RED DE REPOSITORIOS DIGITALES PARA DISEMINACIÓN DE INFORMACIÓN

ADOLFO GUZMÁN Y VÍCTOR-POLO DE GYVES
Centro de Investigación en Computación, Instituto Politécnico Nacional, México

BiblioDigital es un paquete que se instala en uno o más servidores y proporciona servicios de guardar, recuperar, buscar e indexar. Es una federación de bibliotecas locales (llamadas repositorios), independientes pero ligadas entre sí por un índice global. Cada uno es un lugar físico (una computadora) donde se almacenan de manera organizada documentos electrónicos para ser suministrados a los usuarios, quienes podrán acceder a ellos desde cualquier punto de Internet. Cada R tiene su propia taxonomía: los *temas* donde los documentos pueden clasificarse o indizarse. Un lector puede conectarse a cualquier R y tener acceso a *todos* los documentos de la federación. Una búsqueda es sólo acceso y manipulación del índice global, pues ya todo documento está indizado, por lo que no es necesario visitar cada documento para ver si en realidad contiene tal o cual palabra o quién es su autor.

Adolfo Guzmán Arenas es ingeniero de la Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME) del Instituto Politécnico Nacional (IPN), México. Es doctor en computación del Instituto Tecnológico de Massachusetts, donde fue profesor asistente en el Departamento de Ingeniería Eléctrica. Recibió en 1996 el Premio Nacional de Ciencias y Artes de manos del presidente Ernesto Zedillo. Es **ACM Fellow**. En la actualidad trabaja en el Centro de Investigación en Computación (CIC) del IPN, el cual fundó en 1996. Se interesa en procesamiento semántico, representación del conocimiento y aplicaciones de sistemas de información.

Víctor Polo de Gyves Montero es ingeniero en Computación de la Unidad Interdisciplinaria (UPIICSA) del Instituto Politécnico Nacional de México y labora en SoftwarePro International (México). Sus intereses son Unix, *software* libre y la confección de aplicaciones de tecnología avanzada.

RED DE REPOSITORIOS DIGITALES PARA DISEMINACIÓN DE INFORMACIÓN

ADOLFO GUZMÁN Y VÍCTOR-POLO DE GYVES
SoftwarePro International[1]

BIBLIODIGITAL

BiblioDigital es una federación de bibliotecas locales (llamadas repositorios), independientes pero ligadas entre sí por un índice global. Cada uno es un lugar físico (una computadora) donde se almacenan, de manera organizada, documentos electrónicos para ser suministrados a los usuarios, quienes podrán accesarlos desde cualquier punto de Internet. Cada R tiene su propia taxonomía: los *temas* donde los documentos pueden clasificarse o indizarse.

BiblioDigital está formada por un repositorio padre y cero o más repositorios hijos. Cada documento reside en exactamente un R. Cada R reside en una PC con bastante disco, no *break*, anti-virus, conexión a Internet…

El bibliotecario administra un R: da de alta a autores y editores, e inicialmente construye la taxonomía de su repositorio. Los lectores no necesitan, pero pueden, darse de alta.

Un lector puede conectarse a cualquier R y tendrá acceso a todos los documentos de la federación, no sólo a los del R al que está conectado. En BiblioDigital, un lector puede añadir *comentarios* a un documento leído.

Los autores pueden agregar nuevos documentos a su R, actualizar los suyos ya existentes y añadirles documentos complementarios.

También existen *editores*, que son dueños de *colecciones*. Más adelante se proporcionan mayores detalles al respecto.

HOJEANDO LOS DOCUMENTOS

Un lector que desee saber qué documentos contiene la federación puede visitar la taxonomía temática de cualquier R y pulsar en un nodo de ella; la descripción (metadatos) de los documentos que yacen en ese tema aparecerán en la pantalla. Lo mismo puede hacer visitando la taxonomía de conceptos. Ver figura 1.

ACCESO A LOS DOCUMENTOS (BÚSQUEDAS)

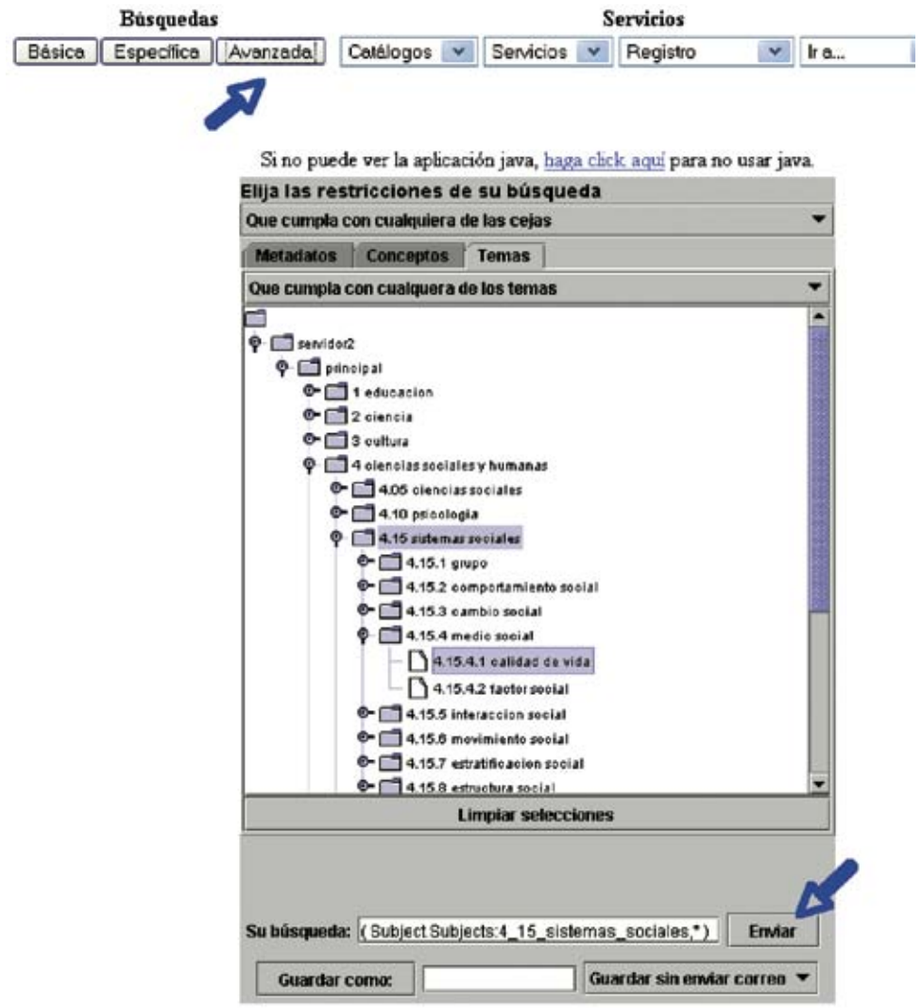
- **Por tema.** La taxonomía temática de un R la define su bibliotecario. Cada autor clasifica su documento en uno o varios de estos temas predefinidos (vocabulario controlado).
- **Por concepto.** La taxonomía de conceptos la da el sistema. Éste (a través de Clasitex ®) clasifica [de manera automática] cada documento en los temas que aborda.
- **Por las palabras y frases temáticas** ("Por mi raza hablará el espíritu") que contiene en el texto completo. En este caso la búsqueda se hace mediante el uso de lematización (énfasis en las raíces de cada vocablo), por lo que una búsqueda de la frase "escuela danesa" también encontrará "escuelas danesas". Ver figura 3 y figura 4.
- **Por metadato.** Es decir, por autor, por título, por lenguaje del documento…
- **Mediante combinaciones** de las opciones anteriores. Esto dota al usuario de un poderoso sistema de búsqueda de documentos.

Los documentos están disponibles en su R el día de su publicación y en forma global al siguiente día: la sincronización entre repositorios ocurre cada madrugada.

1 BiblioDigital es propiedad de SoftwarePro International. Adolfo Guzmán es investigador del CIC-IPN (México).

FIGURA 1.

HOJEANDO LA TAXONOMÍA TEMÁTICA. AL PULSAR "BÚSQUEDA AVANZADA" APARECE UN ÁRBOL CON LA ESTRUCTURA TEMÁTICA DEL R. EL LECTOR PUEDE SELECCIONAR UN NODO, ABRIRLO Y MOSTRAR SUS NODOS HIJOS, ETC. AL PULSAR "ENVIAR", APARECERÁN A LA DERECHA DE LA PANTALLA LOS METADATOS DE LOS DOCUMENTOS CONTENIDOS EN LOS TEMAS SELECCIONADOS ("SISTEMAS SOCIALES" Y "CALIDAD DE VIDA", EN EL EJEMPLO)



COLECCIONES

Una colección (ver figura 2) es un conjunto de documentos a los que se les asocia un nombre (por ejemplo, "Documentos para el curso de Química Básica", "Concurso de Poesía", "Revista Digital Computación y Sistemas"), por ser conveniente considerarlos colectivamente. Un editor puede ser editor de varias colecciones. Un documento puede pertenecer a cero o más colecciones. Las colecciones pueden ser abiertas (cualquier autor puede introducirle documentos de su autoría) o cerradas (sólo el editor puede introducirle documentos).

Cada colección reside en exactamente un R; una colección contiene en realidad apuntadores a documentos ya existentes en cualquier R, dados de alta con anterioridad por su autor.

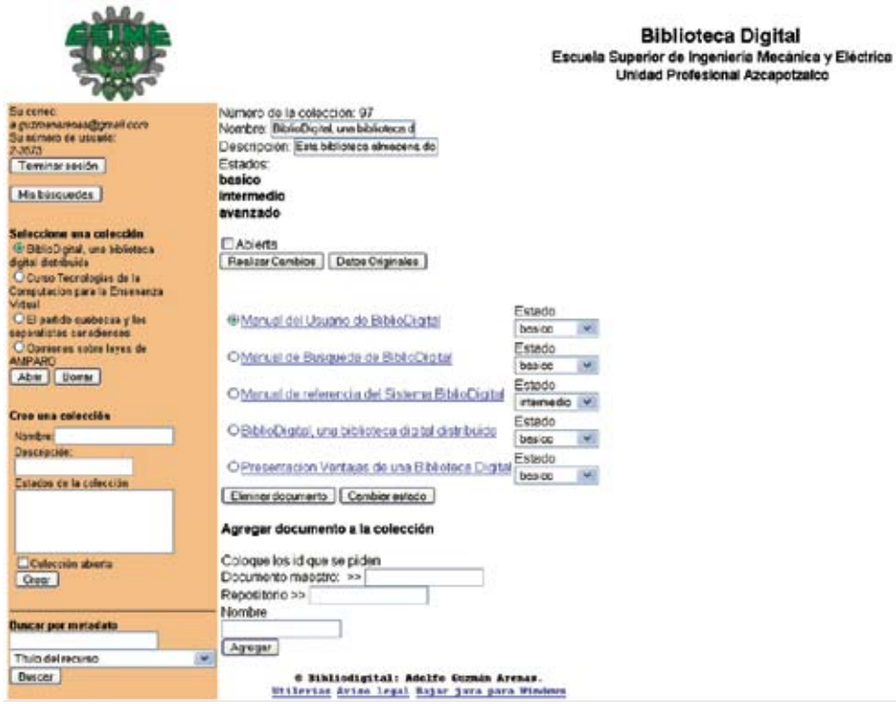
Cada documento en una colección está en un estado ("recibido", "aprobado", "publicado"...). El editor define estos estados para los documentos de su colección. El editor cambia de modo manual el estado de un documento de su colección, por ejemplo, de recibido a aprobado, para que los autores del documento se enteren de su situación actual en la colección. Por ejemplo, ¿cuál es el estado de mi tarea que envié a la colección "Tareas del curso de Física"?

SUSCRIPCIÓN A BOLETINES PERSONALIZADOS (ALERTA TEMPRANA)

Un lector puede indicar su perfil (o perfiles) de temas, conceptos y palabras clave que le interesan. Entonces el sistema le envía una vez a la semana o a la quincena... por correo electrónico, un boletín de noticias donde aparecen los títulos y resúmenes de los nuevos documentos que coinciden con su perfil, para que él pueda consultarlos en un momento oportuno. Para esto, un autor debe registrarse con el bibliotecario.

FIGURA 2.

COLECCIÓN "BIBLIODIGITAL, UNA BIBLIOTECA DIGITAL DISTRIBUIDA". UNA COLECCIÓN PERMITE CONSIDERAR VARIOS DOCUMENTOS EN CONJUNTO. EL EDITOR (A.GUZMANARENAS@GMAIL.COM) HA ENTRADO A ESTA COLECCIÓN SUYA. LA COLECCIÓN NO ES ABIERTA. CONTIENE CINCO DOCUMENTOS, CUATRO SON BÁSICOS Y UNO ES INTERMEDIO. EL EDITOR PUEDE AGREGAR O ELIMINAR DOCUMENTOS A SU COLECCIÓN



ARAÑAS. DOCUMENTOS EXÓGENOS

No importa cuántos documentos pueda haber en una federación, siempre habrá más documentos afuera (en la Web). Por ello, y para aprovechar esta riqueza exógena, BiblioDigital puede leer e indizar (por conceptos y por palabras contenidas) los documentos "fuera de BiblioDigital". Cada R posee una araña (crawler), que busca de manera automática en la Web documentos a partir de una colección inicial de sitios dados por el bibliotecario. Si los documentos pertenecen a la temática del R, serán indizados en ella y opcionalmente copiados a R. De esta forma las arañas enriquecen cada repositorio, pues le agregan documentos relevantes.

CARACTERÍSTICAS IMPORTANTES DE BIBLIODIGITAL

- Un lector puede conectarse a cualquier R y tener acceso a todos los documentos de la federación.
- Una búsqueda recobra documentos de todas las R. Una búsqueda es meramente acceso y manipulación del índice global, dado que ya todo documento está indizado. Por tanto, no es necesario visitar cada documento para ver si en realidad contiene tal o cual palabra o quién es su autor.
- Un bibliotecario (administrador de un R) da de alta a autores y a editores; los lectores no necesitan darse de alta; los documentos son en principio gratuitos, sin encripción y públicos (cualquiera puede leerlos).
- Permite versiones de un documento, documentos accesorios (ejercicios, software, etcétera).
- Permite indizar (y entregar texto completo) documentos que yacen fuera de la federación de R.
- Un autor proporciona una ficha descriptiva (metadatos en el estándar Dublin Core, por ejemplo) del documento que él aporta.
- Puede almacenar documentos en cualquier formato, aunque está diseñada para manejar los formatos populares (Word, Excel, texto plano, PDF, PowerPoint, imágenes, .mpg...).
- Cada R tiene su propia taxonomía temática y al mismo tiempo utiliza otra taxonomía global de conceptos (impuesta por Clasitex).
- Los documentos pueden ser públicos o tener niveles de seguridad.
- Los repositorios comparten un índice global que se actualiza cada madrugada.
- Es posible saber la localidad (URL) donde yace un documento o colección, y enviar esta dirección a colegas interesados (en vez de enviar el documento completo).
- Un lector puede enviar correos al bibliotecario, autores y editores.
- Además de la búsqueda avanzada, existen búsquedas básicas (sencillas), catálogos de autores, de obras, de temas.

FIGURA 3.

BUSCA DOCUMENTOS CUYO CONTENIDO (TEXTO COMPLETO) CONTENGA TORO, REDONDEL, BUREL O CAPOTE

FIGURA 4.

SE ENCONTRARON 149 DOCUMENTOS COMO RESULTADO DE LA BÚSQUEDA DE LA FIGURA 3, ENTRE ELLOS "EL CUENTO, LAS CURAS Y LOS PRODIGIOS DE LA MEMORIA", DE SERGIO PITOL

BÚSQUEDAS AVANZADAS O DE MARKOV

Existe una versión inicial de localización de documentos según la dinámica o patrón de consultas de un usuario. Cada R observa la secuencia de búsquedas de un usuario y lleva estadísticas de lectura, para ofrecerle de manera espontánea documentos que podrían serle útiles. Por ejemplo, "Muchos lectores que consultan el documento A y luego el B a continuación visitan el C; aquí está el documento C". Estas búsquedas son proactivas; el lector no tiene control sobre ellas, aunque puede ignorar sus sugerencias. El bibliotecario puede desactivar estas búsquedas, si considera que atentan contra la privacía de sus lectores.

TAXONOMÍAS Y EL BIBLIOTECARIO

La manera más fácil de crear una taxonomía es partir de un archivo de texto que la contenga. Una utilería de BiblioDigital toma ese archivo y crea la taxonomía. Existen buenas taxonomías del dominio público y se recomienda su uso.

Se dispone de un editor de taxonomías para modificar en forma interactiva las taxono-mías de temas y conceptos: crear o eliminar nodos, o cambiarlos de lugar.

La taxonomía temática se usa para que el autor clasifique su documento en uno o varios temas y para que el lector localice documentos según su temática. El autor lleva a cabo esta clasificación cuando ingresa su documento al R. Es posible pero engorroso reclasificar un documento mal clasificado: habrá que darlo de baja (borrarlo) y volverlo a ingresar. Es un proceso intencionalmente penoso.

La taxonomía de conceptos la maneja Clasitex; el autor o bibliotecario no tienen control sobre ella. Sirve, como ya se dijo, para buscar documentos según sus conceptos.

No es recomendable que el bibliotecario haga cambios frecuentes a la taxonomía temática, ya que desconciertan a los lectores y hacen que los documentos viejos continúen clasificados en temas ahora inexistentes.

Es imposible que un lector registre (indize, clasifique) un documento suyo en un tema inexistente en la taxonomía temática.

Es deber del bibliotecario revisar de modo ocasional los documentos ingresados por los autores, para detectar documentos irrelevantes, contrarios a la política editorial de la empresa, o mal clasificados. Otro deber suyo es respaldar periódicamente la información de su R.

ALGUNOS SERVIDORES DE BIBLIODIGITAL:

(Cada instalación tiene su responsable y sus políticas de uso).

http://148.204.20.100:8080/bibliodigital,
Biblioteca Digital, ESIME Atzcapotzalco en el CIC.

http://148.204.74.44:8080/bibliodigital,
Biblioteca Digital, ESIME Azcapotzalco.

http://148.204.245.235:8080/bibliodigital,
Biblioteca Digital, ESIME Culhuacán.

http://148.204.46.43:8080/bibliodigital,
Centro de Formación e Innovación Educativa.

http://148.204.197.10:8080/bibliodigital,
Centro de Tecnología Educativa del IPN (Ágora).

http://148.235.144.117:8080/bibliodigital,
Universidad Tecnológica de la Sierra de Hidalgo.

http://201.134.139.88:9080/bibliodigital,
Centro Cultural del México Contemporáneo.

Más información: en http://alum.mit.edu/www/aguzman, en http://aguzman.blog.com.