAs no suele variar tanto como la de los BOEs, especialmente en cuanto a anexos se refiere (en el BOA suelen aparecer únicamente las plantillas a cumplimentar).

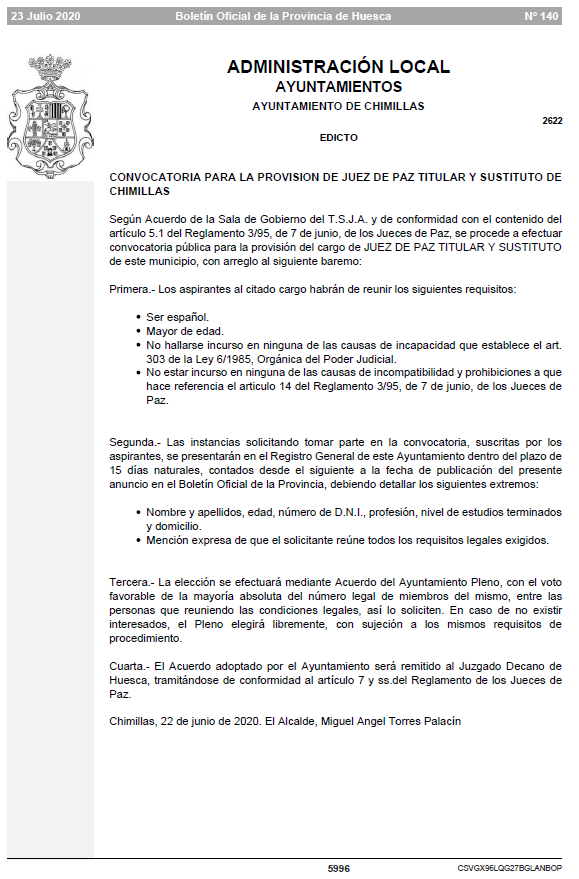


Ilustración 6. PDF del BOPH

En cuanto al BOPH, el encabezado es similar al del BOA pero con un orden contrario y otro formato de fecha. Posteriormente, en la primera página, cuenta con un número muy diverso de títulos, referidos a órganos convocantes, y uno al tipo del artículo. Posteriormente aparece el resumen, y finalmente el título. En el pie aparece el número del csv. Para obtener el texto, se corta por debajo del encabezado, por encima del pie y a la derecha de la banda gris.

En todos los boletines provinciales la estructura del texto es muy variable ya que, mientras que la administración de la provincia realiza convocatorias bastante pormenorizadas, ayuntamientos pequeños en ocasiones redactan un par de párrafos como bases. No suelen aparecer anexos.

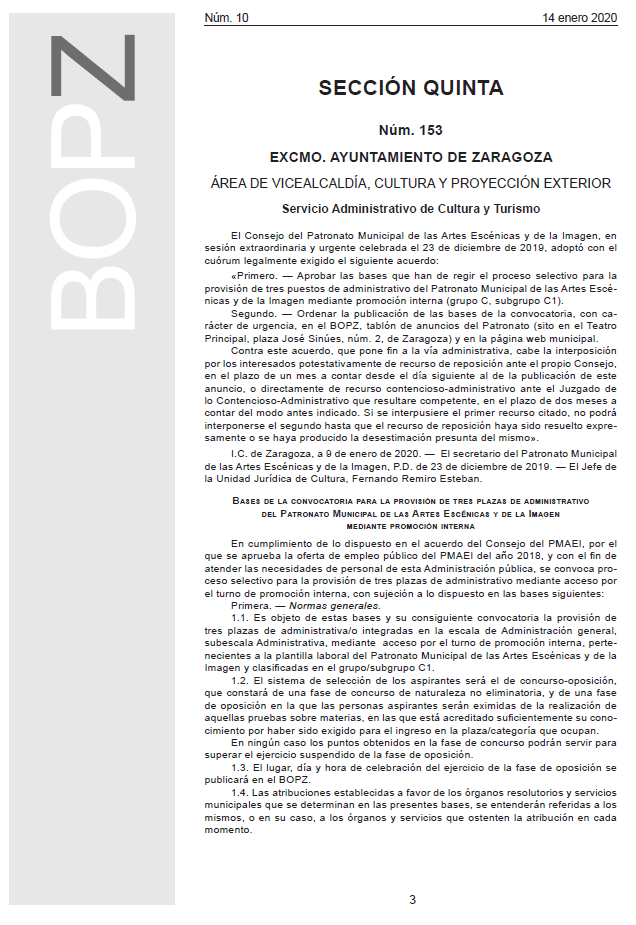


Ilustración 7. PDF del BOPZ

El encabezado del BOPZ no cuenta con texto identificándolo y es más pequeño, ya que la banda izquierda abarca más espacio. Tampoco hay número de csv. En cuanto a la primera página, aparece información sobre la sección (diferente a la sección que hemos mencionado antes), el número y uno o múltiples órganos convocantes. No hay resumen. Para obtener el texto, se corta por debajo del encabezado, por encima del pie y a la derecha de la banda gris.



Ilustración 8. PDF del BOPT

El caso del BOPT es el más diferente de todos, y aumenta la complejidad. En su encabezado muestra el número, la fecha de publicación y la página del sumario en la que aparece el artículo. Tampoco aparece título. Se corta evitando el encabezado y teniendo en cuenta la problemática que se va a explicar.

El punto diferenciador es que, cuando se descarga un artículo, lo que se descarga es la sección (entiéndase aquí sección como corte) del sumario del BOPT del día indicado de todas las páginas en las que haya parte del artículo tratado. Es decir, si una convocatoria empieza a mitad de página en el sumario (como en la Ilustración 8), aparece texto del artículo anterior. Por simetría, si termina a mitad de página, aparece el inicio del artículo siguiente y, si dos artículos caben en una misma página, el PDF resultante para ambos será idéntico, como sería el caso de la Ilustración 9 (que realmente forma parte del artículo de la propia Ilustración 8):

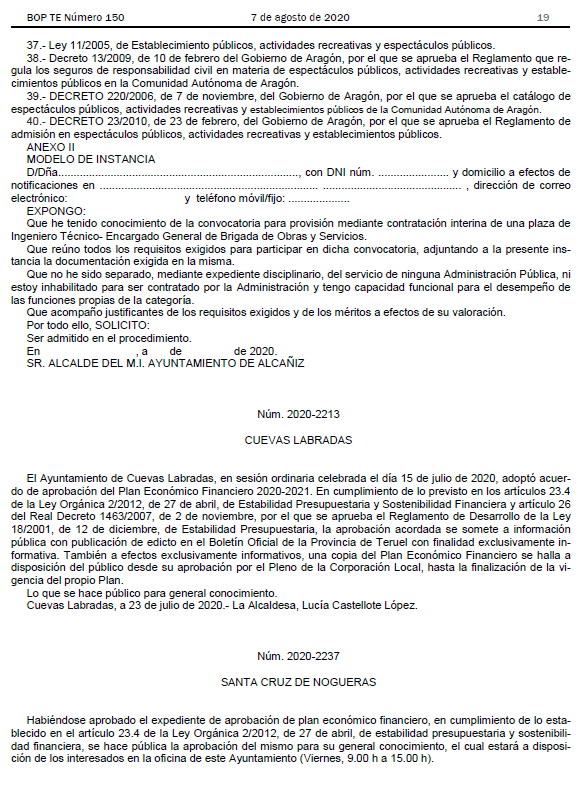


Ilustración 9. PDF del BOPT que aparece en tres artículos

El BOA y los provinciales no tienen formato XML de forma natural, pero sí es posible obtener un XML Sumarizado a partir de la parametrización de la búsqueda. La estructura de dicho XML es la que sigue:



Ilustración 10. XML de boletines aragoneses

A partir de este XML se crean varios (uno por registro), donde la estructura es similar: simplemente desaparece el nivel de la etiqueta <registro>. Podemos ver que hay campos completos como el órgano convocante o el título, y también se tiene el texto o la url al pdf, que necesitan de procesamiento.

El HTML de estos cuatro tipos de boletines es como el de la Ilustración 11. Cabe destacar que, en el caso de los provinciales, la sección viene en blanco. En el caso de Teruel, el título es diferente aquí (en el PDF no hay título). En el caso de Huesca, el título es el inicio del título, finalizado con “…”, y en el caso de Zaragoza el título es el inicio del texto, finalizado con “…”.



Ilustración 11. HTML de boletines aragoneses

# Ingesta de los artículos

En esta sección se van a explicar las herramientas que BOE y BOA ofrecen para recuperar los artículos de sus boletines, y cómo se va a realizar la ingesta utilizándolas.

## Portal del Boletín Oficial de Aragón

### Investigación y obtención manual de los artículos

Este sitio web (<http://www.boa.aragon.es/>) ofrece los boletines oficiales de Aragón y los provinciales de dicha comunidad autónoma además de otros servicios, como puede verse en la figura siguiente:



Ilustración 12. Home del Portal del BOA

De forma manual, es posible realizar búsquedas por fechas, de texto libre o avanzadas. Estas últimas ofrecen diversos parámetros modificables:

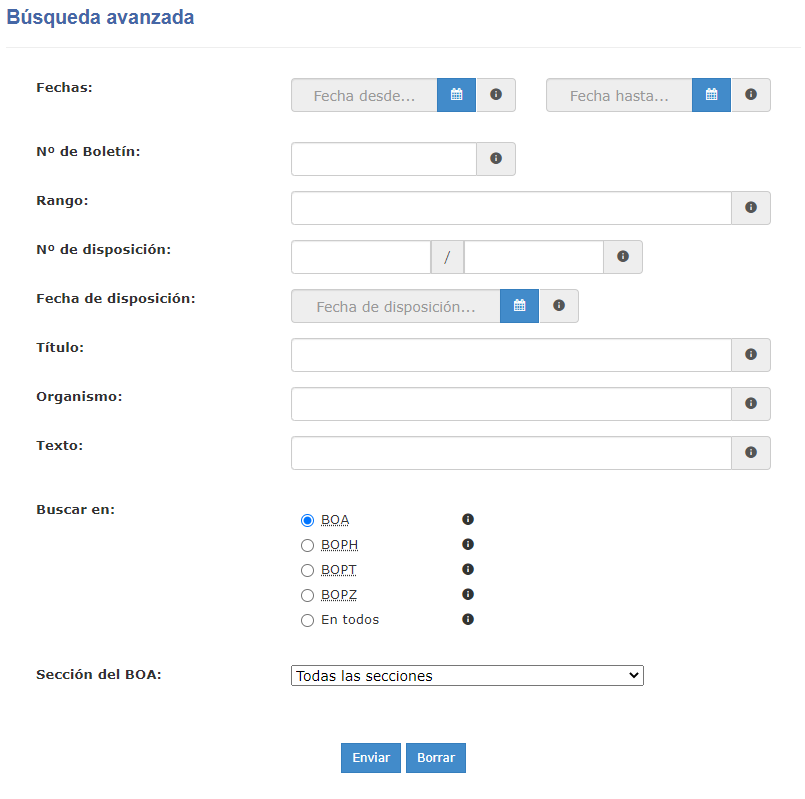


Ilustración 13. Buscador avanzado del BOA

Quizá el parámetro que más interesa para este proyecto es el de la Sección. Al indicar la sección “II. Autoridades y personal”, aparece un nuevo parámetro, que corresponde a la Subsección, y cuyos valores son “II.a. Nombramientos, situaciones e incidencias” y “II.b. Oposiciones y concursos”. Son de estas dos subsecciones de donde se van a obtener aperturas y cierres de convocatorias del BOA. (Estos parámetros de secciones no sirven para los provinciales, que no dividen sus artículos en secciones).

Toda la información indicada en estos parámetros va a ir a parar a la URL resultado, una vez cliquemos “Enviar”. Este enlace nos llevará a una lista en la que aparece información básica de los 100 últimos artículos que coincidan con los parámetros indicados, con enlaces accesibles a dichos artículos:

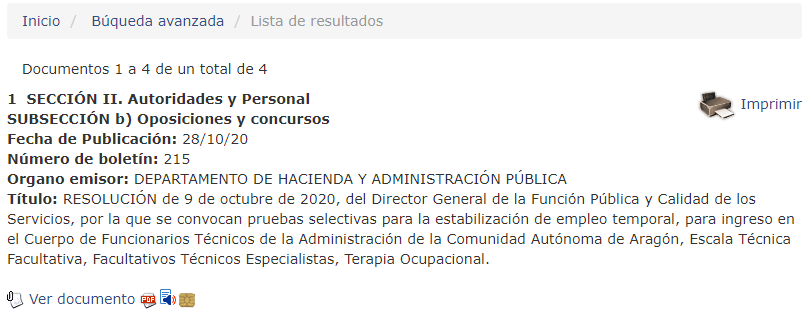


Ilustración 14. Lista de resultados del buscador avanzado del BOA

Si clicamos en “Ver documento”, vamos a una página web donde se nos muestra el artículo al completo, junto con enlaces al pdf y más información:

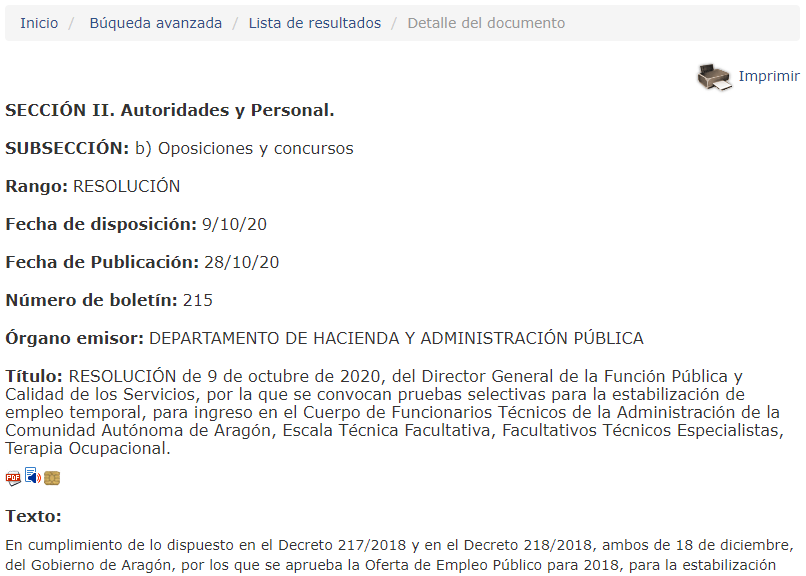


Ilustración 15. Detalles del documento del BOA

La complejidad es la siguiente: los artículos no tienen enlaces únicos que lleven a ese mismo artículo siempre, sino que sus enlaces tienen los mismos parámetros que se han buscado (obteniendo una lista de artículos), y añade un parámetro indicando la posición del artículo en la lista. Esto lo que quiere decir es que, por ejemplo, si se buscan todos los BOAs (sin indicar parámetros) obtendremos una lista y, al acceder al primer artículo de la lista, obtendremos un enlace que devolverá el artículo más reciente del BOA. De esta forma, si nos guardamos el enlace y accedemos a él al día siguiente, el artículo que se nos mostrará será otro: el más reciente del BOA en ese momento.

Para sortear esta dificultad en lugar de obtener los artículos mediante la búsqueda avanzada se ha procedido a obtenerlos mediante la fecha. De esta forma también se obtienen enlaces relativos, pero es mucho menos probable que se introduzca un nuevo artículo en una fecha anterior.

### Obtención automática de los artículos

Continuando la investigación, se encontró el [Buscador Open Data](http://www.boa.aragon.es/#/opendatabuscador) del sitio web del BOA, que también ofrece una búsqueda avanzada, pero el resultado es devuelto en un fichero XML o en un fichero JSON. Estudiando la URL que utiliza, se ha visto que la única diferencia con el buscador del apartado anterior es el valor de un parámetro, por lo que el estudio que se realizó de los parámetros seguía siendo útil:

#### Parámetros y valores principales de las URLs

En este apartado se van a enumerar y explicar los parámetros más útiles, así como los valores que más interesan, de las URLs del sitio web del BOA. El prefijo de la URL al que se deben incorporar los parámetros es: [*http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI*](http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI).

* CMD
  + VERLST: Cuando se quiere mostrar o guardar una lista.
  + VERDOC: Cuando se quiere mostrar o guardar un documento.
* DOCS
  + Indica qué documentos se va a mostrar (si SEC es VERLST), con el formato desde-hasta (por ejemplo, para ver los primeros 100 documentos, sería 1-100)
* DOCR
  + Indica la posición del documento que se quiere ver (si SEC es VERDOC).
* SEC
  + OPENDATABOAJSON: Cuando se quiere obtener el resultado en un JSON.
  + OPENDATABOAXML: Cuando se quiere obtener el resultado en un XML.
  + BUSQUEDA\_AVANZADA: Cuando se está utilizando la búsqueda avanzada, como se ha visto en el apartado 2.1.1.
  + BUSQUEDA\_FECHA: Cuando se está utilizando la búsqueda por fecha.
* OUTPUTMODE (acorde con el SEC en sus dos primeros casos, aunque parece no utilizar realmente este parámetro).
  + XML
  + JSON
* PUBL-GE / PUBL-LE
  + Son los límites superior e inferior, respectivamente, de las fechas de publicación. Recupera las fechas intermedias o iguales. En formato AAAAMMDD.
* NUMB, TITU-C, FDIS-C, TITU, ORGA-C, TEXT-C
  + Con valores textuales, son los parámetros textuales que pueden verse en la Ilustración 13.
  + Número de boletín, Número de disposición, Fecha de disposición, Título, Organismo y Texto, respectivamente.
  + Si se dejan en blanco no los tiene en cuenta.
* SECC-C
  + BOA%2Bo%2BDisposiciones%2Bo%2BPersonal%2Bo%2BAcuerdos%2Bo%2BJusticia%2Bo%2BAnuncios
  + BOA u otras variantes quitando secciones del anterior ítem dan diferente salida, incierta.
  + zaragoza
  + huesca
  + teruel
* SECC (solo si SECC-C es del BOA, no recomendada con BOPs ya que no clasifican por secciones)
  + Personal: Es la sección “II. Autoridades y personal”
* SUBS-C (dependiente de la anterior)
  + Nombramientos: Es la subsección “II.a. Nombramientos, situaciones e incidencias”.
  + Oposiciones: Es la subsección “II.b. Oposiciones y concursos”
* MATE-C
  + Son los temas de los que tratan los artículos.
  + No es recomendado, ya que hay muchos artículos para los que no se indica el tema.
  + 02CEP para Convocatorias de empleo público.

#### Fichero XML devuelto

Una vez estudiados los parámetros, se va a estudiar el fichero XML que devuelve el buscador open data.

Lo primero que se observa es que está codificado con *ISO-8859-1* y que el elemento raíz es <documento>. Dentro del mismo, cada artículo que coincidía con los valores de los parámetros indicados aparece aquí bajo una etiqueta <registro>, con información básica del artículo:

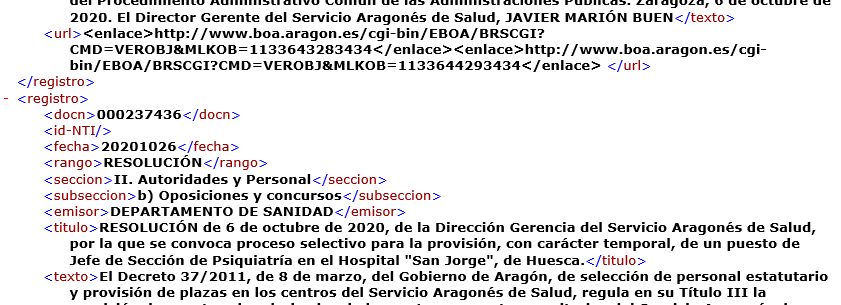


Ilustración 16. XML Sumarizado del BOA

Como se puede ver en la Ilustración 16, ya se nos adelanta la fecha, el rango, la sección, la subsección, el órgano convocante, el título, la URL al PDF y se nos muestra el texto.

El campo más problemático es la URL que, como se puede ver, además de venir mal formateada (siempre sale entre etiquetas <enlace>, que no son etiquetas del XML propiamente dicho, si no que vienen como texto plano), a veces viene más de una. En todas las ocasiones que se ha comprobado, la que dirige al PDF es la primera, mientras que la segunda o no funciona o descarga un fichero llamado BRSCGI que no tiene formato. Por esta razón, se extraerá únicamente el primer enlace.

#### Flujo de recuperación de artículos

Una vez estudiados los parámetros del enlace y los campos del XML devuelto, se procede a explicar el flujo que va a seguir la recuperación automática de los documentos en los tres formatos (XML, HTML y PDF). Para hacerlo más entendible, se explicará mediante un ejemplo.

Pongamos que se quieren recuperar los artículos del BOA de la subsección II.b. del día 3 de septiembre de 2020:

1. Se configura la URL con los parámetros indicados en el enunciado del ejemplo, además de los que se pondrán por defecto, como la recuperación por XML o el comando VERLST: <http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD=VERLST&BASE=BZHT&DOCS=1-100&SEC=OPENDATABOAXML&OUTPUTMODE=XML&SORT=-PUBL&SEPARADOR=&%40PUBL-GE=20200903&%40PUBL-LE=20200903&NUMB=&RANG=&TITU-C=&FDIS-C=&TITU=&ORGA-C=&TEXT-C=&SECC-C=BOA%2Bo%2BDisposiciones%2Bo%2BPersonal%2Bo%2BAcuerdos%2Bo%2BJusticia%2Bo%2BAnuncios&SECC=Personal&SUBS-C=Oposiciones&MATE-C=>
2. Al lanzarlo, se recupera el XML sumarizado. Para cada <registro> del mismo, se guarda por separado un XML, de forma que ya se dispondría del formato XML de cada artículo.
3. Asimismo, por cada uno de estos, se recupera el PDF utilizando la etiqueta <url> y procesándola debidamente, como ya se ha comentado. Al lanzar este enlace, se recupera y guarda el artículo en formato PDF.
4. Para recuperar, por cada registro, el HTML y su URL (que servirá como enlace a la convocatoria), es necesario modificar la URL anterior, con los siguientes cambios:
   1. CMD pasa de ser VERLST a VERDOC.
   2. Se sustituye el parámetro DOCS por DOCR, cuyo valor vendrá dado por la posición del artículo dentro del XML Sumarizado, siendo 1 el primero.
   3. SEC pasa de ser OPENDATABOAXML a BUSQUEDA\_AVANZADA.
   4. Se quita el parámetro OUTPUTMODE.

De esta forma, se obtiene el enlace al artículo en formato HTML:

<http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD=VERDOC&BASE=BZHT&DOCR=1&SEC=BUSQUEDA_AVANZADA&SORT=-PUBL&SEPARADOR=&%40PUBL-GE=20200903&%40PUBL-LE=20200903&NUMB=&RANG=&TITU-C=&FDIS-C=&TITU=&ORGA-C=&TEXT-C=&SECC-C=BOA%2Bo%2BDisposiciones%2Bo%2BPersonal%2Bo%2BAcuerdos%2Bo%2BJusticia%2Bo%2BAnuncios&SECC=Personal&SUBS-C=Oposiciones&MATE-C=>

Así, se han obtenido todos los artículos definidos en el enunciado, en los tres formatos requeridos, finalizando la ingesta automática.

## Portal del Boletín Oficial del Estado

Este sitio web (<https://boe.es/>) ofrece, entre otros, el Boletín Oficial del Estado. De forma manual, es posible obtener el BOE del día o buscar artículos sueltos, tanto mediante calendario como utilizando una búsqueda avanzada (Ilustración 17), que ofrece parámetros similares al del portal del BOA, y mediante el cual se fueron obteniendo artículos de forma manual para el análisis exploratorio. Cabe destacar que los campos tienen un límite de caracteres desconocido (no es posible introducir un título completo en el campo título para encontrar el artículo rápidamente). Además, busca artículos que coincidan en parte con lo parametrizado, no se adhiere mucho a los valores escritos.

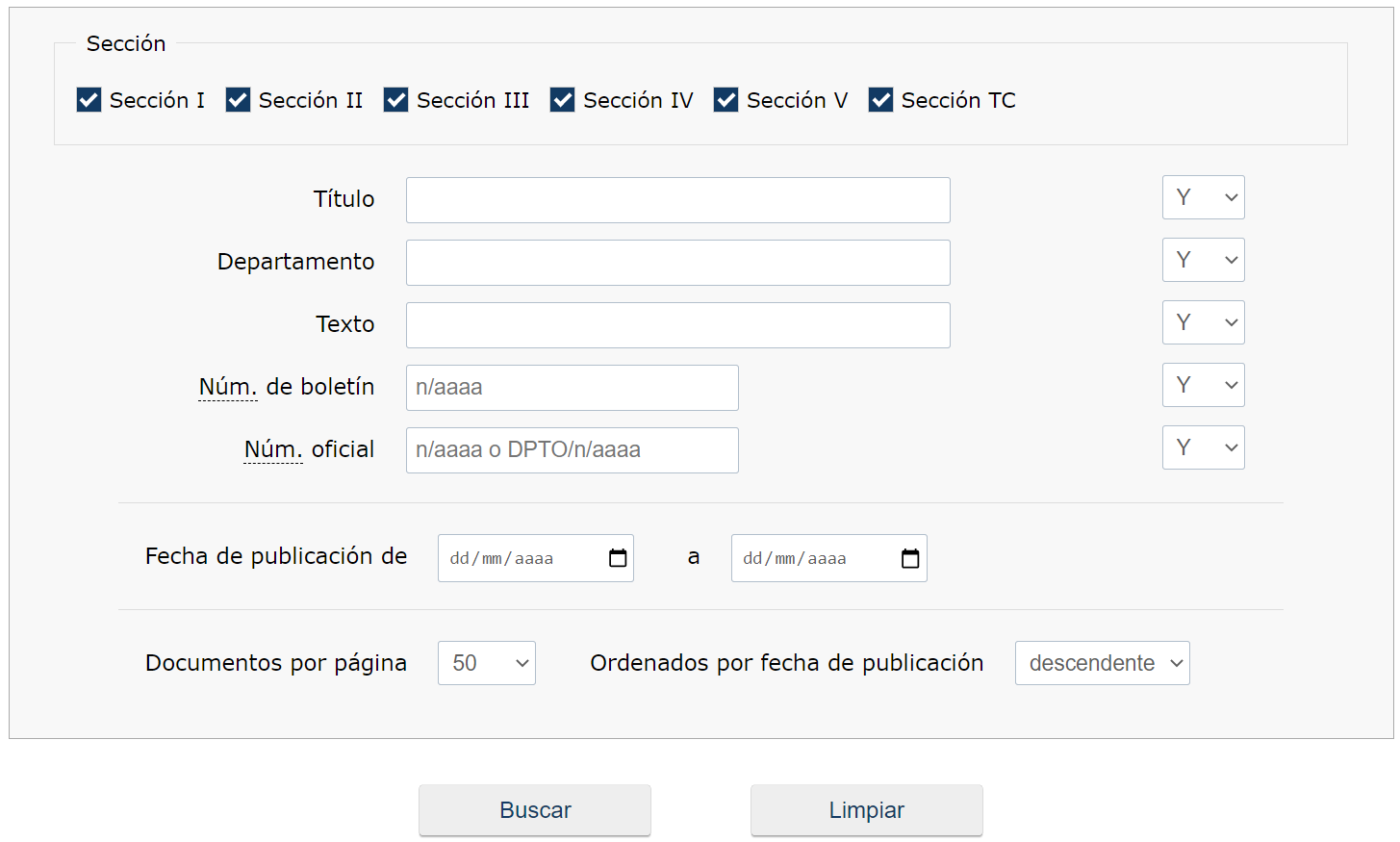


Ilustración 17. Buscador avanzado del BOE

Sin embargo, y a diferencia de éste, el portal del BOE también ofrece un API REST, mediante la que es posible obtener un XML sumario, con todos los artículos del día indicado. Esta API REST es poco flexible (solo permite decidir el día), pero es suficiente para nuestros intereses. Como alternativa, se podría haber realizado *HTML Scrapping* en los resultados del buscador del portal, algo que habría aumentado el riesgo sin obtener ventajas reales.

Se ha estudiado la información del XML Sumario que se recupera como resultado de la llamada a la URL gracias a la [documentación encontrada](https://boe.es/datosabiertos/documentos/SumariosBOE_v_1_0.pdf). En resumen:

* El elemento raíz es <sumario>.
* Dentro del mismo, tenemos el elemento <meta>, donde se almacenan los metadatos del mismo, y <diario>, que tiene un subelemento <sumario\_nbo>, referido al BOE de dicho día, y subelementos <seccion>, con el que será posible recuperar únicamente los de las secciones que nos interesen.
* Dentro de cada sección, se van dividiendo por departamentos y epígrafes. Dentro de estos epígrafes, ya se ubican los elementos <item>. Cada uno de ellos es un artículo.
* Dentro de <item> tenemos el título y los sufijos de las URLs a cada uno de los tres formatos que nos interesan.

Conocido todo esto, el flujo de recuperación no es complejo de seguir (pongamos que queremos recuperar los artículos del 3 de septiembre de 2020):

1. Se recupera el XML Sumario a partir de la URL parametrizada con la fecha, que quedaría: <https://boe.es/diario_boe/xml.php?id=BOE-S-20200903>.
2. En dicho fichero, basta con, por cada ítem (en la ruta sumario/diario/seccion/departamento/epigrafe), concatenar el prefijo “https://boe.es” y el sufijo indicado en el ítem para cada uno de los formatos, y recuperar el resultado. Ya se tendrían los tres formatos indicados para cada artículo del día.
   * En caso de querer filtrar la subsección “2B” (Oposiciones y concursos), bastaría con coger únicamente los ítems que estuvieran dentro de la <seccion> cuyo atributo num es igual a “2B”. Similar para “2A” (nombramientos).

# Corpus utilizado para el análisis

En esta sección se va a detallar el corpus utilizado para realizar el análisis del presente documento. Cabe destacar que se han estudiado más documentos, ya que constantemente se buscaban distintos casos, formatos y temas en ambos buscadores, observando y anotando detalles que pueden no aparecer en los documentos que se han descargado y utilizado como corpus. Finalmente, y de entre todos los documentos comprobados, se guardó un corpus rico en formatos, casos e incluso fechas (a lo largo de 2020). En la tabla que se expone a continuación, se muestra un resumen del número de artículos finalmente guardado y analizado en detalle para este análisis exploratorio.

|  |  |  |  |
| --- | --- | --- | --- |
|  | EMPLEO | OTRO | TOTAL: |
| BOE | 14 | 6 | 20 |
| BOA | 14 | 6 | 20 |
| BOPH | 5 | 3 | 8 |
| BOPZ | 5 | 3 | 8 |
| BOPT | 5 | 3 | 8 |
| TOTAL: | 43 | 21 | **64 documentos** |

Por cada uno de estos artículos, se almacenó su pdf, html, xml y url al propio artículo. Todos estos formatos se recogían de forma manual, teniendo que obtener todos los xml aragoneses extrayéndolos del XML sumarizado, que se comenta en la sección de ingesta.

La categoría “OTRO” refiere a los artículos que no contenían nada relacionado con ofertas de empleo. Estos también se han escogido con riqueza de temas y forma de sus textos.

Además, se almacenaron entre 3 y 5 sumarios de cada uno de los boletines para comprender mejor tanto lenguaje utilizado como estructura.

# Estudio de los campos

En esta sección se van a enumerar y explicar los campos a extraer e insertar en la base de datos. También se va a comentar sucintamente algunas de las formas más habituales de aparición de estos en los diferentes documentos estudiados.

## Campos básicos

En esta sección los campos que se van a comentar son los que mencionaba el pliego explícitamente, en el apartado 2.2.c, siendo totalmente necesarios para el objetivo del proyecto.

### Fuente de los datos

Es el campo que indica de qué tipo de boletín se ha extraído la convocatoria, pudiendo ser BOE, BOA, BOPH, BOPZ o BOPT. La obtención de este campo es trivial, ya que en la ingesta se sabe de qué tipo de boletín se están recuperando los artículos.

### Fecha de publicación

Es el campo que indica la fecha de publicación del artículo que contiene la convocatoria, con el formato “dd/mm/aaaa”. Su obtención también es trivial, ya que es necesario indicar la fecha de los artículos a recuperar en la ingesta. Sin embargo, también es posible recuperarla de los ficheros XML, HTML o PDF. La extracción de la fecha en este último es más compleja, pero también se ha desarrollado esta capacidad para los cinco tipos de boletines. Se basa en ubicarla visualmente en cada tipo de boletín, estudiar los puntos de corte para extraer únicamente dicho atributo y, ya de una forma automática, recortarla con la herramienta de extracción teniendo en cuenta dichos puntos de corte.

En cuanto a su posición dentro de los diferentes boletines:

* En el BOE, la fecha viene indicada en el centro del boletín, en su parte superior, pero debajo del encabezado “BOLETÍN OFICIAL DEL ESTADO”. Viene precedida por el día de la semana y con el mes indicado textualmente, con preposiciones incluidas.
* En el BOA, viene con el formato “dd/mm/aaaa” en la esquina superior derecha.
* En el BOPH, viene en la esquina superior izquierda, con el mes escrito de forma textual.
* En el BOPZ, viene en la esquina superior derecha, con el mes escrito de forma textual.
* En el BOPT, viene en la parte superior central, con el mes escrito de forma textual y con preposiciones.

### URI ELI de la norma

Es la URI que identifica unívocamente una norma. Se almacena únicamente si está definida, algo que ocurre un número muy reducido de veces en el BOE, nunca en los provinciales y en pocas ocasiones en el BOA (y ninguna de ellas relacionada con convocatorias de empleo).

#### Formación

Se ha [estudiado](https://elidata.es/documentacion_tecnica/especificacion_tecnica.php) el formato de las mismas, así como los parámetros que las conforman para poder posteriormente formarlas de forma manual en caso de necesitarlo.

El formato habitual es el siguiente:

/eli/{jurisdiction}/{type}/{year}/{month}/{day}/{number}/{version}/{version\_date}/{language}/{format}

En cuanto a los parámetros y valores que más interesan para este proyecto:

* jurisdiction. Existe una tabla de vocabulario controlado.
  + ‘es’ para Estado.
  + ‘es-ar’ para Aragón.
* type. Existe una tabla de vocabulario controlado.
  + ‘o’ para orden.
  + ‘a’ para acuerdo.
  + ‘res’ para resolución.
* year, month, day
  + En formato AAAA, MM y DD, respectivamente.
  + Es siempre la fecha de firma, no la de publicación.
* number
  + Práctica variable.
    - Órdenes: Sin pautas uniformes de numeración.
    - Resoluciones: No numeradas ni por el Estado ni por la mayoría de comunidades.
  + El año no se incluye en el number.
  + En casos como las órdenes, con número oficial, se considera como number el código alfanumérico completo, en minúsculas y eliminando la barra intermedia.
  + En el caso de números oficiales duplicados (mismo type, no único, en misma fecha), se añadirá al segundo número el sufijo (b), al tercero(c), etc. No se utilizarán ni (a) ni (ñ).
  + En casos como las resoluciones, sin número oficial, se utilizará una secuencia para cada día que inicie en 1, siendo el number (1),(2), (3)...
* version. Existe una tabla de vocabulario controlado.
  + ‘dof’ para recurso legal inicial (el que más nos interesa, ya que las convocatorias no se consolidan).
  + ‘cer’ para recurso legal corregido.
* version\_date
  + Formato AAAAMMDD.
  + Solo para recursos consolidados.
* language. Existe una tabla de vocabulario controlado.
  + ‘spa’ para español.
  + ‘mul’ para multilingüe.
  + Los bilingües van con guión intermedio (‘cat-spa’ para bilingüe catalán-español).
* format. Sigue los valores según [IANA](http://www.iana.org/assignments/media-types).

Un par de ejemplos que se crearon para comprobar que se había entendido (que no funcionan, ya que dichos artículos no tienen URIs definidas):

* BOA: “ORDEN HAP/815/2020, de 17 de agosto, por la que se convoca la provisión, por el sistema de libre designación, de un puesto vacante en el Departamento de Ciudadanía y Derechos Sociales.”
  + /eli/es-ar/o/2020/08/17/hap815/dof/spa/pdf
* BOE: “Resolución de 26 de agosto de 2020, de la Subsecretaría, por la que se convoca proceso selectivo para ingreso, por el sistema general de acceso libre y promoción interna, en el Cuerpo de Ingenieros Técnicos de Obras Públicas.”
  + /eli/es/res/2020/08/26/(1)/dof/spa/pdf

La falta de identificador unívoco de resoluciones es un gran peligro, ya que no se sabe a priori qué resoluciones forman parte de la secuencia numérica y el orden de estas en ella.

#### Modo de obtención

En el BOE, tanto en el XML como en el HTML (no así en el PDF, donde no aparece), hay un campo específico reservado para la misma, de forma que se obtiene del mismo.

* En el XML, la etiqueta bajo la que aparece es <url\_eli>.
* En el HTML, aparece tras la etiqueta de enlace con título “Permalink ELI”. Por ejemplo:



Ilustración 18. Detalle de la ELI URI en HTML del BOE

También se ha comprobado que, conforme se van reduciendo los directorios de esta URI, se van obteniendo listados de URIs coincidentes, a modo de API, algo que con las del BOA no ocurre.

En el BOA, sin embargo, lo único relacionado con la URI ELI es el [listado de artículos que sí tienen URI ELI](http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD=VERLST&BASE=BOLE&DOCS=1-1000&SEC=ELI&SORT=-PUBL&SEPARADOR=&SECC-C=generales&CODR-C=11&@FDIS-GE=20160101). A ésta, se puede acceder desde la [página de inicio](http://www.boa.aragon.es/#/) del Boletín Oficial de Aragón > Base de datos del BOA > ELI – Aragón > Órdenes (desde 2016). Este último paso se debe a que las convocatorias son únicamente resoluciones u órdenes, nunca decretos o leyes. Podemos observar que, a día 20 de octubre de 2020, hay un total de 1041 órdenes identificadas unívocamente desde principios del 2016. Por ello, y para poder tener el listado en una única página HTML, es posible modificar la URL, cambiando el valor del parámetro DOCS (por defecto 1-1000), a 1-100000 o algo de un orden similar. En este listado, aparecen las URI ELIs existentes de forma textual, con hipervínculos que de hecho ni siquiera las utilizan, por lo que no existe una garantía por su parte de que las URIs mostradas funcionen. En la Ilustración 19 puede verse un caso a modo de ejemplo. En él se ve que el enlace (indicado en la etiqueta a, en el atributo href, en azul), es un enlace normal del BOA, en lugar de utilizar el que se muestra más abajo tan solo textualmente, que es la URI ELI.

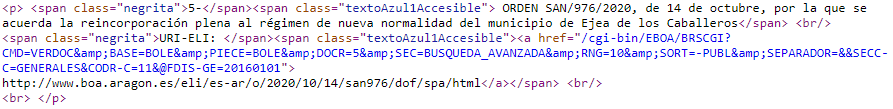


Ilustración 19. Detalle de la ELI URI en HTML del listado del BOA

### Título

El título, que habitualmente es el primer párrafo de texto del artículo, funciona a modo de *abstract* o resumen de lo que trata el artículo en su totalidad. Sin embargo, no es necesario recurrir a su extracción del texto, ya que existe un campo en los XML del que se puede obtener de forma sencilla y sin lugar a error:

* En los XML de los artículos recuperados del BOE, la ruta del XML a este atributo es documento/metadatos/titulo.
* En los XML de los artículos recuperados tanto del BOA como de los provinciales, su ruta es simplemente documento/titulo.

Por otra parte, para obtenerlos de los HTML:

* En el BOE, aparece un título h3 cuya clase es “documento-tit”, en la que aparece como texto el título:



Ilustración 20. Detalle del título en HTML del BOE

* En los cuatro boletines de Aragón en cambio, no hay una etiqueta que nos indique que se está hablando del título, si no que se nos indica mediante texto que el título viene a continuación:



Ilustración 21. Detalle del título en HTML de los boletines aragoneses

En cuanto a la obtención mediante PDF:

* En el BOE, aparece bajo los títulos de sección, subsección y órgano convocante, y tras el número del artículo.
* En el BOA, puede aparecer lo primero o bajo títulos (se han llegado a encontrar incluso 4 títulos).
* En el BOPH, también puede aparecer bajo diferente número de títulos previos, en función de la especificación del órgano convocante.
* En el BOPZ hay también un número variable de títulos previos, pero realmente no existe un título como el que se busca (el resumen del artículo). En los otros formatos la administración simplemente utiliza el principio del texto e incluye unos puntos suspensivos.
* En el BOPT tampoco hay título y, como se ha comentado, los artículos no comienzan por el principio del documento. En estos, aparece el número identificativo y el ayuntamiento u órgano convocante, y a continuación pasa al texto. En los otros formatos, y a diferencia del BOPZ, aparece un título que no aparece en el PDF y que resume el artículo.

### Puesto ofertado

Uno de los campos más complejos de obtener es el puesto ofertado. Es especialmente crítico debido, además, a su presencia en el identificador de las convocatorias. Se van a detallar los posibles puntos de aparición de este a lo largo de los diferentes tipos de boletines, así como las mayores complejidades encontradas. Primero se van a enumerar y explicar sus ubicaciones en el formato PDF de los distintos boletines.

En el BOA, los principales puntos de aparición del puesto son los siguientes:

* En el título, indicado sin introducción.

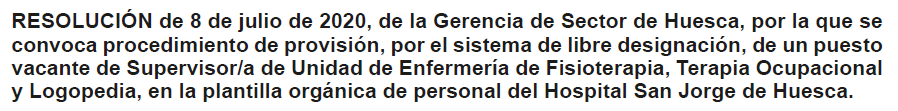


Ilustración 22. Detalle de puesto en BOA

* En el título, tras indicar cuerpo y escala.

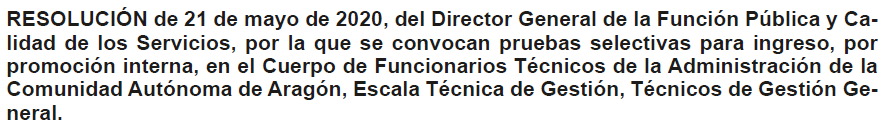


Ilustración 23. Detalle de puesto con cuerpo y escala en BOA

* Al principio del texto, etiquetado con “Denominación:”.

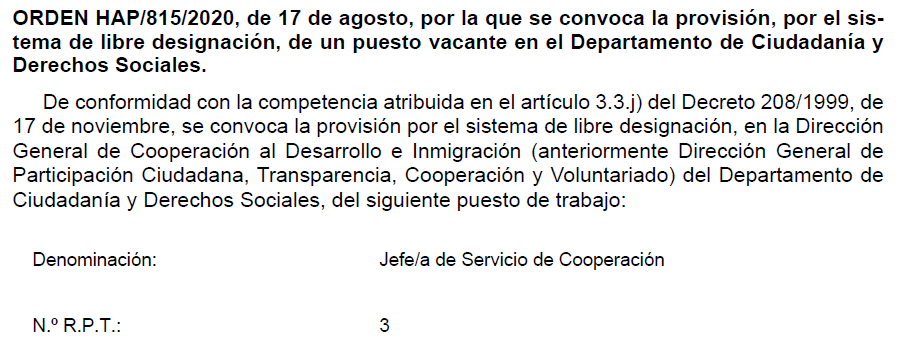


Ilustración 24. Detalle de puesto etiquetado en BOA

* En ocasiones, lo que se ofrece es la incorporación a un cuerpo o a un departamento. En estos casos, hay tres posibilidades:
  + Que no se mencione nada más, relacionado con puestos (se entiende que la incorporación al propio departamento es lo ofertado en sí).
  + Que rediriga a otro sitio para ver los puestos y el número de plazas que se ofertarán.

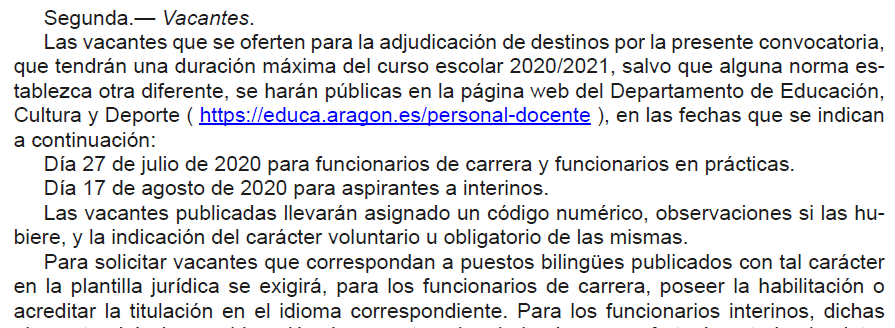


Ilustración 25. Detalle de puestos dirigidos a web en BOA

* + Que aparezcan los detalles de las plazas en tablas en los anexos. Esto ocurre especialmente cuando se ofrece la entrada a algún cuerpo docente. Con los cuerpos docentes, el mayor problema es que no se muestra el puesto como tal, si no que se mencionan materias relacionadas con el mismo, lo que complica la extracción.

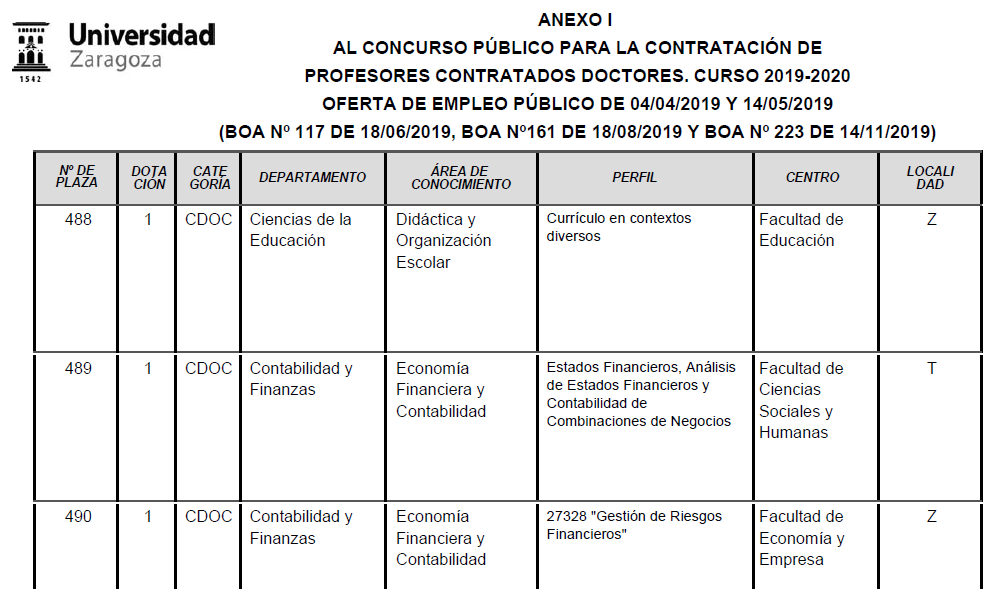


Ilustración 26. Detalle de puestos docentes en tabla de BOA

En cuanto a los boletines provinciales, la información puede venir bastante dispersa (en muchos puntos del documento), pero suelen ser puestos concretos, no incorporación a cuerpos, especialmente en los artículos publicados por ayuntamientos pequeños. En estos, la complejidad radica en que las frases además son muy variables, no existiendo patrones que abarquen un amplio número de casos.

En el BOE los puestos varían más todavía que en el BOA:

* En ocasiones, el puesto viene en el propio título, sin indicaciones previas.

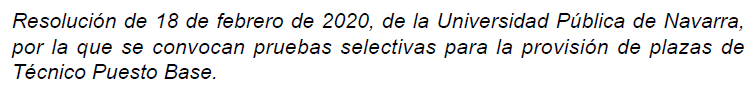


Ilustración 27. Detalle de puesto en BOE

* En otras, viene en el título, precedido por escala o cuerpo.

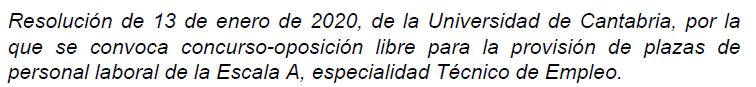


Ilustración 28. Detalle de puesto con escala en BOE

* A veces, en el título se indica únicamente el ingreso al cuerpo o a la escala:
  + En el caso de los cuerpos docentes, a veces indica en anexos mediante tablas o texto los puestos. Como en el BOA, a veces no indica el puesto como tal si no la materia o el departamento. En la Ilustración 29 se muestra texto en el anexo en el que sí aparece el puesto como tal y, en la Ilustración 30, otro caso en el que no es tan sencillo de identificar.

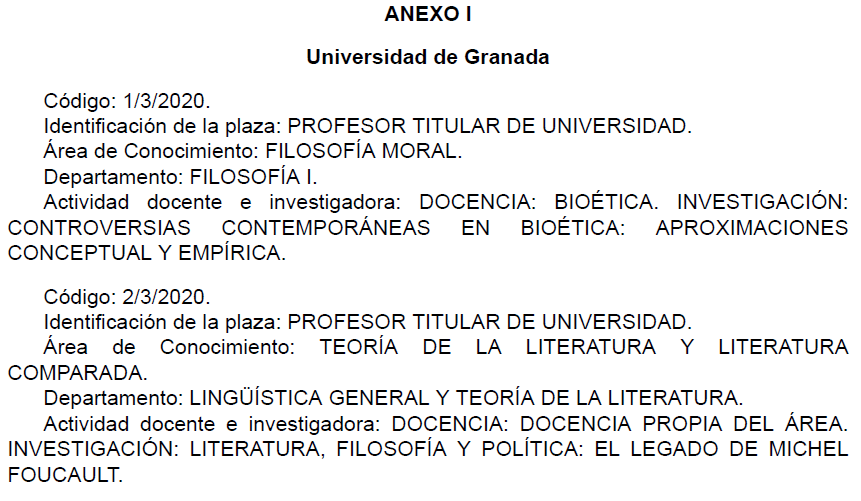


Ilustración 29. Detalle de puestos explícitos de cuerpos docentes en anexo textual de BOE

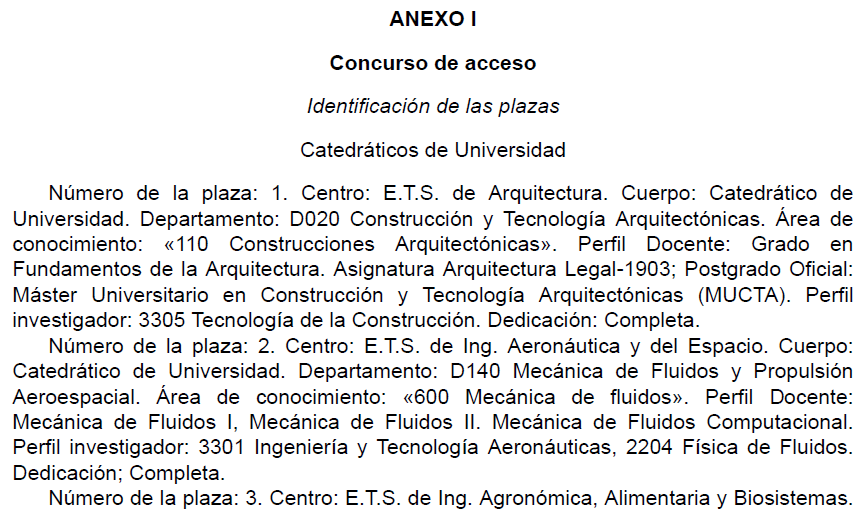


Ilustración 30. Detalle de puestos no explícitos de cuerpos docentes en anexo textual de BOE

* + En otras ocasiones se detallan las plazas en medio del texto, de formas bastante rudimentarias. Esto es muy poco habitual.

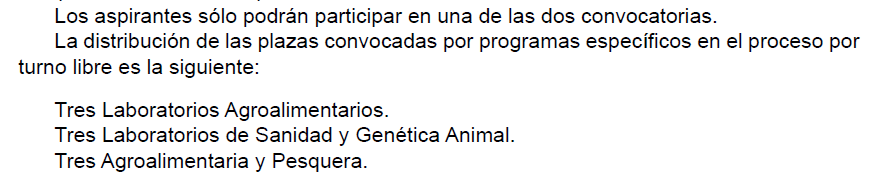


Ilustración 31. Detalle de puestos en medio del puesto, mal explicado, en BOE

* + En otros casos, no se llega a detallar nada sobre los puestos en todo el documento, quedándose únicamente la información del cuerpo o escala al que se ingresaría.

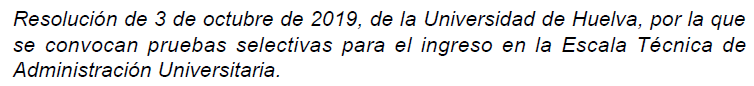


Ilustración 32. Detalle de puesto en el título de BOE sin más información

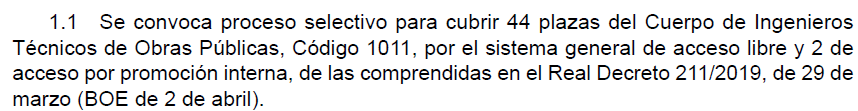


Ilustración 33. Detalle de texto con cuerpo como puesto en BOE

* + En cuanto a los anexos, el texto que identifique al puesto puede ser:
    - Textual:

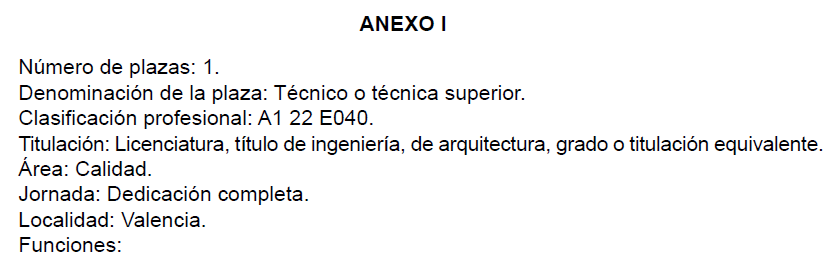


Ilustración 34. Detalle de puesto en anexo textual en BOE

* + - Una tabla horizontal. Hay que identificar correctamente en qué columna aparecen los puestos, mediante la extracción del texto del encabezado.

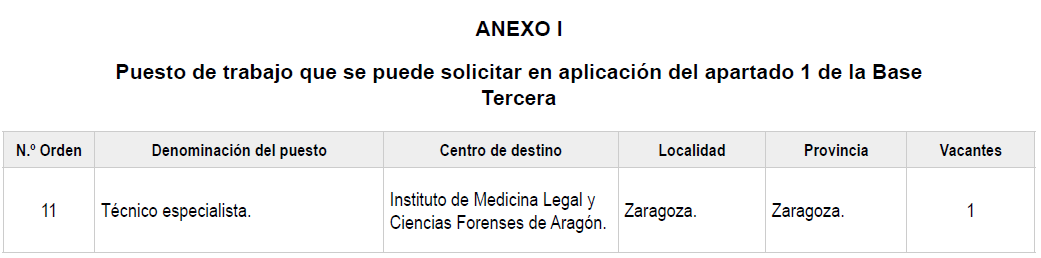


Ilustración 35. Detalle de puesto en anexo con tabla horizontal en BOE

* + - Una tabla vertical. Mismo problema de identificación que tabla horizontal, pero además hay que rotar el PDF para leerlo bien. El problema de que haya tablas de los dos tipos es que a priori no se puede decidir si leer la tabla rotada o sin rotar. Además, en celdas de puestos aparecen frecuentemente textos que no son puestos, si no que sirven para dividir los puestos por gerencias, agencias, territorios, etc (Ilustración 36). En otras ocasiones las tablas no son regulares y tienen celdas que ocupan diversas columnas (Ilustración 37). Además, las diferentes tablas están creadas presuntamente de diferentes formas, ya que presentan diferentes formatos al leerlas.

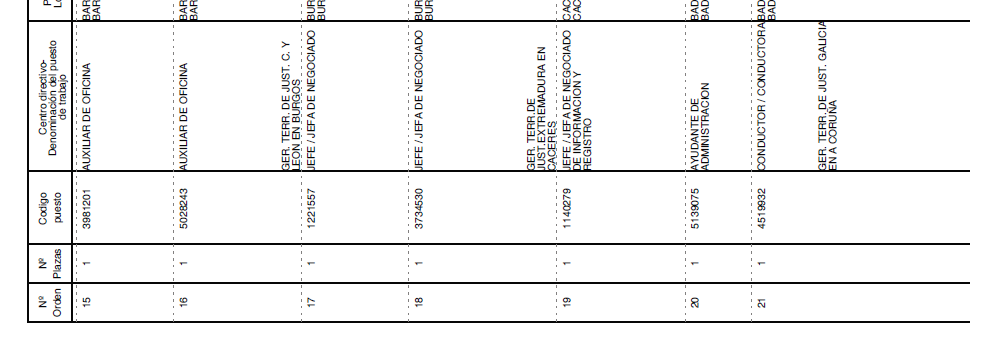


Ilustración 36. Detalle de puestos en anexo con tabla vertical con gerencias intermedias en BOE

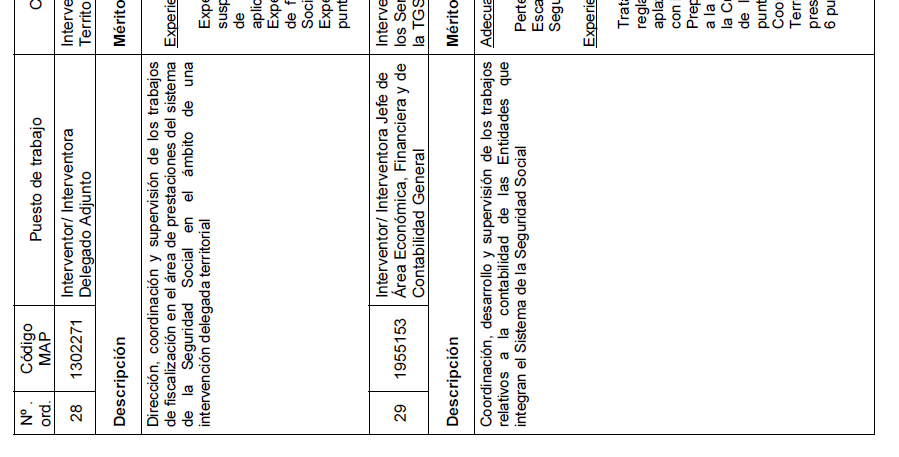


Ilustración 37. Detalle de puestos en anexo con tabla vertical con celdas de varias columnas en BOE

* En ocasiones, si la tabla no está bien cerrada por abajo (con línea visible), se puede perder información, ya que el extractor de texto de tablas puede no interpretar esa última fila.

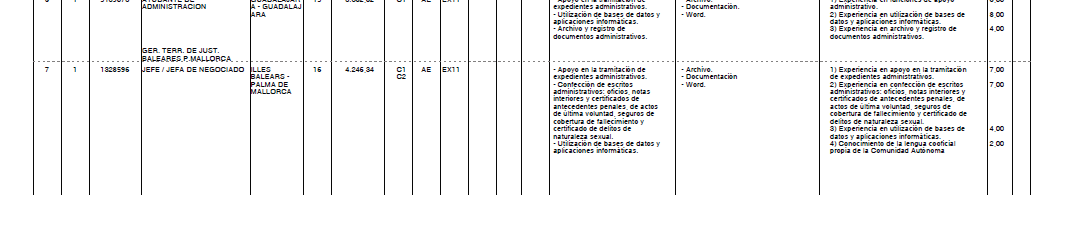


Ilustración 38. Detalle de tabla que no está bien cerrada por la parte inferior

* También se han observado casos en los que, al no caber una tabla en horizontal, en el pdf aparece la propia página apaisada, adaptando encabezado y pie de página.



Ilustración 39. Página apaisada de BOE

Las directrices más generales explicadas hasta ahora sirven también para encontrar los puestos ofertados en los formatos XML y HTML, ya que en general el título y el texto es el mismo. A continuación, se van a detallar algunos casos en los que no lo son o que podrían dar problemas:

* En ocasiones, cuando ocurre lo mencionado acerca de los puestos aparecidos tras la etiqueta “Denominación:” en el PDF puede haber un problema, y es que resulta que esta información bien separada es una tabla sin bordes, y esta tabla desaparece completamente del texto en otros formatos. En la Ilustración 40 puede verse este ejemplo con una parte del PDF y la equivalente en el XML, en la Ilustración 41. Aparecen subrayados los dos mismos trozos de texto.

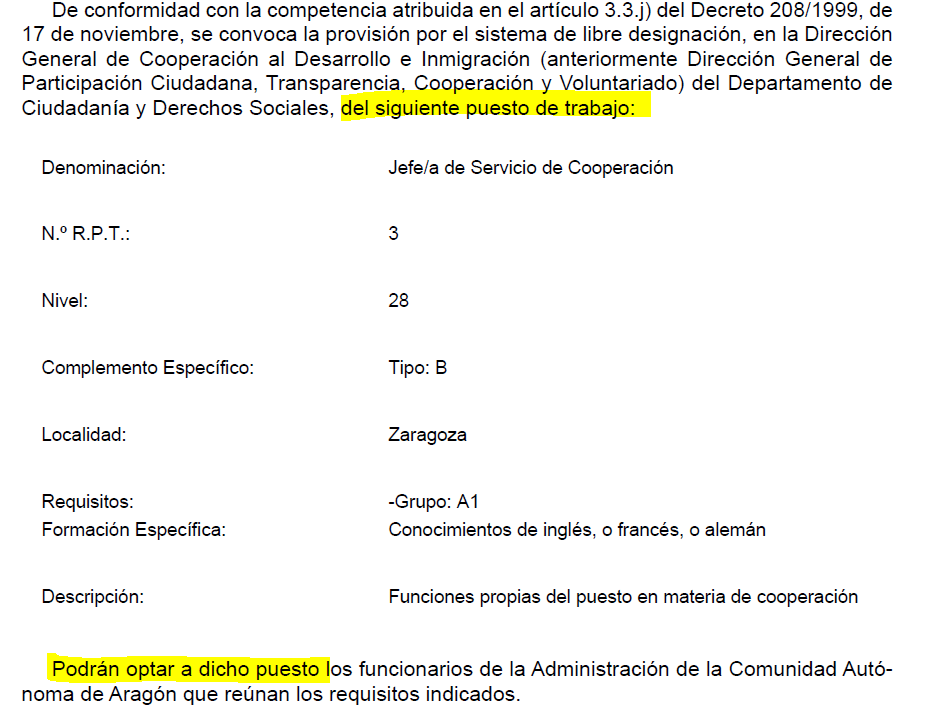


Ilustración 40. Detalle de tabla no explícita desaparecida en el XML de BOA (I)



Ilustración 41. Detalle de tabla no explícita desaparecida en el XML de BOA (II)

Por casos como éste (ya que en los BOAs este formato es bastante habitual, y la información es importante), por defecto se va a leer el texto de los PDFs.

* Las tablas en vertical que aparecen en los anexos de los PDF de los BOEs no aparecen como texto en los otros dos formatos, sino que se han guardado capturas de las mismas y se inserta el sufijo del enlace desde el que se puede recuperar dicha imagen. En el caso del HTML, la clase imagen\_girada se encarga de rotarla 90º.

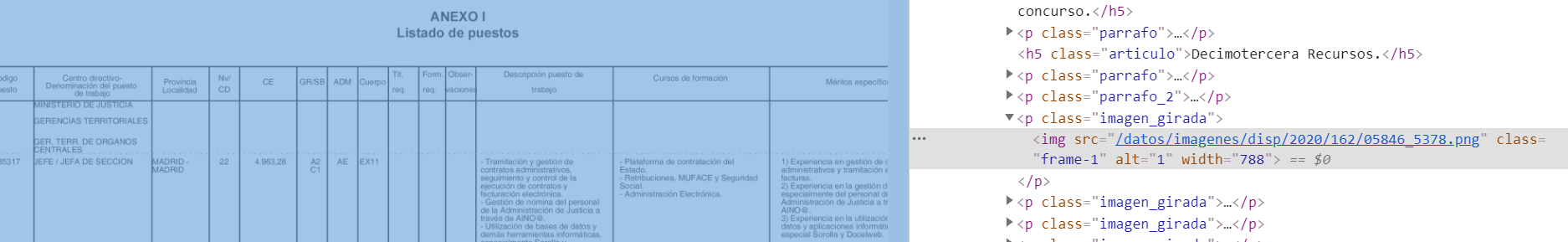


Ilustración 42. Detalle del HTML de la tabla vertical del anexo de BOE

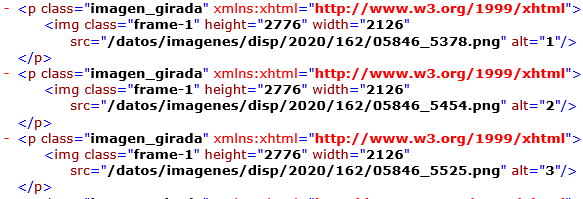


Ilustración 43. Detalle del XML de la tabla vertical del anexo de BOE

* En cambio, las tablas que aparecen en horizontal en los PDF del BOE a veces sí están bien formateadas, mediante tablas de HTML y XML, que se muestran a continuación (ambas son la tabla de la Ilustración 35).

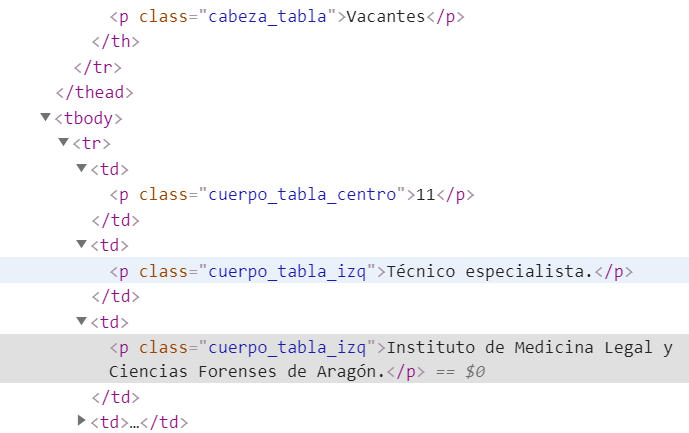


Ilustración 44. Detalle de tabla horizontal en HTML de BOE



Ilustración 45. Detalle de tabla horizontal en HTML de BOE

* Por el contrario, las tablas que aparecen en los PDF del BOA (como la de la Ilustración 26), no aparecen de ninguna manera en los otros dos formatos. Ni se las referencia, ni se las nombra, lo que podría llevar a confusión.

Por estas razones, se opta por leer las tablas de los ficheros PDF.

### Fecha de inicio de presentación de solicitudes

Este campo se forma a partir del campo plazo (apartado 4.3.6) y del campo fecha de publicación (apartado 4.1.2). Del plazo, interesa especialmente saber si se indica que el mismo empieza el día de la publicación o el siguiente, utilizando la fecha de publicación o la siguiente. Se guardará con el formato “dd/mm/aaaa”.

### Fecha de fin de presentación de solicitudes

Este campo se forma a partir del campo plazo (apartado 4.3.6) y del campo fecha de inicio de presentación de solicitudes (apartado 4.1.6). Del plazo, interesa saber el número de días que dura el plazo, y si estos son hábiles o naturales. Si son naturales, este campo será el resultado de la suma del número de días al campo fecha de inicio de presentación de solicitudes. Si son hábiles, hay que realizar dicha suma sorteando los días no hábiles. Se guardará con el formato “dd/mm/aaaa”.

### Enlace a la convocatoria de la fuente oficial

En este campo se almacena la URL a la convocatoria del artículo presente. Las URLs y cómo obtenerlas ya se han analizado en el apartado 2, ya que son necesarias para recuperar todos los formatos. Por ello, esta sección va a centrarse únicamente en la forma de obtención dichos enlaces.

* En el XML Sumario del BOE, durante la ingesta, para cada artículo se indica la url al HTML (al sitio web del artículo), con el texto de las etiquetas ‘urlHtm’. La ruta hasta cada una de ellas es “sumario/diario\_nbo/sección/departamento/epígrafe/item/urlHtm”.
* Para el BOA y los provinciales, con el XML Sumarizado, es necesario crear la URL de forma propia e incluir el parámetro de posición del artículo dentro del mismo.

### Órgano convocante

En este campo se almacena el órgano emisor que ha realizado la convocatoria. Se va a detallar la ubicación de esta información en los diferentes tipos de boletines en todos sus formatos.

Respecto a los PDFs:

* En el BOA, habitualmente, aparece como título, tras el título que indica la sección a la que pertenece el artículo, que en ocasiones no aparece. En otras ocasiones, no aparece ningún título, por lo que no se recuperaría el órgano convocante de este formato.
* En el BOE aparece como tercer título, bajo los de sección y subsección. No es habitual que aparezca más de un título que se refiera al órgano.
* En el BOPH aparece como primer título (sin secciones por encima), pero es habitual que haya varios títulos que podrían considerarse órgano convocante, aumentando en especificidad. Por ejemplo:

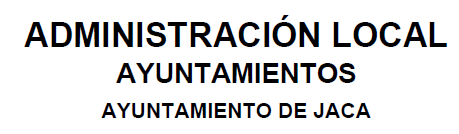


Ilustración 46. Detalle de órgano convocante en PDF de BOPH

* En el BOPZ es similar al BOPH, solo que también aparecen como títulos superiores la sección del artículo en el boletín y el número, por lo que el tercer título es el considerado órgano convocante, aunque, como en el otro, puede haber cuarto, quinto o hasta sexto título:

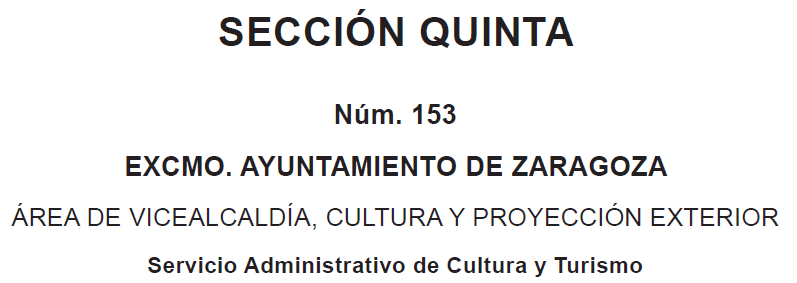


Ilustración 47. Detalle de órgano convocante en PDF de BOPZ

* En el BOPT, el órgano convocante es el título que aparece bajo el número del artículo.

Por otra parte, en los ficheros XML los órganos convocantes aparecen en etiquetas específicas, facilitando mucho su obtención:

* En el BOE, aparece en la etiqueta <departamento>, en la ruta documento/metadatos/departamento.
* En los cuatro boletines aragoneses, aparece en la etiqueta <emisor>, en la ruta documento/emisor.

El caso del HTML es una mezcla entre ambos: viene bastante estandarizado en cuanto a su posición respecto a etiquetas, pero no tiene una etiqueta identificativa.

* En el BOE, aparece como texto entre las etiquetas <dd> tras las etiquetas <dt> cuyo texto es Departamento.



Ilustración 48. Detalle de órgano convocante en HTML de BOE

* En los cuatro boletines aragoneses, aparece dentro del <h3> que tiene un <span> en el que se indica como texto “&Oacute;rgano emisor:”.

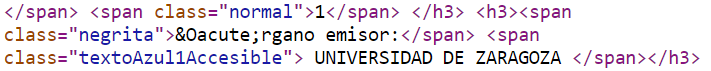


Ilustración 49. Detalle de órgano convocante en HTML de boletines aragoneses

### Datos de contacto

Mencionada su opcionalidad, se consideran datos de contacto las páginas web y los correos electrónicos que aparecen en la convocatoria. Se almacenará uno de los obtenidos. Para ambos casos, algunos de ellos aparecen con espacios en blanco en medio de la URL, lo que complica su extracción completa.

Como problema ocasional, la web a veces es cortada por un fin de línea, por lo que no se une bien con el inicio de la siguiente (ya que suele haber un espacio en blanco entre ellas). Esto no solo impide recuperar el enlace correctamente por nuestra parte: también parece que dificulta la escritura en el XML, ya que el mismo enlace aparece cortado, aunque por otro carácter:

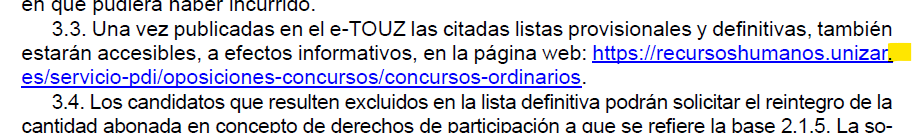


Ilustración 50. Detalle de página web incorrecta por blankspace intermedio

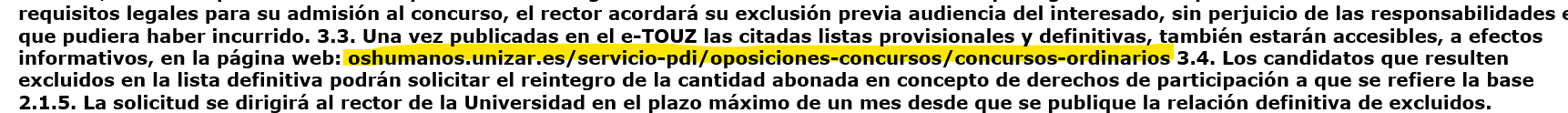


Ilustración 51. Detalle de página web incorrecta por corte en el XML

## Campos necesarios para las vistas

En esta sección los campos son los necesarios para poder mostrar las tres vistas mencionadas en el pliego (evitando reiterar los campos básicos).

### Estado de la convocatoria

Necesario para poder mostrar las convocatorias de empleo público abiertas mediante una vista como se indica en el punto 2.1.c.b del pliego, este campo indica el estado de la convocatoria, con valores principales “abierto” y “cerrado”.

El procedimiento habitual de apertura y cierre de convocatorias sería el siguiente:

1. En cuanto se detecte una nueva convocatoria, se incorpora a la base de datos con estado “abierto”.
2. Para cerrar la convocatoria:
   1. Obtención de artículo de nombramiento.
   2. En el mismo, detectar los campos que identifican una convocatoria (apartado 4.2.2).
   3. Con el identificador creado con estos campos, comprobar si existe una convocatoria abierta con dicho identificador.
   4. Cerrar la convocatoria identificada.
      1. En caso de existir más de una, cerrar solo una para minimizar posibles errores.

En la Ilustración 52 pueden verse dos artículos, el de apertura de una convocatoria y el de cierre de esta. En la Ilustración 53, lo mismo pero siendo una convocatoria realizada con una orden con identificador único.

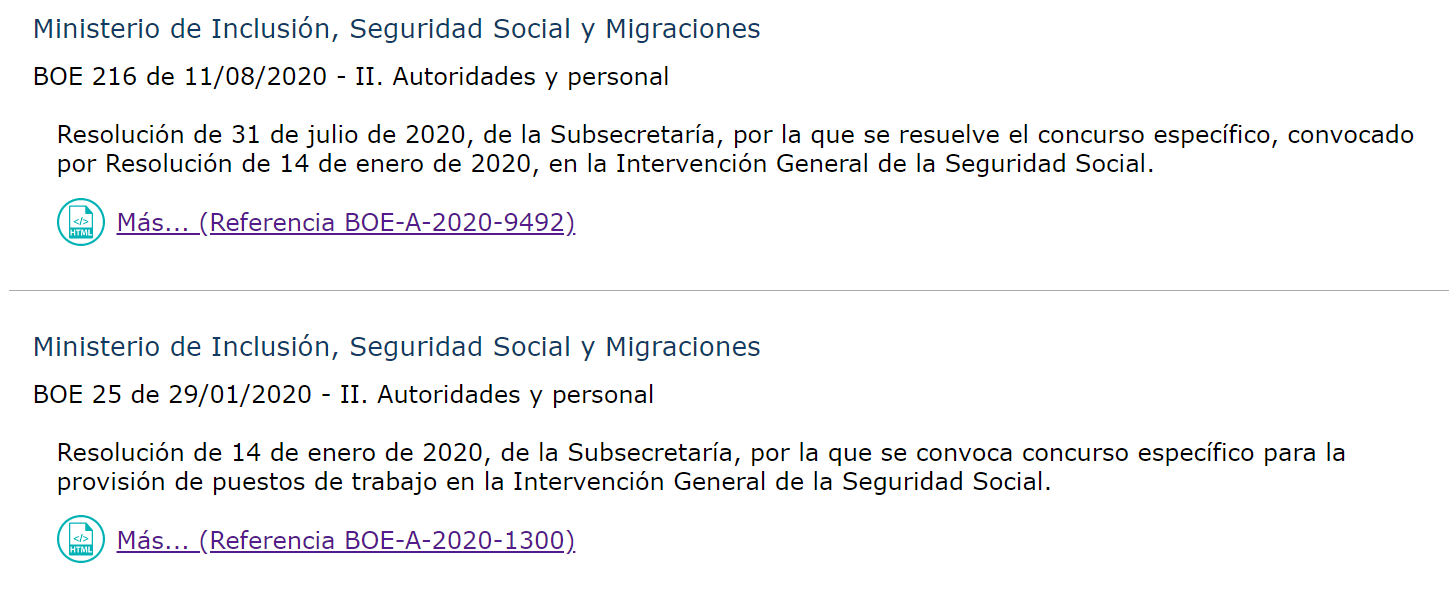


Ilustración 52. Ejemplo de artículos de apertura y cierre de convocatoria

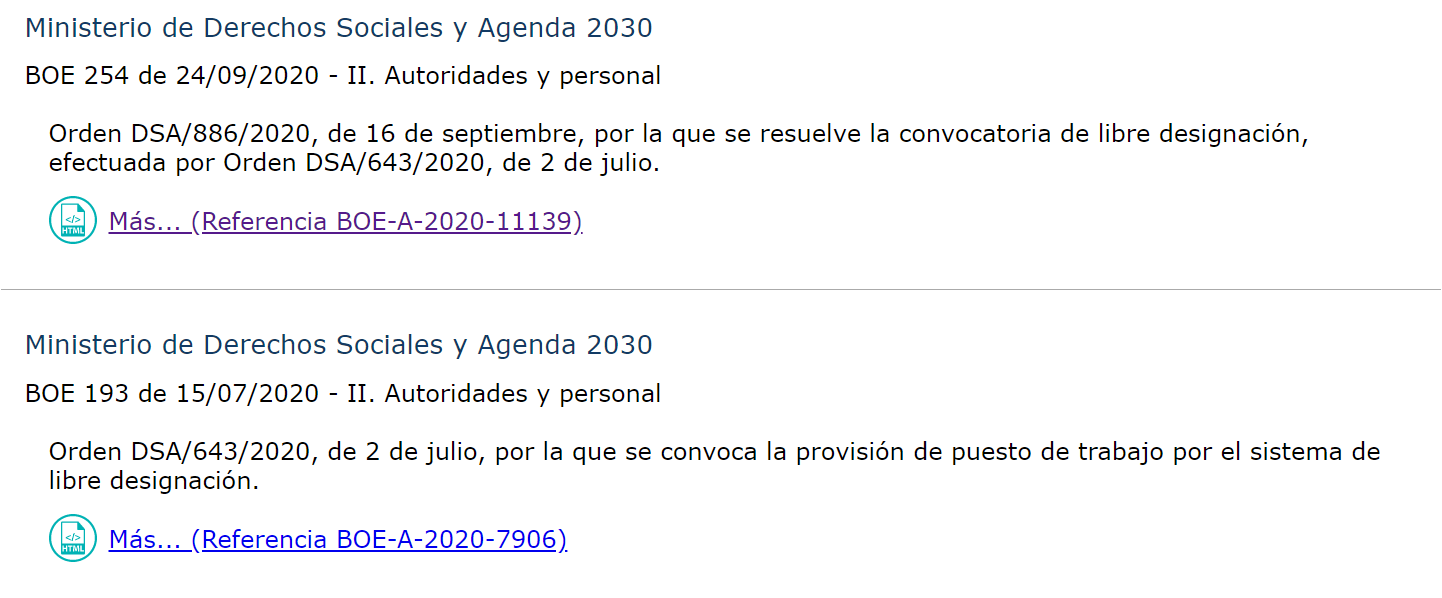


Ilustración 53. Ejemplo de artículos de apertura y cierre de convocatoria unívoca

Estudiando los estados de convocatorias se han encontrado, como problemas generales, que hay poca información sobre este estado (el que suele interesar es el estado de inscripción, es decir, si es posible todavía realizar solicitudes), y que la nomenclatura en cuanto a estado abierto/cerrado no está estandarizada, ya que eso se suele nombrar para el estado de inscripción mencionado.

#### Posibles cierres de convocatorias

En este apartado se van a enumerar los posibles cierres de convocatorias observados en los diferentes tipos de boletines, especialmente BOA y BOE.

* En ocasiones, se cierra mediante el nombramiento de los elegidos, exponiendo sus nombres.

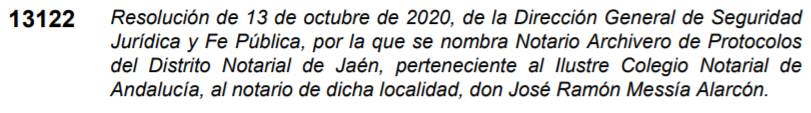


Ilustración 54. Nombramiento en BOE

* En otras ocasiones, se cierra mostrando una lista de admitidos en el texto o en un anexo en formato de tabla.

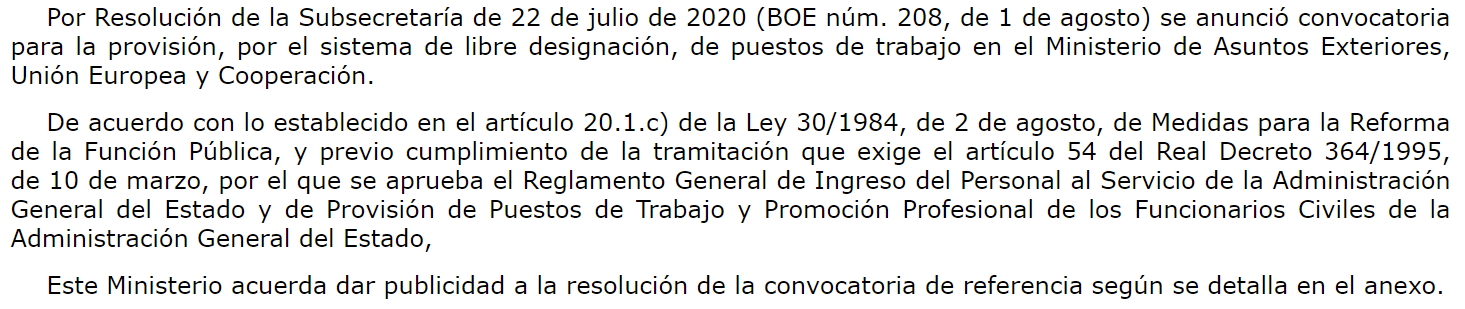


Ilustración 55. Referencia a anexo para lista de admitidos

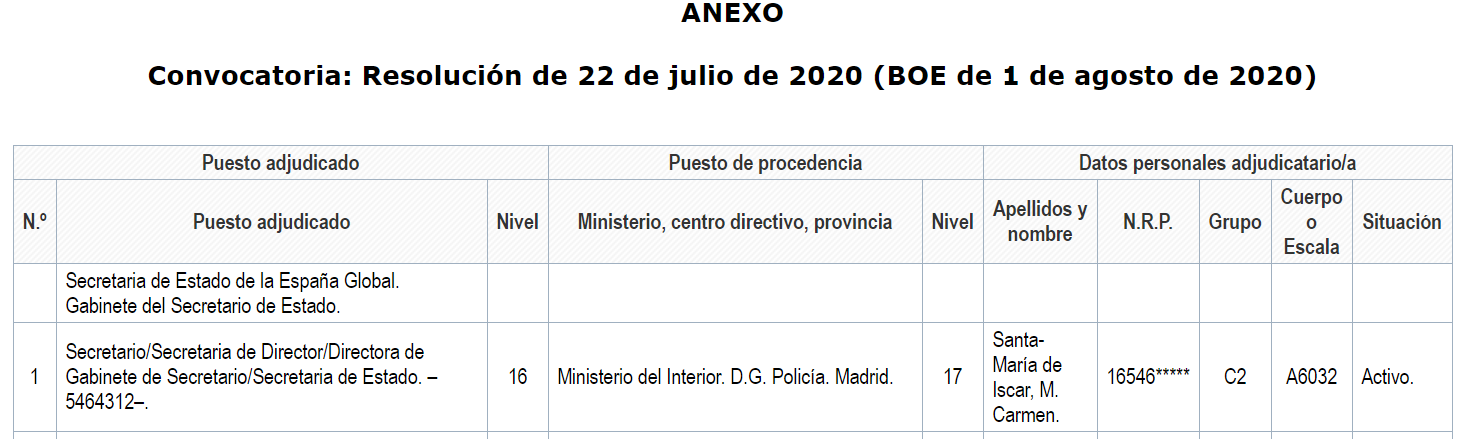


Ilustración 56. Lista de admitidos en anexo de BOE

* También puede ocurrir que se declaren plazas desiertas. En el BOE, de la misma forma que en la Ilustración 53, puede referenciar a tablas en anexos.

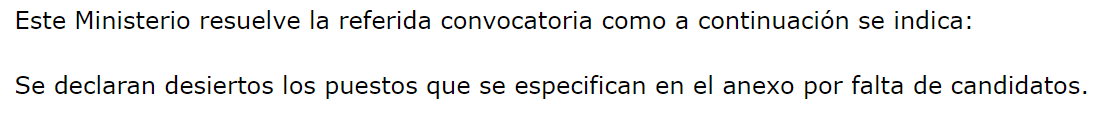


Ilustración 57. Referencia a anexo para plaza desierta

* Otro caso observado es dejar sin efecto una convocatoria. Esto podría incluso ocurrir en el artículo de apertura de otra convocatoria.

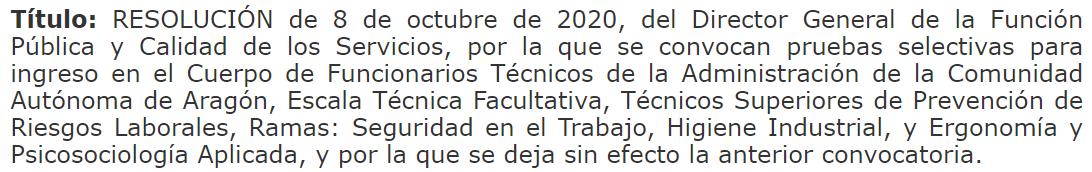


Ilustración 58. Convocatoria sin efecto y apertura de una nueva

#### Posibles problemas con los cierres de las convocatorias

A pesar de lo visto en el apartado anterior, en muchas ocasiones hay problemas con el cierre de las convocatorias:

* En los boletines provinciales a veces no se nombra de forma oficial al elegido, especialmente para ayuntamientos pequeños. En la Ilustración 59 puede verse la apertura de una convocatoria que, si se busca dicho ayuntamiento, todavía no se ha cerrado.

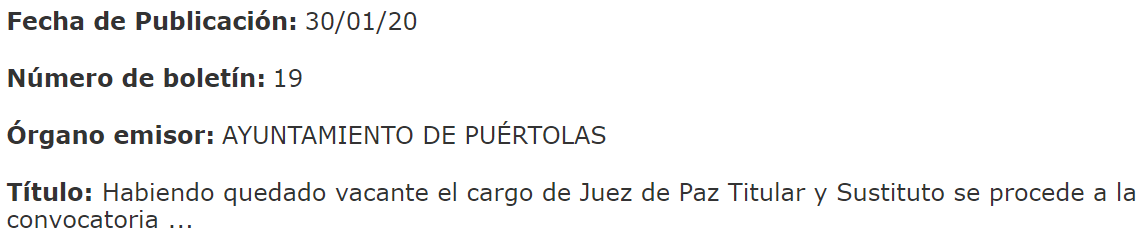


Ilustración 59. Apertura de convocatoria que todavía no se ha cerrado en el BOPH

* A veces se redirige a una página web para ver la lista de admitidos.

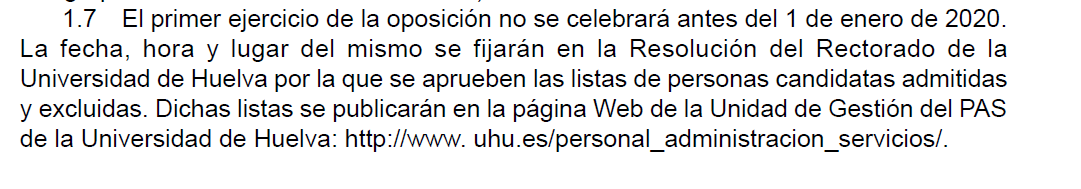
****

Ilustración 60. Redirección a página web para lista de admitidos

* A veces se redirige a un tablón de edictos para ver la lista de admitidos.

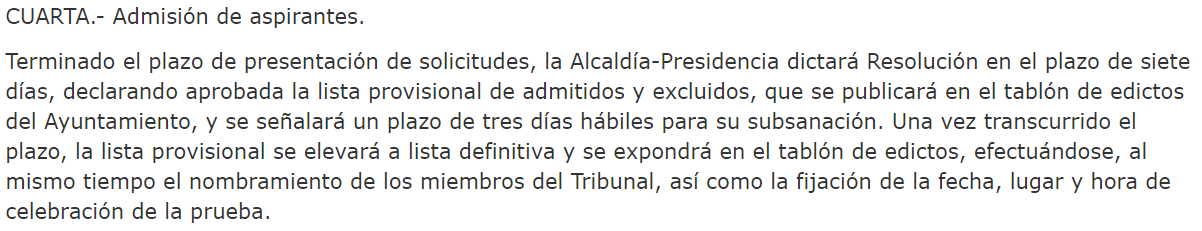


Ilustración 61. Redirección a tablón de edictos para lista de admitidos

* Dificultad de identificación unívoca de una convocatoria:
  + En ocasiones no se ve muy claramente qué convocatoria habría que cerrar porque hay pocas diferencias entre varias. Esto ha tratado de solucionarse con el identificador compuesto de la convocatoria.

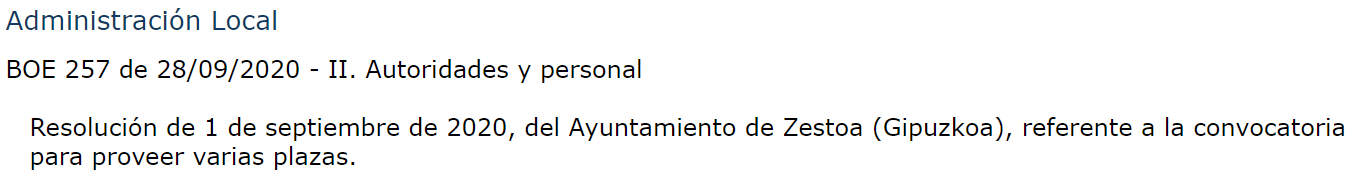
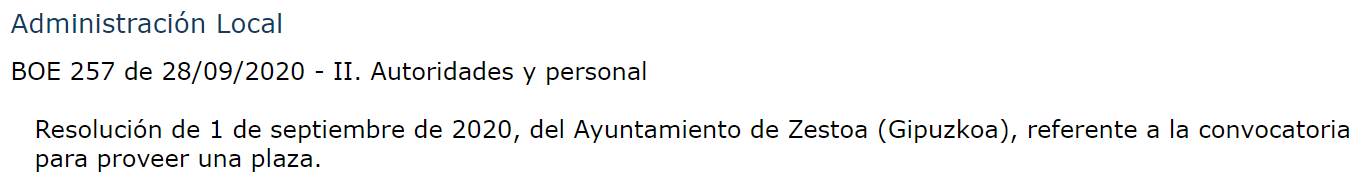


Ilustración 62. Apertura de convocatorias similares

* + En otras ocasiones, las diferencias son tan mínimas que un pequeño error en la extracción de los datos podría causar algún error.

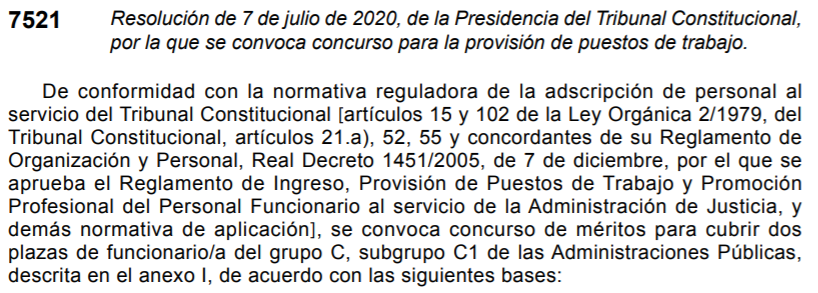
****

Ilustración 63. Apertura de convocatorias con diferencias mínimas I

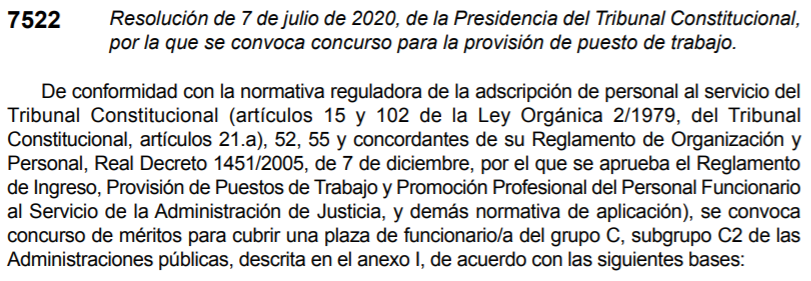
****

Ilustración 64. Apertura de convocatorias con diferencias mínimas II

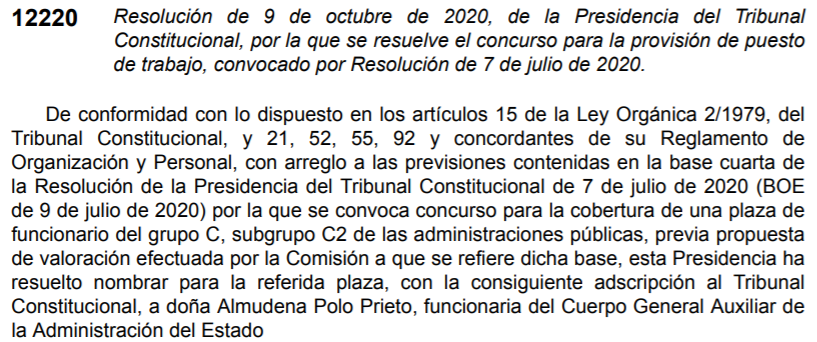
****

Ilustración 65. Cierre de convocatoria con diferencias mínimas

* + Además, en caso de haber escritos de manera diferente cualquiera de los campos que forman el identificador de la convocatoria en el cierre respecto a la apertura, sería necesaria la búsqueda difusa (*fuzzy search*) para poder identificar que se refieren a la misma convocatoria. Esto ocurre, de hecho en el caso que se acaba de ver: en la apertura (Ilustración 64) el puesto es “funcionario/a del grupo C, subgrupo C2 de las Administraciones públicas”, mientras que en el cierre (Ilustración 65) el puesto pasa a ser “funcionario del grupo C, subgrupo C2 de las administraciones públicas”. Las minúsculas de las dos últimas palabras no evitan la identificación, sino que es el “/a” de la primera palabra la que no se puede crear a partir del cierre. Además, si se aplicara dicha búsqueda difusa, también se cogería la apertura de la Ilustración 63 como posibilidad.

### Identificador de la convocatoria

Necesario para poder cerrar una convocatoria, este campo artificial es el que va a identificar de forma única a la convocatoria del presente artículo. Va a estar formado por el tipo de boletín del artículo, la fecha de firma de la apertura de la convocatoria, el órgano convocante y el puesto ofertado en la misma. Se han buscado diversos casos de convocatorias para tratar de asegurar que la conjunción de estos cuatro atributos es única, y hasta ahora no se ha visto ningún caso que indique lo contrario. Eliminando cualquiera de los cuatro atributos, en cambio, sí que dejaría de ser única esta identificación.

### Fecha de planteamiento/firma de la convocatoria

Necesario para poder identificar una convocatoria, este campo se refiere a la fecha en la que se acordó crear la convocatoria de la oferta de empleo.

En todos los artículos de BOA y BOE, el comienzo del título es del estilo “Resolución de 8 de octubre de 2020”, variando el tipo de artículo (si es orden, le acompaña su identificador) y la fecha, que se mantiene con esos “de” a modo de separadores. De esta forma, cuando el artículo sea uno que abre convocatoria, bastará con obtener dicha fecha de ese lugar en concreto. Buscarla en todo el texto sería contraproducente, ya que en ocasiones aparecen multitud de fechas, justificando las ofertas de empleo público o por otras causas.

Para el cierre de las convocatorias, en el caso del BOE aparece la misma cadena de caracteres pero en este caso al final del título, pudiendo venir precedido por distintos verbos pero de una forma bastante estandarizada al fin y al cabo. Por el contrario, el caso del BOA es más complejo, ya que se han observado ocasiones en las que:

* La fecha de firma era el comienzo del texto (Por Resolución de 1 de marzo de 2019 […] se convocó proceso selectivo”).
* La fecha de firma aparece por el texto (“la convocatoria […], publicada por Resolución de 16 de julio de 2020”).
* La fecha de firma aparece por el texto, sin ir precedida por su tipo (“la convocatoria de 21 de agosto de 2020”).
* La fecha aparece, por el texto, separada del tipo: (“Mediante Resolución de la Gerencia de […], de 19 de junio de 2020”).

Por otra parte, los boletines provinciales no cuentan con estas fechas.

### Enlace al artículo de cierre

Similar al campo del apartado 4.1.8, se guardaría la URL que lleva al sitio web del artículo que cierra la convocatoria presente.

## Otros campos mencionados

En esta sección se explican los campos que se han ido mencionando en distintas reuniones con el cliente y no están en el pliego, que van a tratar de incorporarse en la medida de lo posible.

### Tipo de la convocatoria

Este campo se refiere a si las plazas ofertadas en la convocatoria son para cualquier persona que cumpla los requisitos (lo que se va a llamar de ahora en adelante convocatoria libre), si son para personas que actualmente están trabajando para el estado, pero se quieren presentar para cambiar de empleo (que llamaremos convocatoria interna) o si se ofrecen plazas para ambas variaciones (lo que se va a llamar convocatoria libre+interna).

Suele venir indicado, en el BOA y especialmente en el BOE, en el título del artículo, con una variada nomenclatura. En ambos, cuando no viene indicado en el título, viene indicado en el texto, siendo nombrado en múltiples ocasiones (por ejemplo, cuando se detallan las distintas plazas, cuando se mencionan los requisitos, cuando se nombra el temario o se detallan los puestos, etc.). En los boletines provinciales en ocasiones no viene indicado este campo en todo el artículo, aunque presumiblemente estas convocatorias son libres. Por dicha razón, y dado que es menos restrictiva, por defecto se entiende que las convocatorias son libres.

Cuando se ofrecen plazas de ambos tipos, habitualmente se indica con frases predefinidas. Por ejemplo, la indicada a continuación, entre comas: “[…] por la que se convoca proceso selectivo para ingreso, por el sistema general de acceso libre y promoción interna, en el Cuerpo de Ingenieros de Caminos, Canales y Puertos del Estado.” Sin embargo, también aparecen los términos de forma separada por el texto en ocasiones, por lo que es necesario detectar correctamente ambos términos para no interpretar que la convocatoria es de un solo tipo.

Aparte de ese potencial problema, la mayor dificultad de este campo reside en la casuística de la terminología de estos tipos. Se ha realizado un seguimiento intenso de muchos artículos en busca de la nomenclatura de ambos tipos, enumerando en la siguiente tabla los diferentes términos con los que se distingue cada tipo. Esta terminología ha aumentado respecto a la mostrada en la reunión del 20 de octubre.

|  |  |  |
| --- | --- | --- |
| **Convocatoria Libre** | **Convocatoria Interna** | **Convocatoria Libre+Interna** |
| * Turno libre * Turno de acceso libre * Turno libre ordinario * Sistema de acceso libre * Sistema general de acceso libre * Sistema de concurso-oposición en turno libre * Concurso-oposición libre * Acceso libre * Pruebas selectivas para la estabilización de empleo temporal * Todo aquel interesado * Convocatoria abierta * La presente convocatoria tiene carácter de abierta * Concurso público | * Concurso * Concurso general * Concurso específico * Concurso de acceso * Promoción interna * Turno de promoción * Turno de promoción interna * Sistema de promoción interna * Sistema de libre designación * Sistema de provisión: libre designación * Libre designación * Procedimiento de libre designación * Concurso de méritos * Turno de movilidad * Concurso de movilidad interadministrativa * Concurso de traslados * Concurso de destinos * Movilidad voluntaria | * (Combinaciones entre ambas) * Por el sistema general de acceso libre y promoción interna * Mediante acceso libre o promoción interna * Por el turno de acceso libre y el sistema de promoción interna * Turno libre y promoción interna |

### Número de plazas ofertadas

Este campo fue planteado por Icíar en la reunión del 6 de octubre de 2020. Este campo indica el número de plazas totales que se ofrecen en la convocatoria. Es un campo muy variable, ya que la información puede venir muy diferente tanto en continente (cómo aparece la información) como contenido (a qué se refiere esa información).

En cuanto al contenido, se han detectado diferentes divisiones de estas plazas. En ocasiones se menciona explícitamente el número de plazas totales ofertadas, pero en algunas de estas ocasiones se muestran, además, divisiones en base a diferentes criterios: tipo de convocatoria (cuántas plazas son libres y cuántas internas), reservadas (cuántas para población general y cuántas se reservan para discapacitados, para mujeres víctimas de violencia de género, etc. Esta división en ocasiones aparece como subdivisión de la división por tipo de convocatoria, siendo las reservadas parte únicamente de las plazas libres y no de las totales), distribución de diferentes organismos, comunidades, Ofertas de Empleo Público (unas plazas corresponden a la Oferta de Empleo Público del 2019, que quedaron desiertas, y otras de la del 2020), etc. En otras ocasiones, solo aparecen estas divisiones, sin ofrecer el número total de plazas. El gran peligro de calcularlo en la limpieza de datos es que se podrían sumar cifras de otras divisiones, dando lugar a duplicados y, por ende, a información errónea.

En cuanto al continente, el número de plazas viene representado de varias formas y puede aparecer en bastantes lugares diferentes del texto del artículo. En el título, muchas veces se diferencia primeramente si se ofrece una plaza o varias plazas. En el primer caso ya no sería necesario buscar en el texto. Por el texto, el número de plazas puede venir dado de forma redactada (“Se convocan pruebas selectivas para cubrir, por el turno de promoción interna, nueve plazas […]”. A veces de forma numérica), de forma etiquetada (“Número de plazas: Una”. A veces de forma numérica), en tablas de los puestos ofertados en los anexos (de forma que son números de plazas parciales), etc.

En cuanto a las ocasiones en las que solo se oferta una plaza, además de lo comentado en el párrafo anterior, se pueden encontrar los textos “un puesto de Jefe/a […]” o “un puesto vacante de Jefe/a […]”.

En ocasiones el número de plazas ofertadas no aparece (ni de forma total ni de forma parcial), en otras se da por hecho que se ofrece una única plaza (sobre todo en los provinciales) y en otras incluso se redirige a un portal para buscarlas en fechas indicadas:

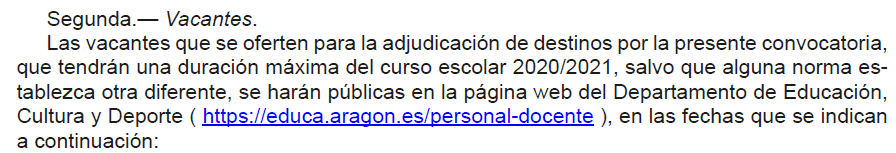


Ilustración 66. Detalle de no aparición de número de plazas

### Cuerpo del puesto ofertado

Este campo fue planteado por Icíar en la reunión del 6 de octubre de 2020. Este campo se refiere al Cuerpo o Agencia a la que pertenece el puesto ofertado en la convocatoria. En ocasiones, ni siquiera se menciona un puesto específico y se ofrece el ingreso al propio Cuerpo. Siempre viene precedido de la palabra Cuerpo (o Agencia), y esta parte va a formar parte del campo. Habitualmente viene con la palabra “de” tras el mismo. No aparece habitualmente pero cuando aparece suele ser en el propio título del artículo. La mayor complejidad que acarrea es que, en ocasiones, aparecen cuerpos en los requisitos a cumplir para solicitar una plaza, y estos cuerpos no tienen por qué coincidir con el ofertado. En estos casos, hay que tener cuidado en no obtener estos cuerpos, ya que no son el cuerpo buscado.

También se estudió el documento enviado por Icíar el 28 de septiembre, donde aparecen una serie de cuerpos y escalas de Aragón. No siempre encajan con los mencionados en dicho documento (cuando aparece un cuerpo, ya que habitualmente en estos no aparece), pero se han observado convocatorias en las que sí. No se ha encontrado un documento similar que refiera a cuerpos o escalas del Estado. En los boletines provinciales no se ha observado su aparición.

### Escala del puesto ofertado

Este campo fue planteado por Icíar en la reunión del 6 de octubre de 2020. Este campo se refiera a la Escala a la que pertenece el puesto ofertado en la convocatoria. Al igual que pasa con el cuerpo, en ocasiones no se menciona un puesto específico y sí la escala. Siempre viene precedido de la palabra Escala, y esta parte va a formar parte del campo. No aparece habitualmente pero cuando aparece suele ser en el propio título del artículo. En ocasiones viene acompañada de la subescala (un organismo más específico), explicada en el apartado 4.3.7.

En el BOA suele aparecer únicamente cuando aparece el cuerpo. En el BOE, sin embargo, aparece más que éste, especialmente en los títulos. También puede aparecer en algunos boletines provinciales.

Se han llegado a observar equivocaciones entre grupo y escala, apareciendo un “Escala A”.

### Grupo del puesto ofertado

Este campo fue planteado por Icíar en la reunión del 6 de octubre de 2020. Indica el grupo profesional al que se adhiere la convocatoria. También se consideran los subgrupos, ya que es más específico, pero conserva la identificación al propio grupo. A continuación, se muestran las titulaciones exigidas a cada grupo profesional, proporcionadas por Icíar el mismo 6 de octubre.

|  |  |  |  |
| --- | --- | --- | --- |
| Personal funcionario | | Personal laboral / Personal estatuario, militar, otro / Pruebas de capacitación profesional | |
| Grupos / | Titulaciones exigidas | Grupos | Titulaciones exigidas |
| subgrupo | Profesionales |
| A1 | Universitario de Grado. Doctor, Licenciado, Ingeniero, Arquitecto o equivalente | GP1 | Doctor, Licenciado, Ingeniero, Arquitecto o equivalente |
| A2 | Universitario de Grado. Diplomado Universitario, Ingeniero Técnico, Arquitecto Técnico o equivalentes. | GP2 | Diplomado Universitario, Ingeniero Técnico, Arquitecto Técnico o equivalentes. |
| B | Técnico Superior |  |  |
| C1 | Bachiller o Técnico. Bachillerato Unificado Polivalente o Formación Profesional de Técnico Superior o Técnico Especialista, o equivalente. | GP3 | Bachillerato, Bachillerato Unificado Polivalente o Formación Profesional de Técnico Superior o Técnico Especialista, o equivalente. |
| C2 | Graduado en Educación Secundaria Obligatoria. Educación General Básica o Formación Profesional de Técnico o Técnico Auxiliar, o equivalentes. | GP4 | Graduado en Educación Secundaria, Educación General Básica o Formación Profesional de Técnico o Técnico Auxiliar, o equivalentes. |
| E | Educación Primaria, Certificado de Escolaridad | GP5 | Nivel de formación equivalente a Educación Primaria, Certificado de Escolaridad o acreditación de los años cursados y de las calificaciones obtenidas en la Educación Secundaria Obligatoria. |
| F | Sin determinar en la convocatoria | F | Sin determinar en la convocatoria |

En cuanto a lo visto en el corpus de análisis, el grupo aparece en contadas ocasiones. A veces aparece en la parte de requisitos, pero no debería ser un problema ya que representa titulaciones requeridas. A veces el grupo e incluso el subgrupo forman parte del puesto. También hay que tener cuidado con peticiones de nivel de idiomas, que pueden coincidir con el grupo profesional requerido. A veces aparece con la forma “grupo/subgrupo C1”, otras indicado por separado (“grupo C, subgrupo C1”), a veces solo uno o el otro…

### Plazo de presentación de solicitudes

Este campo fue planteado por Icíar en la reunión del 20 de octubre de 2020. Es la cadena de caracteres literal que aparece en la convocatoria, que tras procesarla se convierte en las fechas de inicio y fin de presentación de solicitudes.

Para extraer y procesar más fácil y correctamente su información, se estandariza la frase, considerando plazo el trozo de frase que empieza en el número de días que dura y termina en la palabra ‘publicación’. De esta forma, la sintaxis es muy similar en la mayoría de casos:

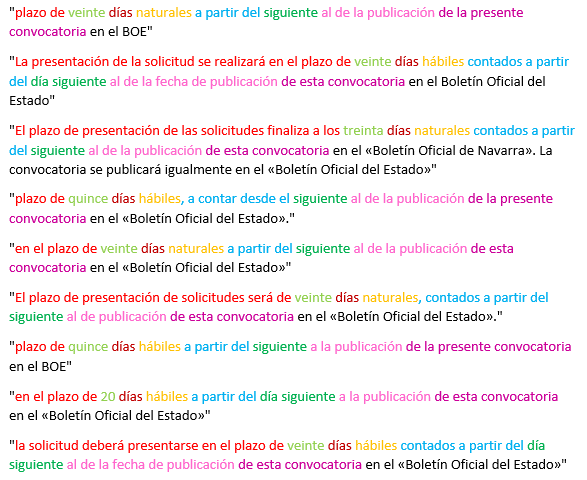


Ilustración 67. Sintaxis estudiada para el plazo de presentación de solicitudes

Como puede observarse, se indica primero el número de días que dura, luego indica si estos son hábiles o naturales, y después indica cuándo comienza el plazo (en el día de la publicación o en el siguiente). Cabe destacar que el número de días a veces aparece indicado en formato numérico, aunque es poco habitual.

En algunas ocasiones, en boletines provinciales, se han observado otras sintaxis de plazos, del estilo “el plazo es hasta el 20 de octubre.” o “el plazo es del 15.04 al 23.04.“.

### Subescala del puesto ofertado

Este campo ha sido planteado por Hiberus en la reunión del 3 de noviembre de 2020.

La subescala es un cuerpo más específico de la escala (explicada en el apartado 4.3.4), que en ocasiones aparece junto a ésta. En las ocasiones en las que aparece, lo hace precedida por la propia palabra “Subescala”, sin preposición posterior. Respecto a su posición con la escala, suele venir tras ella, bien tras una coma, bien entre paréntesis.

### Rango del artículo

Este campo ha sido planteado por Hiberus en la reunión del 3 de noviembre de 2020.

Este campo se refiere al rango del artículo en el que aparece la apertura de la convocatoria. El rango puede ser ley, decreto, decreto-ley, orden… pero en el caso de ofertas de empleo únicamente se han encontrado artículos con rango Resolución u Orden (este último es poco habitual).

En BOA y BOE siempre se encuentra al inicio del título, mientras que los boletines provinciales no poseen rango.

### ID de la orden del artículo

Este campo ha sido planteado por Hiberus en la reunión del 3 de noviembre de 2020.

Cuando el rango de un artículo es Orden, en ocasiones ésta viene acompañada de un identificador alfanumérico único de la misma, que sería lo que se guardaría en este campo. Como se ha comentado en el apartado 4.3.8, las ofertas de empleo en pocas ocasiones tienen rango de Orden, por lo que este campo no tendrá información en la mayoría de los casos.