

Aplicació per una base de dades de documents

Lluís Poves¹, Antoni Barba², Xavier Pouplana³
Dept.. Enginyeria Telemàtica Universitat Politècnica de Catalunya
Jordi Girona 1-3, Mod. C3, 08034, Barcelona
Dept.. Projectes Arquitectònics (UPC)
CDPAC - ETSAB

Abstract

El objetivo de este proyecto es diseñar un sistema de base de datos relacional destinada al almacenamiento de documentos de arquitectura y desarrollar una aplicación web que permita acceder a esos documentos a través de un navegador, con funciones de consulta y catalogación de documentos. La aplicación ha de ser totalmente abierta, flexible y ampliable.

Los recursos necesarios básicos son un servidor de base de datos y un servidor web. Se utilizará un sistema operativo Linux, un servidor web Apache y una base de datos MySQL. Para el desarrollo de la aplicación se utiliza JavaScript en la parte del cliente y PHP en la parte del servidor. PHP es un lenguaje interpretado rápido, multiplataforma y dispone de numerosas funciones de acceso a bases de datos MySQL, además de muchas otras librerías que facilitan el desarrollo de aplicaciones web.

1 Introduccion

La base de datos estará formada por diferentes tablas que contendrán toda la información de arquitectura: documentos, fondos archivísticos, autores, entidades, eventos y referencias bibliográficas. Además se utilizará la misma base de datos para almacenar otro tipo de información como los usuarios registrados y las ayudas de la aplicación. La aplicación tendrá funciones de consulta y de catalogación. Dispondrá de una zona privada para usuarios catalogadores y una zona de administración, con funciones para el mantenimiento de la aplicación, la base de datos, los usuarios registrados y las copias de seguridad. Soportará también el tratamiento de imágenes asociadas a los registros. Otro aspecto fundamental de la aplicación será la seguridad, diseñada a tres niveles: servidor, base de datos y aplicación

Las consultas serán de diferentes tipos: simple, por campos, por listas, por tipologías y experta. Todas las páginas de consulta tendrán elementos como títulos de página, nombre de la base de datos consultada, elección del formato de presentación del resultado, información sobre la consulta, instrucciones y navegación con otras páginas. Los resultados de las búsquedas se presentarán en formato simple, completo o detallado, con navegación sobre los mismos y enlaces directos a la información relacionada.

En la parte de catalogación, los usuarios registrados tendrán opciones de inserción, edición y borrado de registros, a través de formularios. También dispondrán de otras funciones específicas como el envío de imágenes, consultas especiales y acceso a documentos y manuales de catalogación.

En su versión inicial el catálogo estará formado por 6 fondos archivísticos: Gaudí (Cátedra i AMAB), HOLSA, Moragas, Sostres i Arxiu Gràfic ETSAB. Será ampliable a cualquier número de fondos y permitirá la catalogación de fondos privados con finalidad docente.

¹ ETSETB

² ETSETB

³ ETSAB

El sistema de archivo de documentos de arquitectura del CDPAC ha estado basado hasta ahora en CDS/ISIS, sistema de base de datos de libre distribución desarrollado por la UNESCO. La puesta en marcha de este sistema fue responsabilidad de la empresa italiana Nexus, que además proporciona herramientas propias de catalogación y consulta basadas en este sistema. Concretamente en el CDPAC se utilizan dos de estas herramientas: Vim4 para la catalogación y EasyWeb para la consulta online. Vim4 es una aplicación prácticamente obsoleta, diseñada para MS-DOS y que funciona solo en modo texto. Es necesario trabajar en local, no permitiendo introducir datos a través de la red.

EasyWeb es el software que hace de intermediario entre la base de datos y el servidor web. Está instalado en un servidor Linux junto con el servidor web Apache. La aplicación está basada en un conjunto de CGIs que acceden a los archivos de la base de datos ISIS. Nexus proporciona otra herramienta: EasyCat, que permite la catalogación online, pero no está implantada en el CDPAC.

Estos productos son aplicaciones comerciales y lógicamente no son gratuitos. Los inconvenientes de este sistema son básicamente su coste económico, pues una actualización implica una inversión y la falta de flexibilidad a la hora de realizar modificaciones o mejoras en la base de datos o en la interfaz web de consulta.

Objetivos

El objetivo de este proyecto es el de migrar este sistema hacia una solución que evite los inconvenientes descritos en la introducción. Se trata de diseñar un sistema de base de datos relacional y desarrollar una aplicación web que permita acceder a esta información a través de un navegador, mediante la utilización de consultas de diferentes tipos y actualizar los datos online.

La utilización de un sistema de base de datos para almacenar documentación de arquitectura tiene varias ventajas:

- Proporciona un sistema de archivo de la información sólido, seguro y flexible.
- Permite el acceso digital a la información, a través de las redes telemáticas. Este acceso online tendrá será utilizado tanto para operaciones de consulta como para actualizar la información.
- La publicación en la red además, permite preservar los documentos originales de manipulaciones innecesarias.

La aplicación web a desarrollar ha de permitir el acceso a la base de datos a diferentes niveles mediante la autenticación de los usuarios que se conecten. Básicamente existirán tres niveles de acceso: consulta, catalogación y administración.

La aplicación será totalmente flexible y ampliable, pues solo será necesario modificar las páginas o crear nuevas por parte del administrador para realizar mejoras o crear nuevas funcionalidades, sin depender de terceros que impliquen un coste económico.

2 Herramientas y recursos

Para el desarrollo de esta aplicación y su posterior utilización es necesaria una infraestructura hardware y software específica. En este momento se dispone de un PC estándar con Linux RedHat 7. Las herramientas necesarias son un servidor web, un lenguaje de programación y un servidor de bases de datos relacional. RedHat incluye todas estas herramientas. No obstante, la aplicación ha de poder ejecutarse tanto en entornos Unix como Windows. La aplicación será

portable a diferentes entornos si el lenguaje de programación y la base de datos son multiplataforma.

El servidor web que utilizaremos es Apache. Se trata del servidor web más popular, pues más del 50% de los sitios de Internet lo utilizan. Utilizaremos la versión para Linux, aunque existen distribuciones para entornos Unix e incluso para Windows. Apache es un software de libre distribución.

Como lenguaje de programación utilizaremos PHP. Es un lenguaje interpretado que se ejecuta en el servidor. El código de los programas va incrustado en las propias páginas HTML. Su sintaxis es cómoda y similar a la de otros lenguajes de programación, como C, C++ o Java, incluida la programación orientada a objetos. PHP es rápido, multiplataforma y dispone de una gran cantidad de librerías que facilitan el desarrollo de aplicaciones. Permite crear páginas de forma dinámica, enviando al cliente solo código HTML resultado del procesamiento de los scripts, ocultando el código. Es una herramienta muy moderna que está teniendo mucha popularidad en el ámbito de las aplicaciones web. La ventaja de su utilización con Apache es que el intérprete PHP se compila como un módulo de Apache, por lo que la velocidad de ejecución de las páginas es muy superior, así como el consumo de recursos al cargarse el intérprete de PHP una sola vez en memoria.

Como servidor de base de datos disponemos de MySQL. Se trata de un servidor SQL (Structured Query Lenguaje), el estándar más común en bases de datos. Es uno de los servidores de bases de datos “Open Source” más extendidos. MySQL es muy rápido, fiable y sencillo. Su conectividad, velocidad y seguridad lo convierten en un buen servidor para el acceso a datos a través de Internet.

Apache, PHP y MySQL forman un equipo totalmente integrado, rápido y robusto para la creación de sitios web con contenidos dinámicos.

2. 1 Requerimientos de la aplicación

En este apartado se describen los diferentes componentes y las especificaciones que ha de cumplir el sistema de consulta online (Internet) del catálogo de documentos de Arquitectura del CDPAC.

Los componentes principales de este sistema serán:

- Páginas de consulta
- Páginas de lista
- Páginas de resultados
- Páginas de catalogación
- Página de administración
- Otras páginas (ayuda, descripción base de datos, etc)

Páginas de consulta

En primer lugar dividiremos las consultas en dos tipos: simple y avanzada.

La **consulta simple** realizará la búsqueda de unos o más términos en toda la base de datos en sus principales campos (autor, título, proyecto, lugar, etc).

Esta consulta solo requerirá del usuario la introducción del término de búsqueda sin ninguna opción adicional.

La **consulta avanzada** permitirá al usuario una búsqueda más compleja y con mayor control de los resultados, con diferentes opciones. Dividiremos la consulta avanzada en tres opciones: listas, campos y experta.

Tipos de consulta avanzada

Listas

Búsqueda por listas de autor, título, materia, etc. Tras escoger la palabra clave se presenta una lista alfabética de elementos en los que se incluye el término buscado.

Campos

Se presenta una lista de campos en los que el usuario introducirá el contenido de uno o varios. El resultado presentará una lista de títulos correspondientes a las condiciones impuestas.

Experta

La implementación de esta opción en la nueva versión está por decidir. Básicamente se diferencia de la búsqueda por campos en que debería permitir la utilización de operadores booleanos AND, OR, NOT y otros operadores de similitud, truncamiento, adyacencia, etc.

Elementos de las páginas de consulta

A continuación se describen los elementos que formarán parte de las diferentes páginas de consulta.

Los siguientes elementos serán comunes a todos los tipos de página de consulta:

Título de la página o nombre de la base de datos consultada.

Información sobre la base de datos.

Selección del sistema de ordenación de los resultados.

Elección de las opciones de visualización de los resultados: formato de presentación (reducido o completo), número de registros a visualizar, etc.

Botones de ejecutar consulta y borrar.

Información acerca del tipo de consulta que se está realizando e instrucciones para llevarla a cabo.

Mensajes de ayuda breves.

Navegación con otras páginas de consulta.

A continuación se enumeran los elementos a incluir según el tipo de consulta.

Consulta por campos:

Búsqueda global (en todos los campos).

Lista de campos sobre los que puede realizarse la búsqueda con un cuadro de texto donde el usuario introducirá los términos de búsqueda.

Lista desplegable con los diferentes términos en campos cuyo contenido deba ser seleccionado de una lista controlada.

Consulta experta:

Cuadro de texto para la introducción de la sentencia de búsqueda experta. Esta sentencia definirá las condiciones de búsqueda y estará formada por una combinación de palabras clave que definan campos, términos de búsqueda y operadores aritméticos y/o lógicos (=, AND, OR, NOT, etc.)

Otra opción es la de generar esta sentencia de forma automática a partir de selecciones mediante botones, cuadros de texto, selectores, etc. Por parte del usuario.

Como se realizará la consulta

Tras confirmar la consulta pulsando sobre el botón apropiado se ha de generar de forma invisible para el usuario una sentencia SQL que será enviada al sistema gestor de la base de datos. La conexión entre la página web y el gestor, encargado de generar la sentencia SQL será un CGI diseñado a tal efecto o mediante código incrustado en el propio documento web (por ejemplo PHP).

Páginas de lista

Se presenta una lista de autores, entes, proyectos, lugares, dependiendo de la opción utilizada en la página de consulta por listas. Los elementos existentes en la página de lista serán los siguientes:

Título de la página.

Información sobre el término que se busca y el campo (autor, proyecto, lugar, etc).

Lista alfabética de términos. Cada elemento de la lista será un hipervínculo a la página de resultados final.

Selectores de los términos del listado. En combinación con el botón buscar permitirá continuar la búsqueda seleccionando uno o más términos de la lista.

Botones de búsqueda y borrado. El botón de búsqueda continuará la búsqueda a partir del término de la lista seleccionado. El de borrado borrará las selecciones efectuadas.

Número de elementos visualizado.

Resaltar el término de búsqueda.

Botones de navegación en el listado (adelante, atrás, inicio, fin, mostrar todos).

Página de resultados

Esta página mostrará el resultado final de la consulta en forma de lista de registros con dos formatos de representación a escoger: reducido o completo.

En el formato reducido los resultados muestran los campos más importantes del registro. Cada elemento será seleccionable para ampliar los resultados en una página de resultados en formato completo. En formato completo se mostrará de forma más detallada los datos del registro. En los registros

Los elementos que contendrá esta página serán:

Título de la página.

Información sobre la expresión de búsqueda.

Especificación del número de registros encontrados y el número de visualizados.

Lista de resultados en formato reducido o completo.

Botones de selección del formato de visualización (reducido o completo).

Selectores de los términos del listado en el formato de visualización reducido.
Información de la existencia de imágenes adjuntas al registro. Ampliable a otro tipo de ficheros (audio, video, etc).
Cuadros de imágenes (en el formato completo) con una visualización reducida de la imagen adjunta al registro (<5Kb) en formato jpg. Esta imagen reducida será al mismo tiempo un hipervínculo para la visualización de la imagen en tamaño grande en formato jpg.
Términos de búsqueda resaltados.
Botones de navegación en el listado (adelante, atrás, inicio, fin).

Páginas de administración

Se implementará un apartado privado, accesible solo por el administrador de la aplicación. Este apartado ha de incluir las diferentes opciones relacionadas con el mantenimiento de la aplicación, la gestión de los fondos archivísticos, el mantenimiento de los usuarios registrados y la gestión de las copias de seguridad (aplicación, base de datos e imágenes).

Otras páginas

Índices. Lista alfabética de autores, entidades, fondos, lugares.
Página de identificación. Posibilidad de restringir el acceso con login y password a nivel global o en algunas páginas en concreto. Esto será imprescindible en la catalogación online.
Ayudas.
Descripción de la base de datos.
Recorridos guiados, mostrar resultados a partir de consultas ocultas ya preparadas, por ejemplo guías temáticas, listado de novedades, etc., a través de enlaces.

Requerimientos de las páginas de catalogación

Las personas autorizadas por el administrador del sistema y debidamente autenticadas en su sesión mediante login y password podrán realizar actualizaciones en la base de datos. Esta actualización se realizará de dos formas diferentes: en modo edición y en modo inserción. También tendrá la posibilidad de borrar registros que haya introducido él previamente. Un usuario especial, propiedad del administrador del sistema tendrá autorización total para insertar, modificar o borrar cualquier registro.

Modo inserción

En esta modalidad el usuario introducirá registros nuevos en las diferentes tablas. Después de escoger en qué tabla va a insertar el registro se le mostrará un formulario con los elementos típicos (cuadros de texto, listas controladas, casillas de selección, etc.) donde deberá introducir los datos relativos a un registro. Una vez relleno el formulario, el usuario enviará los datos, la aplicación pedirá confirmación y realizará la inserción del nuevo registro en la tabla correspondiente.

Modo edición

Consiste en la modificación de datos de registros ya existentes. Solo podrán modificar un registro los usuarios que pertenezcan al mismo fondo que el usuario que lo catalogó. Para la

modificación de un registro se realizará una consulta previa, y mediante un botón o enlace de edición se tendrá a un formulario para la modificación de los registros resultados de la consulta. El formulario de edición será idéntico al de inserción, pero se presentará al usuario con todos los datos ya contenidos.

Borrado de registros

Cualquier usuario catalogador tendrá la posibilidad de borrar registros pertenecientes a su fondo archivístico. Igual que en el modo edición, antes deberá realizar una búsqueda y una vez tenga el resultado dispondrá de un botón o enlace para borrar el registro, previa confirmación.

3.Elección de la base de datos, justificación del uso de MySQL

El objetivo de este proyecto en su comienzo era el de migrar la base de datos ISIS existente a una solución más potente. En un primer momento se pensó en Oracle, por tratarse de una de las bases de datos más extendida. Sin embargo, después de recopilar información en Internet, descargar, instalar y probar el software, contactar con proveedores, etc. se llegó a la conclusión de que, aunque en la fase de desarrollo se podía disponer de las herramientas gratuitamente, su posterior utilización en un entorno de producción conllevaría un coste económico bastante alto. El volumen de datos inicial, las perspectivas de crecimiento y las futuras necesidades de seguridad e integridad no justificaban la inversión. Tras evaluar otras opciones “open source” encontré numerosas ventajas y decidimos optar por la solución gratuita.

De entre las soluciones open source más comunes se barajaron dos opciones: MySQL y PostgreSQL, por tratarse de los dos motores de base de datos de más popularidad, de las que se dispone mayor información online y con mayor crecimiento en los últimos meses.

Comparación MySQL – PostgreSQL

Buscando por Internet información y comparativas sobre MySQL y PostgreSQL para ver cuál puede ser mejor, encontramos información abundante aunque a veces algo confusa. De entre todo lo leído, he podido sacar estas conclusiones:

MySQL

Su principal objetivo de diseño fue la velocidad. Con este fin se sacrificaron algunas características esenciales en otros sistemas más "serios". Otra característica importante es que consume muy pocos recursos, tanto de CPU como de memoria. Tiene licencia GPL a partir de la versión 3.23.19.

Ventajas:

- Mayor rendimiento. Mayor velocidad tanto al conectar con el servidor como al ejecutar selects y otras operaciones.
- Mejores utilidades de administración (backup, recuperación de errores, etc).
- Aunque se cuelgue, no suele perder información ni corromper los datos.
- Mejor integración con PHP.
- No hay límites en el tamaño de los registros.
- Mejor control de acceso, en el sentido de qué usuarios tienen acceso a qué tablas y con qué permisos.

- MySQL se comporta mejor que Postgres a la hora de modificar o añadir campos a una tabla "en caliente".

Inconvenientes:

- No soporta transacciones, "roll-backs" ni subselects.
- No considera las claves ajenas. Ignora la integridad referencial, dejándola en manos del programador de la aplicación.

PostgreSQL

Postgres intenta ser un sistema de bases de datos de mayor nivel que MySQL, a la altura de Oracle, Sybase o Interbase. Tiene licencia BSD.

Ventajas:

- Por su arquitectura de diseño, escala muy bien al aumentar el número de CPUs y la cantidad de RAM.
- Soporta transacciones y desde la versión 7.0, claves ajenas (con comprobación de integridad referencial).
- Tiene mejor soporte para triggers y procedimientos en el servidor.
- Soporta un subconjunto de SQL92 MAYOR que el que soporta MySQL. Además, tiene ciertas características orientadas a objetos.

Inconvenientes:

- Consume bastantes más recursos y carga más el sistema.
- Límite del tamaño de cada fila de las tablas a 8k!!! (se puede ampliar a 32k recompilando, pero con un coste añadido en el rendimiento).
- Es de 2 a 3 veces más lenta que MySQL.
- Menos funciones en PHP.

En cuanto a consideraciones de estabilidad del servidor, cada comparativa da datos contradictorios. En general parece que MySQL es más estable, y que Postgres tiende a desperdiciar memoria y sobrecargar bastante el sistema.

Como conclusión a la comparación entre MySQL y Postgres, parece que MySQL junto con Apache y PHP forman un buen equipo para servir páginas web con contenido dinámico. En el planteamiento inicial del proyecto dimos prioridad a la velocidad y el número de accesos concurrentes frente a la seguridad o consistencia (en general bastará con hacer backups periódicos que se restaurarán tras una caída del servidor). Por esta razón nos hemos decantado por MySQL como motor de base de datos para este proyecto.

4 Aplicación desarrollada

Software de consulta que permet l'accés ràpid i senzill a les dades del catàleg informàtic d'arquitectura del CDPAC a través de Internet, amb l'únic requisit d'un PC amb navegador. Basat en una base de dades MySQL i un conjunt de scripts php que gestionen la informació. S'executa conjuntament amb un servidor web Apache sobre sistema operatiu Linux però és possible la instal·lació en qualsevol servidor Windows o Unix.

Permet la cerca d'informació per qualsevol persona des de qualsevol ordinador connectat a Internet i la catalogació de informació pels usuaris registrats, que és consultable de manera immediata.

Està disponible a l'adreça <http://cd-lab06.upc.es>

Catalogació

6 fons arxivístics inicials amb la versió 1: Gaudí (Càtedra i AMAB), HOLSA, Moragas, Sostres i Arxiu Gràfic ETSAB. Ampliable a qualsevol nombre de fons.

Permet la catalogació de fons privats.

Suporta la vinculació de 3 imatges per cada registre i l'enviament directe d'imatges al servidor mitjançant la pròpia aplicació.

5 bases de dades: documents, autors, entitats, esdeveniments i referències bibliogràfiques relacionades entre si.

Sistema d'ajuda online i manual de catalogació

Cerca

5 tipus de cerca: **simple, per camps, per tipologies, per llistes i experta.**

- cerca simple per paraules clau
- cerca per camps a totes les bases de dades
- cerca per llistes ordenades d'autors, entitats, esdeveniments i fons
- cerca per tipologies de projecte sobre els documents
- cerca experta sobre els documents

3 vistes pels resultats: reduïda, completa o detallada.

Possibilitat de seleccionar un o més resultats per canviar de vista.

Visualització dels documents relacionats amb qualsevol registre.

Selecció del nombre de resultats per pàgina i de la ordenació.

Navegació per les pàgines de resultats.

Cerca sobre resultats.

3 tamanyos d'imatge.

Àrea privada i àrea d'administració

Dos zones d'accés restringit pels usuaris catalogadors i administradors.

Opcions pels usuaris catalogadors:

- Cerques especials dels registres propietat del fons de l'usuari, per inicials del catalogador, per data de catalogació o data de modificació
- Enviament i gestió de les imatges del fons
- Accés a documents i manuals de catalogació
- Catalogació de registres 'no consultables'
- Fons privats no consultables amb finalitat docent

Eines d'administració:

- Consulta del log d'accessos
- Gestió de la base de dades
- Gestió de les còpies de seguretat
- Gestió dels usuaris de l'aplicació
- Manteniment de l'ajuda de l'aplicació
- Manteniment dels suggeriments i consultes sobre el catàleg
- Importació massiva de dades

Administración de la aplicación

La página de administración solo es accesible por el administrador con su usuario y permite realizar diferentes acciones relacionadas con el mantenimiento de la aplicación. Las opciones implementadas son:

- Creación de nuevos fondos de catalogación. A través de un formulario el administrador puede dar de alta nuevos fondos archivísticos.
- Log de accesos. Permite ver el histórico de las operaciones realizadas contra la base de datos. Es posible visualizar el log aplicando criterios de selección, como usuario, fondo archivístico, fecha del acceso o dirección IP del usuario.
- Gestión de los usuarios de la aplicación. Permite crear usuarios nuevos, vinculados a un fondo archivístico, borrarlos, modificar sus datos, etc.
- Gestión de las copias de seguridad. A través de esta opción el administrador podrá realizar copias de seguridad, tanto de la aplicación (páginas html y php) como de la base de datos MySQL y de las imágenes.
- Ejecución directa de sentencia SQL. Permite ejecutar directamente sentencias sql sobre la base de datos. En el caso de SELECTS sobre la base de datos MAP la respuesta se muestra en pantalla igual que en el resto de consultas y con las mismas opciones de visualización.
- Gestión de la ayuda de la aplicación. Existe una tabla destinada a almacenar los textos de ayuda relativos a diferentes elementos de las páginas de consulta y catalogación. Esta tabla es consultada cuando el usuario hace clic sobre uno de los iconos de ayuda. Este apartado permite actualizar la información de esta tabla.
- Gestión de las sugerencias de los usuarios. Un formulario de sugerencias permite al usuario realizar cualquier tipo de comentario o sugerencia sobre el catálogo, o el funcionamiento de la aplicación. Desde este apartado el administrador puede gestionar estos comentarios, enviando un correo de respuesta al usuario y dándola por resuelta.
- Pruebas de rendimiento. Se han implementado una serie de pruebas de rendimiento sobre la aplicación, que básicamente realizan operaciones sobre la base de datos y se miden los tiempos de respuesta. En este apartado el administrador puede realizar estas pruebas y extraer resultados.

5 Conclusiones

El desarrollo de la aplicación sobre MySQL y el hecho de que sea un software abierto ha permitido avanzar en paso firme a la hora de resolver los problemas de adaptación que iban surgiendo según se iban teniendo en cuentas los diversos requisitos del sistema.

6 Referencias

- Paul DuBois. *Especial MySQL*. Ed. Prentice Hall
- <http://www.mysql.com>
- <http://www.php.net>
- <http://www.apache.org>

