Ciencias de la Información Vol. 44, No. 3, septiembre - diciembre, pp. 51-54, 2013

Indexación de metadatos en bibliotecas digitales mediante protocolos de comunicación

Yurelkys de los Angeles Carreras Riopedre1
Juan Manuel Rey Alvarez

El desarrollo de internet y la disponibilidad de los documentos en formato digital han permitido que las bibliotecas digitales se nutran de fuentes ya existentes, mediante protocolos de comunicación. Este intercambio de metadatos ha posibilitado la recolección y almacenamiento de información desde diversos orígenes permitiendo que los usuarios se interactúen con una única interfaz de búsqueda. El objetivo de este trabajo es abordar las características y funcionamiento de los protocolos más usados mundialmente en el intercambio de información entre bibliotecas digitales, protocolos que permiten la indexación de metadatos de diversas fuentes. Además de exponer los resultados obtenidos en la implementación de un sistema de indexación de documentos para la Biblioteca Digital ¨Alma Mater¨, haciendo uso del protocolo OAI-PMH.

***Palabras clave**: indexación, biblioteca digital, OAI-PMH, protocolo de comunicación.*

RESUMEN

ABSTRACT

The development of the Internet and the availability of documents in digital format have allowed digital libraries nurture existing sources through communication protocols. This metadata exchange has enabled the collection and storage of information from diverse backgrounds allowing users to interact with a single search interface. The aim of this paper is to address the characteristics and performance of the most used worldwide in the exchange of information between digital libraries, protocols that allow metadata indexing from different sources. In addition to presenting the results obtained in the implementation of a system for indexing documents for "Alma Mater" Digital Library, making use of the OAI-PMH.

***Keywords**: Gross standardized stress, Tucker congruency coefficient, non parametric statistical method.*

Introducción

Con el surgimiento y desarrollo de las tecnologías de la información y las comunicaciones, se ha logrado almacenar, compartir y gestionar grandes colecciones de datos en formato digital. Con el uso creciente del acceso en línea, se crearon repositorios de documentos de todo tipo, generados por diferentes instituciones. Repositorios que aumentaron considerablemente su volumen de información y cambiaron su visualización adoptando interfaces cada vez más sencillas de utilizar. Estos avances conllevaron al surgimiento de las bibliotecas digitales, cuyo objetivo es el acceso universal a la información, sin limitantes de tiempo ni espacio, asegurando la disponibilidad, recuperación y autenticidad de los documentos.

La disponibilidad de los documentos en formato digital y el desarrollo de internet, han permitido que las bibliotecas digitales que en principio sus documentos eran introducidos manualmente ahora se nutran de fuentes ya existentes, a través de protocolos de comunicación que permiten el intercambio de metadatos y brindan la posibilidad de recolectar y almacenar información de diversos orígenes.

Desarrollo

Las bibliotecas digitales deben permitir la recuperación de información mediante metadatos, los que posibilitan una búsqueda efectiva y precisa proporcionando valor añadido a la mera acumulación de información. (Dora Pérez, 2007)

Los mecanismos de comunicación que han impulsado la tendencia de las bibliotecas digitales a crecer indexando contenidos, parten de la estandarización de la información. Donde juegan un papel importante los formatos para describir los documentos o formatos para la representación de metadatos. Los metadatos son datos sobre los documentos. Este término se refiere a cualquier dato que se utilice para ayudar a identificar, describir y localizar recursos electrónicos enlazados en red.

Existen varios formatos para la representación de metadatos entre ellos se encuentra MARC que es un conjunto complejo de estándares para describir, almacenar, manipular y recuperar datos bibliográficos legibles por ordenador. Es un estándar altamente desarrollado que se diseñó originalmente en los años 60 para la descripción de libros impresos y que ha seguido adaptándose para proporcionar descripción, acceso y localización de la información de los recursos en la red. (Virginia Ortlz. Replao Jiménez, 1999)

Otro formato para la representación de metadatos es Dublin Core, esta la forma abreviada para el Dublin Metadata Core Element Set (Conjunto básico de elementos metadatos de Dublín). Su objetivo principal es crear un conjunto de elementos de datos que describan los documentos electrónicos de las redes con el fin de facilitar su búsqueda y recuperación. Está compuesto por un conjunto mínimo de elementos que facilitan la recuperación de información en la red. Está diseñado para facilitar la recuperación de recursos en las redes de una forma similar a un catálogo de biblioteca, pero con una estructura mucho más sencilla. Está formado por 15 elementos de datos y su punto más fuerte es que el diseño es tan intuitivo que los propios proveedores de información pueden codificar sus documentos al mismo tiempo que los crean. (Virginia Ortiz. Replao Jiménez, 1999)

La existencia de formatos de metadatos para describir los documentos es de suma importancia, pero no solo basta con que estén descritos, es necesario que esa información se transmita, lo cual se hace a través de los protocolos. Los protocolos son un conjunto de hábitos y procedimientos utilizados en las relaciones interpersonales. Cuando es usado bajo el contexto de redes de comunicación, el termino protocolo tiene un significado similar pero a un nivel más específico. Un protocolo de red es un conjunto de reglas, secuencias, formatos de mensajes y procedimientos bien detallados que posibilitan la transferencia de datos entre dos o más sistemas de computación, el término utilizado para describir como los sistemas de computación se comunican con otros a nivel de bit y de byte.(ing. Carmen L. Duran Gil, 2007)

Existen varios protocolos para el intercambio de información entre bibliotecas digitales, los más usados son el Z39.50 y el OAI-PMH.

El protocolo Z39.50 es un estándar para la recuperación de la información basado en la estructura cliente/servidor, que facilita la interconexión de sistemas informáticos. Uno de los beneficios básicos del protocolo, en el ámbito de las bibliotecas y de los centros de documentación, es que hace posible la comunicación entre sistemas que utilizan diferente hardware y software.

Permite la realización de búsquedas simultáneas a múltiples bases de datos, sin tener que conocer para ello las sintaxis de búsqueda que utilizan dichos sistemas, utilizando una única interfaz de usuario para recuperar la información, ordenarla, y exportar los registros bibliográficos.

Facilita la interconexión entre los usuarios y las bases de datos donde se encuentra la información que necesitan a partir de una interfaz común y facilita el manejo, independientemente del lugar en que se encuentren las bases de datos así como la estructura y la forma de acceso de éstas.

Especifica un conjunto de reglas para gestionar las formas y procedimientos de interconexión remota de computadoras, con el propósito de buscar y recuperar información, aunque su aplicación actual es más amplia puesincluye la consulta y el intercambio de datos bibliográficos, la intercomunicación de índices y resúmenes de: información geoespacial, documentos oficiales, objetos digitales o de metadatos que describen los documentos de las bibliotecas electrónicas y digitales.(Aldo Guajardo Salinas, 2010)

Los servicios o facilidades principales que brinda el protocolo son:

1. La inicialización, precursora del trabajo real, en la que se establecen los parámetros básicos de la sesión que se va a iniciar entre el cliente y el servidor, permitiendo la autenticación del usuario. La negociación incluye la versión del protocolo, las operaciones que podrán efectuarse, juegos de caracteres, lenguas, segmentación y tamaño de la información.

2. La búsqueda, funcionalidad más importante del estándar, que permite realizar búsquedas simples o complejas con la misma herramienta a múltiples bases de datos, agilizando la recuperación de información. Las estrategias de búsqueda pueden utilizar operadores booleanos, de proximidad y otros.

3. La recuperación de la información: una vez realizada la búsqueda, el cliente solicita al servidor los registros que quiere visualizar, que dependiendo del número solicitado, podrán aparecer segmentados en conjuntos de registros.

El protocolo posibilita controlar el acceso, realizar búsquedas utilizando índices, ordenar la información recuperada, y poder acceder a información sobre el servidor y los servicios que ofrece. (William Moen, 1997)

El protocolo OAI-PMH o Iniciativa Abierta de Archivos es una herramienta de interoperabilidad que posibilita el intercambio de metadatos sobre cualquier material almacenado en soporte electrónico. (Aldo Guajardo Salinas, 2010)

La arquitectura del protocolo rechaza la búsqueda distribuida, a diferencia del Z39.50. Utiliza transacciones HTTP para emitir preguntas y obtener respuestas entre un servidor o archivo y un cliente o servicio recolector de metadatos. El segundo puede pedir al primero que le envíe metadatos según determinados criterios como la fecha de creación de los datos. En respuesta el primero devuelve un conjunto de registros en formato XML, incluyendo identificadores de los objetos descritos en cada registro. (José Manuel Barrueco. Imma Subirats Coll, 2003)

Las peticiones se emiten utilizando los métodos GET o POST del protocolo HTTP y constan de una lista de opciones con la forma de pares del tipo: clave = valor. Existen seis peticiones que un cliente puede realizar a un servidor:

- ***GetRecord:*** Utilizado para recuperar un registro concreto. Necesita dos argumentos, identificador del registro pedido y especificación del formato bibliográfico en que se debe devolver.

- ***Identify:*** Utilizado para recuperar información sobre el servidor: nombre, versión del protocolo que utiliza, dirección del administrador, etc.

- ***ListIdentifiers:*** Recupera los encabezamientos de los registros, en lugar de los registros completos. Permite argumentos como el rango de fechas entre los que queremos recuperar los datos.

- ***ListRecords:*** Igual que el anterior pero recupera los registros completos.

- ***ListSets:*** Recupera un conjunto de registros. Estos conjuntos son creados opcionalmente por el servidor para facilitar una recuperación selectiva de los registros. Sería una clasificación de los contenidos según diferentes entradas. Un cliente puede pedir que se recuperen solo los registros pertenecientes a una determinada clase. Los conjuntos pueden ser simples listas o estructuras jerárquicas.

- ***ListMetadataFormats:*** Devuelve la lista de formatos bibliográficos que utiliza el servidor.

El protocolo soporta múltiples formatos para expresar los metadatos, no obstante requiere que todos los servidores ofrezcan los registros utilizando Dublin Core codificado en XML. Además de éste formato cada servidor es libre de ofrecer los registros en otros formatos adicionales como MARC. Un cliente puede pedir que los registros se le sirvan en cualquiera de los formatos soportados por el servidor. (José Manuel Barrueco. Imma Subirats Coll, 2003)

La arquitectura de este protocolo se basa en proveedores de datos y proveedores de servicios; los primeros son los archivos que proporcionan la información y los segundos son los recolectores o servicios que toman los datos, y los presentan a los usuarios finales.

La aplicación de OAI-PMH permite realizar el intercambio de información para que desde puntos centralizados, proveedores de servicios puedan realizar búsquedas conjuntas sobre los metadatos de todos los repositorios asociados, específicamente los Open Access (Acceso Abierto) que promueve eliminar las barreras económicas, legales y tecnológicas, y trata de obtener a cambio, como beneficios, una mayor accesibilidad para los documentos y una mayor visibilidad para los autores. Los documentos que están disponibles en Acceso Abierto son más consultados y tienen más posibilidades de ser citados.

El Ministerio del Poder Popular para la Educación Universitaria en Venezuela se encuentra desarrollando el proceso de municipalización de la Educación Superior, en apoyo a este proceso se creó la Biblioteca Digital "Alma Mater", como espacio para acceder a los repositorios de universidades y a los artículos de revistas científicas de interés para los distintos programas de formación.

Motivado por la necesidad de dicha biblioteca de poseer una herramienta que le posibilitara un crecimiento acelerado de sus fondos a partir de la recolección de metadatos desde fuentes externas, se desarrolló un sistema de indexación de documentos. El sistema de indexación de documentos para la Biblioteca Digital ''Alma Mater'' se basa en el protocolo OAI-PMH y le permite a dicha biblioteca recolectar de forma rápida documentos disponibles en fuentes externas específicamente en canales de información de acceso abierto.

El sistema de indexación de documentos para la biblioteca digital ¨Alma Mater¨, se creó como un módulo para el sistema de gestión de contenidos Drupal, en su versión 6.17, el módulo llamado dataprovider _reader se basó en el protocolo OAI-PMH, específicamente empleó la petición identify**,** para comprobar que la fuente de la que se va a indexar documentos es real y confiable. Se usó el ListMetadataFormats para verificar que se emplea dublin core como formatos de metadatos y List Records para recuperar la información de todos los documentos que se indexan en la biblioteca.

En el proceso de investigación y desarrollo del módulo se utilizó el método analítico -sintético, para hacer un análisis detallado de los componentes de los protocolos de intercambio de metadatos y utilizar los resultados del análisis como base para el diseño del sistema. Este método también fue usado en el estudio y selección de las tecnologías y herramientas empleadas para desarrollar sistemas de indexación de metadatos.

Se hizo uso de método histórico-lógico para realizar un análisis del surgimiento y desarrollo de las bibliotecas digitales fundamentalmente la evolución en su funcionamiento y las tendencias actuales del mismo. Se realizaron además encuestas para identificar preferencias y facilidades de uso para los usuarios de catálogos de bibliotecas digitales, con diversas interfaces.

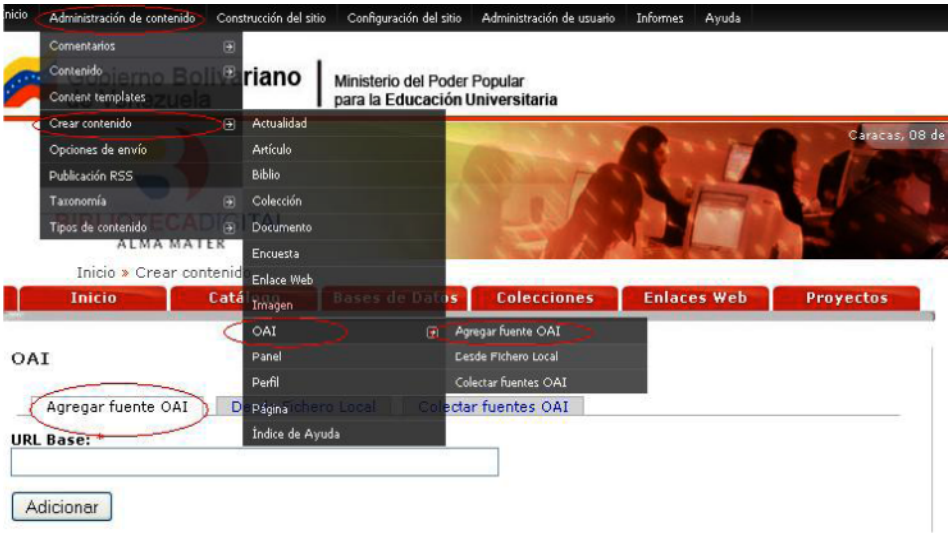

Fig. 1. Funcionalidad agregar fuente OAI.

En las figuras 1 y 2 se muestran imágenes del módulo dataprovider _reader en funcionamiento.

Conclusiones

Los protocolos para el intercambio de información entre las bibliotecas digitales posibilitan un crecimiento acelerado de los fondos de las mismas, al indexar documentos que brindan otras fuentes.

El Z39.50 es un protocolo para la recuperación de información basado en la estructura cliente-servidor que facilita la interconexión de sistemas informáticos, por lo que constituye un gran avance en la interconexión entre sistemas bibliotecarios al permitir superar las enormes barreras que conlleva operar con diferentes sistemas informáticos. Mientras que el protocolo OAI-PMH rechaza la búsqueda distribuida, pero proporciona metadatos, sujetos a criterios de alcance bastante simples, con lo que brinda mayor rapidez en la recuperación de información.

El módulo desarrollado dataprovider _reader para la Biblioteca Digital "Alma Mater" es una herramienta flexible y configurable que se puede integrar a cualquier sistema desarrollado con el CMS Drupal, para importar documentos desde fuentes externas, emplea el protocolo OAI-PMH para la indexación de documentos lo que brinda mayor rapidez en la recuperación de información, asegura un crecimiento acelerado de los fondos de la biblioteca y la rápida recuperación de información en las búsquedas del catálogo.

Figura 2. Funcionalidad importar documentos desde archivo local.

Bibliografía

Guajardo Salinas, Aldo. (2010). Z39.50 y OAI-PMH: Protocolos de Transferencia y Recuperación de Información,.Disponible en: http://102novadoc.es/masinfo/oai-chile.pdf.

Duran Gil, Carmen. (2007). Protocolos, Disponible en: http://www.fortunecity.es/imaginapoder/nada/617/PT111.htm.

Pérez, Dora (2007). La biblioteca digital, Disponible en:http://www.uoc.edu/web/esp/articles/La_biblioteca_digital.htm.

Barrueco, José Manuel, Subirats Coll, Imma. (2010). OAI-PMH: Protocolo para la transmisión de contenidos en Internet, Disponible en: http://www.uv.es/=barrueco/cardedeu.doc .

Ortiz, Virginia, Jiménez, Repiso. (1999). Nuevas perspectivas para la Catalogación: Metadatos versus MARC, Revista española de Documentación Científica, Vol 22 (No 2): 198 – 219.

Moen, William. (1995)The ANSI/NISO Retrieval in the Information Infrastructure. Disponible en: http://old.cni.org/pub/NISO/docs/Z39.50-brochure/50.brochure.toc.html.

Recibido: 10 de septiembre de 2013
Aprobado en su forma definitiva:
14 de noviembre de 2013

Yurelkys de los Angeles Carreras Riopedre1
Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños Km 2 1/2, La Habana, Cuba.

Juan Manuel Rey Alvarez
Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños Km 2 1/2, La Habana, Cuba.