

Extracción y búsqueda de información en boletines oficiales

Autor: Gómez García, Tomás

Directores: Fernández Gavilanes, Milagros y Fernández García, Norberto.

Contacto: tomasg@guardiacivil.es, formacion.tomas@gmail.com

Resumen:

El objetivo de este trabajo es la creación de una aplicación web en un servidor virtualizado desde la que se podrá llevar el control de los componentes de una Unidad militar para poder saber sus vicisitudes, tales como ascensos, condecoraciones, cursos, cambios de destino, etc.

Las Planas Mayores de las Unidades son las encargadas de llevar este control y comprobar si los componentes de la Unidad se mencionan en los distintos boletines, haciendo esta comprobación manualmente. Con la aplicación, lo que se pretende es descargarlas de este trabajo burocrático y optimizar mejor el tiempo de las mismas.

La aplicación se descargará los distintos boletines, buscando en ellos a los componentes de la Unidad (o a aquellas otras personas que por interés se puedan especificar) y palabras clave configuradas por los gestores.

La aplicación guardará las búsquedas realizadas, los boletines descargados y mandará correos electrónicos con las coincidencias del día. También permitirá realizar búsquedas históricas sobre los boletines previamente descargados.

La aplicación web tendrá varias secciones, desde las que se configurarán los usuarios, las personas o palabras clave a buscar, las coincidencias encontradas, un gestor de contenido y una sección de administración.

Por otro lado, habrá una serie de scripts (pequeños programas que ejecutan una determinada tarea) encargados de las descargas, conversión a texto, generación de índices, hacer búsquedas y mandar correo, que se ejecutarán automáticamente y que también son llamados por la aplicación web.

La aplicación completa ya ha sido terminada y testeada por lo que próximamente será puesta a disposición de los usuarios.

Palabras clave: búsqueda, BOGC, BOD, BOE, publicaciones

1. Introducción y objetivos

1.1. Motivación de este TFM

Todas las Unidades militares tienen una Plana Mayor (en adelante PIM), y entre sus objetivos está el poder sugerir enfoques o ayudar a tomar decisiones bien informadas que gestionen y conserven eficazmente los recursos de la unidad. Además de generar información, el personal también gestiona el flujo de comunicación dentro de la unidad y alrededor de ella. Han de conocer en todo momento el personal del que disponen y las incidencias sobre el mismo.

Teniendo en cuenta lo anterior, todas las PIM tienen entre sus cometidos el de estar pendientes de la publicación de boletines y de buscar en los mismos la aparición de sus componentes para poder saber sus vicisitudes, tales como ascensos, condecoraciones, cursos, cambios de destino, etc.

En la actualidad se trata de un trabajo manual, donde se busca, por ejemplo, el nombre de la Unidad en ese boletín y de ahí se va obteniendo información de los componentes.

Lo que pretende con este TFM es descargar a las PIM de todo este tedioso trabajo, aumentar el rango de aciertos y poder así mantener informando al Mando de todo lo que sucede en su Unidad con relación al personal.

1.2. Objetivos a conseguir

El objetivo con este TFM es crear un servidor virtualizado con una aplicación que se encargue de todo este trabajo, sea capaz de descargarse automáticamente los boletines de interés y buscar en ellos las menciones al personal de la Unidad y demás palabras claves definidas por cada uno de los usuarios de las PIM, Además, debe informar mediante un correo electrónico de las coincidencias encontradas y mantener las mismas en el servidor para posibles consultas posteriores.

Con el histórico de descargas se podrá también hacer búsquedas de una persona en concreto, para poder conocer su trayectoria, o también para poder buscar una palabra clave.

Por otro lado, desde la aplicación se podrá mandar hacer una descarga de boletines que no se hayan podido descargar automáticamente (por posibles fallos de comunicaciones, publicaciones que sufren retrasos, etc.).

Para la toma de requisitos se ha contado con la experiencia de una PIM, que tras varias reuniones ha sabido indicar lo que realmente necesitaba.

La implementación se ha hecho con un lenguaje de programación orientado a generación de páginas web (PHP) (1), una base de datos relacional y programas de conversión a texto de archivos con formato PDF.

En la validación y pruebas se ha contado con el apoyo de los integrantes de la PIM, que han comprobado todas las funcionalidades y el correcto funcionamiento de la aplicación.

2. Estado del arte

2.1. Antecedentes de búsquedas en Boletines

Hasta ahora, las PIM lo que hacen es buscar manualmente en los boletines a aquellos mandos de su Unidad más relevantes, o buscar el nombre de su Unidad en los boletines, pues cuando se produce

un cambio de destino, aparece el destino anterior que esa persona ocupaba y el nuevo destino. En algunas otras situaciones, como ascensos, también sucede lo mismo.

Valiéndose de estas búsquedas manuales es como pueden saber las vicisitudes del personal de su Unidad.

3. Análisis de la aplicación

3.1. Entrevistas con la PIM

Para que la aplicación llegase a un buen fin se hicieron reuniones con el personal encargado de llevar a cabo esta tarea, a los que se les explicó lo que se podría hacer y se les planteó que propusiesen funcionalidades a implementar en esta aplicación.

La primera cosa que se tuvo en cuenta fue el acceso. No se podía dejar un acceso libre a esta información, por lo que se debía crear una gestión de usuarios y permitir el acceso identificando al usuario de forma fehaciente. Se descartó el acceso con usuario/contraseña y se optó por un acceso que exigiese un certificado digital, en particular el certificado de la tarjeta de identificación profesional de cada uno.

El siguiente problema que se identificó era tener un listado actualizado del personal que estaba a su cargo y cómo poder actualizarlo. Se eligieron dos estrategias: una era la carga automática mensual (ya que la base de datos del personal se encuentra accesible en una base de datos central, a la que se accede solo con autorización y que solo se actualiza mensualmente) y la otra, complementaria a la anterior, era la carga manual de datos por los componentes de la PIM. Lo que se haría sería subir un archivo CSV y actualizar la base de datos de la aplicación con esos datos.

También se expresó la necesidad de poder realizar búsquedas automáticas sobre determinadas palabras que apareciesen en alguno de los boletines o en un boletín concreto, y estas búsquedas serían distintas dependiendo de la actividad de la que cada persona de la PIM estuviese encargada. De esta forma se añadió una nueva búsqueda que se ejecutaría diariamente y dependería de cada usuario de la aplicación.

3.2. Búsquedas mediante línea de comandos

Se sopesaron varias estrategias para realizar las búsquedas dentro de los boletines, la estrategia finalmente adoptada fue la de búsquedas mediante línea de comandos.

Los comandos de Linux están muy optimizados para realizar búsquedas dentro de textos, incluso con archivos grandes, de varios centenares de megabytes, realizar un *grep* (comando para buscar una cadena dentro de un texto) sobre el mismo es casi inmediato.

Esta velocidad de búsqueda puede ser muy útil para encontrar palabras dentro de grandes volúmenes de texto sin necesidad de ningún tipo de indexación.

Una posible estrategia de búsqueda consistiría en pasar los archivos PDF de un boletín a `txt` y a continuación juntar todos los `txt` en un único archivo. El contenido de cada línea de este archivo sería la ruta del archivo origen y a continuación, en la misma línea (eliminando los saltos de línea) el contenido de cada uno de los archivos `txt`.

4. Diseño de la aplicación

4.1. Elección del software

A continuación, detallaremos el software elegido para la creación de la aplicación y explicaremos el porqué de su elección.

Después de haber sopesado varios virtualizadores se vio que VirtualBox (2) tiene muchos problemas con las unidades extraíbles y sobre todo con las interfaces de red. Por tanto, dado que lo que se pretende es que esta aplicación pueda funcionar sobre cualquier equipo o incluirla en un hipervisor, al final se ha optado por usar VMWare Player (3).

Como sistema operativo se optó por Linux ya que cada distribución suele incluir todo el software adaptado ya a cada versión.

Por otro lado, los usuarios no necesitan interactuar con el servidor, si no con la aplicación, por lo que la elección del sistema operativo Linux no supone un problema.

Dentro de Linux se ha elegido Debian (4) (la versión 12, *bookworm* (5), que en el momento de escribir este texto es la última versión de esta distribución), por considerar que es más estable que el resto de distribuciones de Linux.

Como servidor web se ha elegido Apache (6), la versión que viene en la distribución de Debian (versión 2.4.62), ya que Apache es el más usado de los servidores web y sobre el que mejor documentación hay.

Como sistema gestor de bases de datos relacionales (SGBD) se ha elegido MariaDB (7), que también viene por defecto con Debian (versión 15.1).

El lenguaje de programación elegido es PHP (versión 8), tanto para el *frontend* como para el *backend*, ya que es uno de los más utilizados y tiene un amplio soporte y si es necesario un mantenimiento de la aplicación será más fácil hacerlo por su baja curva de aprendizaje.

Para el modelado de todas las funcionalidades se ha elegido el modelador Bizagi (8), ya que posee una interfaz intuitiva, que facilita su uso, además de utilizar el estándar *Business Process Model and Notation* (BPMN) (9) y disponer de una amplia comunidad y documentación de apoyo. No se han usado todas sus funcionalidades, solo las de modelado.

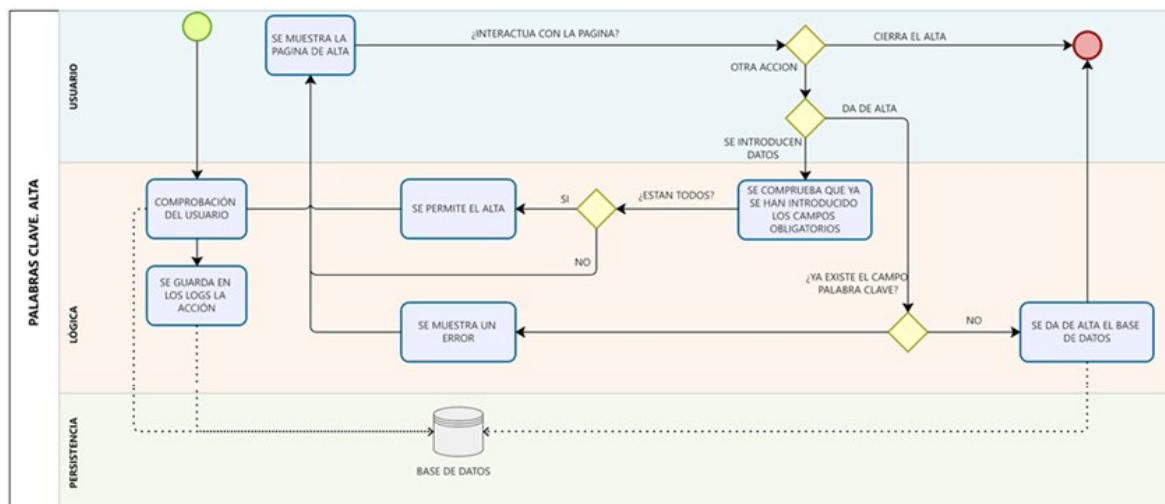


Figura 1. Ejemplo de modelado de una funcionalidad.

4.2. Componentes de la aplicación

A continuación, se explica el uso de los distintos componentes empleados en la aplicación.

Para las hojas de estilo se buscó en Internet un CSS (*Cascading Style Sheet*) que tuviese propiedades *responsive* (adaptable, significa hacer que un sitio web se pueda visualizar en todos los dispositivos: tabletas, smartphones, etc.) se usó para ello la plantilla denominada *editorial* (10) proporcionada por htmlup.net (11).

Para simplificar las operaciones de JavaScript se ha usado jQuery (12), no ya solo por las funcionalidades que provee, sino porque otras funcionalidades que se han ido añadiendo a la aplicación dependen de él.

Para la carga dinámica del contenido se ha usado una librería de PHP denominada Jaxon (13) que permite el uso de AJAX con PHP. Esta librería permite crear funciones en PHP, que pueden ser llamadas como si fuesen funciones de JavaScript y que se ejecutan en el *backend*.

También se han usado otras funcionalidades para hacer más visual la aplicación, como por ejemplo una extensión de jQuery denominada *draggable* (14), para la creación de las modales (ventana o cuadro de diálogo que se superpone al contenido principal de una página o aplicación) y permitir su movimiento. Otra funcionalidad que se ha incluido es *dmUploader* (15) un JavaScript basado en jQuery y que usa AJAX para la subida de archivos.

5. Implementación

5.1. Estructura de directorios

El directorio raíz (llamado *DocumentRoot*) de Apache es el lugar a partir de donde se hace visible el servidor web y en nuestro caso es el *path*: `/var/www/htdocs/BO`. Es decir, nuestra aplicación sería llamada como <https://127.0.0.1>, y ya nos aparecería la página web generada por el *index.php*.

5.2. Scripts

Existe una serie de *scripts* que se encuentran en `/var/www/htdocs/scripts`, fuera del acceso desde el *frontend* para que no se pueda tener acceso directo a ellos y ser ejecutados desde él, pero accesibles por las funciones de *backend*,

5.3. Páginas web

Las páginas se encuentran en el *DocumentRoot* `/var/www/htdocs/BO`, donde está el *index*, los ejecutables de PHP y jQuery, hojas de estilo CSS, imágenes (img), etc.

5.4. Navegación por los Boletines

En el directorio `/var/www/htdocs/BO/Boletines` se encuentran los boletines descargados y sus archivos en `txt`, accesibles desde el navegador, por si algún gestor quiere acceder a ellos.

5.5. Acceso a la aplicación

Solo se accede en modo HTTPS (el puerto 80 se ha deshabilitado). Para acceder en modo HTTPS al servidor web, se usan los certificados de la tarjeta de identificación profesional y también se ha generado una CA autofirmada.

Una vez que el usuario ha accedido a la aplicación entran en juego los tres posibles roles que se han creado para la misma:

- 1) Superadministrador: Puede ver a todos los usuarios y personas buscadas, así como modificar todos los campos, también gestiona ciertos parámetros de la aplicación como el acceso a bases de datos.
- 2) Administrador: Solo puede ver y modificar a los usuarios de su grupo y personas buscadas, así como subir archivos.
- 3) Gestor: Solo tiene visibilidad de los usuarios y personas buscadas, pero no puede modificarlos. Sí puede dar de alta y modificar palabras clave.

En esta aplicación se están tratando datos de carácter personal y la aplicación accede a distintas bases de datos y relaciones del personal de la Guardia Civil. Los usuarios que accedan a esta aplicación han de tener los mismos permisos y la misma autorización necesaria como si accediesen a los datos originales.

6. Estructura de la aplicación

La aplicación se encuentra dividida en secciones, que son:

- 1) La cabecera, donde aparece el usuario y su identificador, que coincidirá con las credenciales del certificado utilizado para acceder a la aplicación. También aparecen las distintas secciones de las que consta la aplicación. Esta barra de secciones estará siempre visible, para poder navegar por ellas.
- 2) Gestores: En esta sección se muestran los gestores que pueden acceder a la aplicación.
- 3) Correos: Se muestra a quién se mandaría un correo con las coincidencias de los distintos boletines.
- 4) Buscados: Se muestran a todas aquellas personas que se buscan en los distintos boletines.
- 5) Palabras clave: En esta sección cada gestor podrá definir aquellas palabras que son de su interés para que se puedan buscar automáticamente de forma diaria en los boletines que se seleccione.
- 6) Coincidencias: Se muestran las coincidencias de buscar a todo el personal marcado para la búsqueda en cada uno de los boletines.

PALABRAS CLAVE EN BOG, BOB, BOE y OTRO

2024

Agosto - 2024

Lu	Ma	Mi	Ju	Vi	Sa	Do
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Resultado de buscar JEFATURA DE ASUNTOS ECONOMICOS DE LA GUARDIA CIVIL en el día 10/08/2024

BOE JEFATURA DE ASUNTOS ECONOMICOS DE LA GUARDIA CIVIL en Notificaciones/BOE-B-2024-28795.pdf	Ver en bdt
BOE JEFATURA DE ASUNTOS ECONOMICOS DE LA GUARDIA CIVIL en Notificaciones/BOE-B-2024-28797.pdf	Ver en bdt

Resultado de buscar CIBER en el día 10/08/2024

BOE CIBER en Notificaciones/BOE-A-2024-16672.pdf	Ver en bdt
BOE CIBER en Disposiciones/BOE-A-2024-16672.pdf	Ver en bdt

BUSQUEDAS DE **Ver las que NO se buscan** **Incluir**

Buscar por los criterios ...

Texto	BOG	BOB	BOE
Análisis, análisis, análisis	SI	SI	SI
BASICO DE CIBERDEFENSA	SI	SI	NO
CIBER	SI	SI	SI
CURSO DE SEGURIDAD DE TECNOLOGIAS DE LA INFORMACION Y COMUNICACION	NO	SI	NO
DIPLOMA DE INFORMATICA MILITAR	SI	SI	NO
DIRECCION STIC	SI	SI	NO
DTIC	SI	NO	NO
EFT	SI	SI	NO
FACULTATIVA TECNICA	SI	SI	NO
INGRESO EN LA ORDEN AL MÉRITO POLICIAL	SI	NO	NO

1 2 SIGUIENTE

Figura 2. Captura de pantalla de una sección de la aplicación.

- 7) Contenido: Se muestran tres tablas que sirven de soporte para los desplegables de las secciones de “Gestores” y “Buscados”.
- 8) Cargas: Sección en la que los gestores pueden realizar cargas locales, desde un fichero CSV.
- 9) Significado del email: sección informativa, en la que se explica el significado de lo que aparece entre paréntesis en el correo electrónico.
- 10) Administración: Aquí se encuentran los parámetros que necesitan algunos *scripts* para funcionar, la gestión de las secciones y la información del sistema.

7. Validación y Pruebas

Una vez creada la aplicación se sometió a una serie de pruebas para comprobar su correcto funcionamiento. Como resultado de algunas de estas pruebas se encontraron una serie de errores que fueron subsanados y probados nuevamente para asegurar su correcto funcionamiento.

Se hicieron pruebas de:

- Todos los *scripts* de descarga y de búsqueda. Comprobando que según qué parámetros se pasasen hacían correctamente su función.
- Los dos *scripts* de subida, comprobando con distintos archivos y distintos tamaños su correcto funcionamiento.
- Todas y cada una de las secciones, su visibilidad, acceso, altas, bajas y modificaciones.

8. Conclusiones y líneas futuras

8.1. Conclusiones

La aplicación, tal como ha sido expuesta en este TFM, está ya terminada, y lista para ser puesta en producción, por lo que se han alcanzado todos los objetivos que se pretendían conseguir.

La aplicación ya ha sido presentada a los usuarios de la PIM de mi Unidad y próximamente se podrá utilizar.

8.2. Líneas futuras

Aunque se ha llegado a terminar la aplicación con todas las funcionalidades inicialmente previstas, se podrían añadir alguna serie de mejoras que, por tiempo de desarrollo, no se han podido implementar, tales como:

- 1) Sección de sugerencias. Para que los usuarios de la aplicación planteen posibles mejoras, nuevas funcionalidades o avisen de errores que hayan pasado desapercibidos.
- 2) Sección de registros. Todas las acciones que se llevan a cabo en la aplicación y errores producidos quedan registradas en la base de datos, por lo que se podría ver lo que ha ocurrido en un momento determinado, o por un usuario concreto, pero no se ha creado un buscador o visualizador de los mismos. Posiblemente sea uno de los siguientes pasos para una mejora de la aplicación.
- 3) Sección con estadísticas. A través de los registros también se podrían crear unas estadísticas de uso de la aplicación, secciones más usadas, número de coincidencias diarias, etc.
- 4) Las coincidencias encontradas se guardan en una tabla específica (“*COINCIDENCIAS*”) en la base de datos. Esta tabla podría crecer mucho por lo que sería interesante ir guardando un histórico o creando tablas nuevas, como, por ejemplo, “*COINCIDENCIAS _2023*”, “*COINCIDENCIAS _2024*”, etc.

- 5) Existen excepciones sobre las búsquedas, de tal forma que el nombre de una persona que venga después de “el general”, “el teniente general”, “jefe interino de”, no se incluyen como coincidencias, ya que normalmente son los que firman cada orden o curso. Una mejora sería que estas excepciones se encontrasen en alguna tabla de gestión por si es necesario añadir más excepciones.
- 6) También existe la posibilidad de añadir más boletines o documentos en los que buscar, como por ejemplo algún boletín de la Comunidad Autónoma en la que se encuentre la Unidad que usa la aplicación.

Referencias

1. PHP. [Internet]. [3 Sep 2024]. <https://www.php.org>.
2. VirtualBox. [Internet]. [5 Sep 2024]. <https://www.virtualbox.org/>.
3. VMWare. [Internet]. [3 Sep 2024]. <https://www.vmware.com/>.
4. Debian. [Internet]. [13 Sep 2024]. <https://www.debian.org/>.
5. Debian. [Internet]. [1 Sep 2024]. <https://www.debian.org/distrib/>.
6. Apache. [Internet]. [14 Oct 2024]. <https://httpd.apache.org/>.
7. MariaDB. [Internet]. [4 Sep 2024]. <https://mariadb.org/>.
8. Bizagi. [Internet]. [5 Nov 2024]. <https://www.bizagi.com/es>.
9. BPMN. [Internet]. [24 Oct 2024]. <https://www.bpmn.org/>.
10. HTML5UP. [Internet]. [3 Sep 2024]. <https://html5up.net/editorial>.
11. HTML5UP. [Internet].: Editorial Genios [2 Sep 2024]. <https://htmlup.net>.
12. jQuery. [Internet]. [23 Nov 2024]. <https://code.jquery.com/>.
13. Jaxon-php. [Internet]. [5 Sep 2024]. <https://github.com/jaxon-php/jaxon-core>.
14. jQuery. widgets. [Internet]. [21 Oct 2024]. <https://jqueryui.com/download/>.
15. DanielMG. dmUploader. [Internet]. [26 Sep 2024]. <https://github.com/danielm/uploader>.