El concepto de metadato. Algo más que descripción de recursos electrónicos

José A. Senso
Doctor en Documentación. Profesor del Departamento de Biblioteconomía y Documentación. Universidad de Granada. España.
E-mail: jsenso@ugr.es

Antonio de la Rosa Piñero
Licenciado en Documentación. Software engineer and consultant. Wisdom. Amsterdam. Holanda.
E-mail: antonio@wisdom.nl

Resumen

Resulta evidente la necesidad de establecer mecanismos que permitan una descripción más exhaustiva de los recursos electrónicos. En este trabajo se propone como solución el uso de metadatos. Por ese motivo se estudia el concepto de metadato con el fin de determinar tanto su campo de acción como los diferentes presupuestos subyacentes en este.

Palabras clave

Recuperación de información; Metadatos; Objetos; Descripción de recursos.

The metadata concept. Something more than description of electronic resources

Abstract

It is evident the necessity to establish mechanisms that allow a more exhaustive description of the electronic resources. In this work we propose as solution the use of metadata. For that reason the concept of metadata is studied for the purpose of determining their action field and the underlying different meanings in this.

Keywords

Information retrieval; Metadata; Objects; Resource description.

INTRODUCCIÓN

Desde hace varios años, los profesionales de la información tienen que afrontar el reto de solucionar los problemas de recuperación de información provocados por la aparición de las bibliotecas digitales y el aumento vertiginoso de la información en formato electrónico.

Como resultado de este crecimiento, cada vez es más difícil el proceso de localizar información. Se han propuesto distintos mecanismos capaces de, por un lado, superar las limitaciones de los sistemas de recuperación basados en la navegación hipertextual – recordemos que el Web no estaba planeado para permitir esto (Husby, 1997) – y, por otro, facilitar la recuperación de la información a texto completo. Las propuestas se han materializado en tres grandes líneas de acción[*]:

1. Índices compilados manualmente

2. Bases de datos creadas por robots o arañas

3. Métodos de indización distribuida

Vamos a analizar detenidamente estos tres modelos.

Índices compilados manualmente

Se trata de grandes bases de datos donde los usuarios o creadores de las páginas Web sugieren su ubicación dentro de unas categorías mediante un formulario. En la mayoría de ocasiones esta organización es la que se traslada directamente al servicio (caso de Terra). En Yahoo, por el contrario, son profesionales quienes las evalúan, organizan y clasifican en la categoría adecuada. Para realizar la consulta, un gestor de páginas Web se encarga de hacer de pasarela entre la base de datos y el usuario que consulta (Montes Hernández, 1999).

[*] Obviamos los servicios que se engloban dentro del grupo de los "metabuscadores" ya que no se pueden considerar como un sistema de información completo (no realizan tareas de indización, y el trabajo de recuperación de la información lo llevan a cabo por medio de llamadas a procesos generados por otros sistemas).

En cuanto a la forma de realizar la búsqueda, la información está clasificada en varios grupos conceptuales encabezados por términos generales, y cada grupo se encuentra subdividido, a su vez, en más subcategorías a través de las cuales se va descendiendo en niveles de especificidad.

Bases de datos creadas por robots o arañas

Partiendo del esquema clásico: una interface, un motor de búsqueda, y una base de datos, los buscadores utilizan un robot para la alimentación automática de su base de datos. El robot – también llamado araña – es un programa de ordenador que está diseñado para recorrer de forma automática la estructura hipertexto de un servidor Web con el fin de alimentar bases de datos textuales a partir de documentos HTML, así como otro tipo de formatos de edición electrónica, distribuidos en diferentes servidores.

Tomando como punto de partida una URL inicial, el robot recupera un fichero en formato HTML que transfiere al sistema local, de forma similar a como lo hace un cliente Web, pero, una vez recuperado, en lugar de proceder a su visualización, se sirve de él para generar nuevos registros en una base de datos. Cada entrada de esta base de datos recogerá la URL completa del documento y una serie de palabras significativas extraídas, bien de los fragmentos con un mayor contenido informativo (< TITLE >, < H1 >, etc.), o bien a partir de su frecuencia de aparición en el documento (Harvest, 2000).

Una vez indizado el documento, el robot identifica las referencias hipertextuales que contiene y que nos dirigen a otras unidades informativas en el mismo o en otros servidores de la Red. De forma recursiva, el robot recupera los documentos referenciados en estos nexos, procediendo a su indización, obtención de nuevas referencias, etc.

Complementando al robot encargado de la extracción/ indización de documentos, encontraremos un motor de búsqueda que permite interrogar estas bases de datos desde los clientes WWW mediante programas de distinta complejidad accesibles mediante la especificación CGI*.

Evidentemente las estrategias de selección de URLs iniciales, extracción de contenido de los documentos y asignación de valores a estos términos de indización están abiertas a numerosas posibilidades, y cada implementación ha podido optar por distintas alternativas.

Métodos de indización distribuida

El ejemplo clásico de este sistema lo encontramos en el servicio Harvest, que surge a finales de 1993 dentro de la línea de trabajo del IRTF-RD (Internet Research Task Force Research Group - Resource Discovery).

Si bien es posible identificar similitudes con Aliweb*, Harvest no se basa en un "esfuerzo humano distribuido", sino en una arquitectura hardware y software repartida entre distintos servidores Web. Distinguiremos dos elementos principales en el modelo Harvest:

– Gatherers: un software instalado en un servidor Web que periódicamente extrae información relativa a los ficheros disponibles (en ese mismo servidor) para la comunidad de usuarios de Internet.

– Brokers: recuperan automáticamente la información extraída por uno o más gatherers y la integran en índices sobre los que se podrán lanzar ecuaciones de búsqueda.

– La comunicación entre brokers y gatherers utiliza como protocolo un sistema de metadatos denominado SOIF (Summary Objetc Interchange Format). En la actualidad se está trabajando para que el fichero de intercambio se genere también en formato RDF (Resource Description Format) (WebTop, 2000).

La solución al problema

Independientemente del sistema utilizado para alimentar la base de datos, siempre nos encontraremos con varios problemas. Por un lado los servicios de búsqueda recuperan gran cantidad de documentos que, en la mayoría de los casos, no satisfacen las necesidades de información al no ser pertinentes. Y esto es debido a que los documentos de la Red carecen de datos suficientes para la descripción (Gill, 1998; Ortiz-Repiso, 1999).

* Common Gateway Interface, especificación técnica que posibilita la mayor interacción entre clientes y servidores WWW. La principal prestación de CGI se centra en la generación de documentos HTML de forma dinámica, es decir, enviando al cliente un documento previamente inexistente. El documento puede consistir en una página HTML, una imagen, texto plano, etc., pudiendo incluir información procesada por el servidor como resultado de un cálculo o de la consulta a una base de datos.

* El modelo Aliweb propone que sean los administradores de servidores web quienes alimenten las bases de datos. Para ello, el administrador de cada servidor debe generar un fichero en un formato estándar (IAFA) donde incluiría el nombre de cada uno de los ficheros que forman su web así como una serie de palabras claves que identifiquen su contenido.

Gran parte de consultas realizadas sobre un motor de búsqueda cualquiera de la Red genera una excesiva cantidad de ruido en la recuperación a menos que el usuario sea capaz de formular complejas ecuaciones de búsqueda. Aun así, el nivel de precisión es relativamente bajo (Olvera Lobo, 2000). Por ejemplo, a mediados de agosto de 2000, la empresa británica WebTop hizo pública una encuesta según la cual el 82% de los internautas británicos no lograba encontrar la información que buscaba en la Red (WebTop, 2000).

El hecho de que la mayoría de páginas Web apenas utilice descripciones básicas para informar del contenido de las mismas (The search engine report, 1997), así como las limitaciones de los sistemas de recuperación a texto completo utilizados en la actualidad, imposibilita acceder de forma directa e instantánea a los documentos por campos concretos (autor, instituciones, materias...).

A esto hay que sumar que gran parte de motores de búsqueda usa métodos de ponderación poco eficaces – bien por las restricciones del software utilizado en la recuperación, bien por la pobre implementación de sus sistemas de indexación – lo que repercute en una baja tasa de eficiencia (Husby, 1997).

Por otra parte hay que reseñar la sobrecarga de tráfico en la Red, que causa, además del constante deambular de las arañas (Koster, 1995), la necesidad que los usuarios realicen gran cantidad de búsquedas en un mismo servicio hasta encontrar la información deseada (Elsen, 1998; NetGambit, 1999).

Una de las soluciones propuestas, los agentes inteligentes, no es capaz de resolver tampoco estos problemas. Para que uno de estos programas funcione correctamente debe generar una pequeña base de conocimiento del entorno que le viene dada por su propia experiencia (generada a partir de las peticiones del usuario y de la información que localiza en la Red) y por la de otros agentes que se comunican con él (Vargas-Quesada; Hípola, 1999).

Si bien es cierto que en cuestiones de recuperación de información los agentes inteligentes pueden resultar más precisos que los motores de búsqueda (Hípola; Vargas-Quesada; Montes Hernández, 1999), el hecho de que su productividad tenga una relación directamente proporcional con el tiempo que están vagando por la Red (con el fin de generar parte de la base de conocimiento antes mencionada) hace que su uso ralentice en exceso el resto de operaciones a realizar en ella. Sin hacer mención que, al igual que los servicios de búsqueda de Internet, los agentes más utilizados hasta ahora tampoco son capaces de realizar búsquedas por campos concretos en documentos a texto completo – ya que éstas se lanzan, en la mayoría de ocasiones, sobre las bases de datos de los buscadores –.

A tenor de lo expuesto aquí resulta evidente la necesidad de establecer mecanismos permitan una descripción más exhaustiva de los recursos electrónicos. En la actualidad se cuenta con sistemas de metadatos que pueden ser las herramientas que permitan la realización de estas descripciones. En este trabajo nos centraremos en el estudio del concepto de metadato con el fin de determinar tanto su campo de acción como los diferentes presupuestos subyacentes en este.

EL CONCEPTO DE METADATO

Los metadatos, en sí, no suponen algo completamente nuevo dentro del mundo bibliotecario. Según Howe (1993), el término fue acuñado por Jack Myers en la década de los 60 para describir conjuntos de datos. La primera acepción que se le dio (y actualmente la más extendida) fue la de dato sobre el dato, ya que proporcionaban la información mínima necesaria para identificar un recurso. En este mismo trabajo se afirma que *puede incluir información descriptiva sobre el contexto, calidad y condición o características del dato*. La evolución del término desde esta fecha hasta 1997 ha sido descrita por Lange y Winkler (1997) revelando que no existen demasiadas novedades.

Atendiendo a la definición antes mencionada, podríamos considerar la catalogación como un proceso de generación de metadatos. Teniendo en cuenta que la mayoría de sistemas de metadatos ha sido creada no sólo por profesionales de la información sino también por informáticos, diseñadores de programas, técnicos de sistemas, etc., la utilización de este término puede conllevar una carga excesiva (por ejemplo, reglas de catalogación, clasificaciones de materias...). El concepto de metadato se utiliza como un término neutral (Caplan, 1995), que permite alejarnos de posibles prejuicios por parte de todas aquellas personas menos cercanas al mundo bibliotecario, y que coloca a todos los grupos profesionales implicados en su desarrollo en una posición de igualdad.

Por otra parte, y si se analiza desde el punto de vista de la información distribuida, metadato, como concepto, aporta más información que el término catalogación, tal y como veremos más adelante.

Incidiendo sobre el aspecto básico de la definición, otros autores amplían el concepto de "dato sobre el dato" al afirmar que incluyen información sobre su contexto, contenido y control así como todo lo que tenga que ver con "el dato" (Pasquinelli, 1997).

En el informe de Biblink* (Heery, 1996) el metadato se define como *información sobre una publicación en oposición a su contenido. No sólo incluye descripción bibliográfica, sino que también contiene información relevante como materias, precio, condiciones de uso, etc.*

Ercegovac (1999), por su parte, afirma que un metadato describe los atributos de un recurso, teniendo en cuenta que el recurso puede consistir en un objeto bibliográfico, registros e inventarios archivísticos, objetos geoespaciales, recursos visuales y de museos o implementaciones de software. Aunque puedan presentar diferentes niveles de especificidad o estructura, el objetivo principal es el mismo: describir, identificar y definir un recurso para recuperar, filtrar, informar sobre condiciones de uso, autentificación y evaluación, preservación e interoperatividad.

En resumen, la mayoría de funciones descritas por estos autores las podemos encontrar agrupadas en el trabajo de Iannela y Waugh (1997):

• Resumir el significado de los datos

• Permitir la búsqueda

• Determinar si el dato es el que se necesita

• Prevenir ciertos usos (PICS**)

• Recuperar y usar una copia del dato

• Mostrar instrucciones de cómo interpretar un dato

• Obtener información sobre las condiciones de uso (derechos de autor)

• Aportar información acerca de la vida del dato

• Ofrecer información relativa al propietario/creador

• Indicar relaciones con otros recursos

• Controlar la gestión

El padre del Web, Tim Berners-Lee, se percató rápidamente de la importancia de los metadatos. Para él, su concepto no debía limitarse a la descripción de recursos Web. Más bien se debía ampliar, englobando las particularidades de gente, cosas, conceptos e ideas (Berners-Lee, 1997). Si bien es cierto que su definición es ambiciosa, Berners-Lee no contempló la posibilidad de extrapolar sistemas de metadatos a otros recursos electrónicos que no fueran Web*.

Para Berners-Lee existen tres tipos de metadatos en el Web:

• El primero de ellos es el que se encuentra dentro del documento mismo (por ejemplo aquellos que se pueden encontrar en cualquier documento generado por un procesador de textos).

• El segundo es el que se produce durante una transferencia HTTP (HyperText Transfer Protocol) – cliente y servidor se envían información sobre el objeto que están transmitiendo por medio de metadatos –.

• El último es más difícil de encontrar, ya que el metadato se utiliza cuando se consulta en otro documento (para comprobar si se puede acceder a él –o al sitio Web-, verificar derechos de autor...).

Este último caso es especialmente peculiar, ya que determina un papel "activo" por parte del metadato, y no "pasivo" (esperar a ser visto), como suele ser habitual. En realidad, este sistema de verificación se ha sustituido en la actualidad por otros mecanismos más precisos y complejos como pudieran ser las cookies o la realización de páginas HTML utilizando ASP (Active Server Pages) o cualquier otro lenguaje de programación (Visual Basic Script, JavaScript...).

Una de las ideas que se encuentra subyacente en la definición de Berners-Lee es la del trabajo con "objetos"**,

* Nombre del proyecto puesto en marcha por iniciativa de un grupo de bibliotecas nacionales europeas que tenía como principal objetivo el estudio del rol de las bibliografías nacionales en relación con las publicaciones electrónicas.

** Platform for Internet Content Selection. Mecanismo que utiliza metadatos para controlar el acceso a determinado tipo de páginas atendiendo a un sistema de clasificación previamente establecido.

* Nos estamos refiriendo a sistemas de metadatos como IAFA (Internet Anonymouys FTP Archive) utilizado para la descripción de ficheros en servidores FTP anónimos o SOIF (Summary Object Interchange Format) para el intercambio de descripciones de ficheros en la arquitectura Harvest.

** Entidad informativa que puede ser manipulada individualmente. Cabe que sea información "primaria" de cualquier tipo o información sobre otra información (metadatos).

tal y como se entiende en programación. Para Rosa (1999) muchos de los conceptos que pertenecen a la orientación a objetos existen desde hace mucho tiempo y se trabaja con ellos en Internet.

Apoyando esta teoría encontramos a Miller (1996), que incide en la necesidad de incluir el término "objeto" dentro de la definición al afirmar que *existen medatados para la mayoría de objetos o grupos de objetos concebibles, se almacenen en formato electrónico o no*; y a Husby (1997); quien presentó un trabajo en el congreso ELAG'97 con el que define los metadatos como atributos que describen un objeto. Estos objetos pueden ser documentos en papel, dentro de la Red o información de otro tipo. Otros autores que han desarrollado ideas similares son Hakala – que aportó el concepto de *documento como objeto*[*] – (1999) y Drewry (1997).

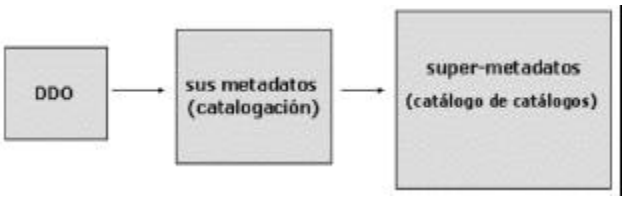
En estas definiciones podemos observar un salto cualitativo importante pues consideran que los documentos (así como sus partes: líneas, párrafos, imágenes...) se pueden tratar como objetos, y los metadatos como los atributos que definen las características de cada uno de ellos, sin limitarse a su descripción simple (lo que hasta ahora venía siendo la catalogación). Ésta es una de las ideas que se desprende del proyecto Desire (Dempsey; Heery, 1997).

Dentro del intento de teorizar sobre el concepto de metadato junto al de objeto, destaca la aportación de Chilvers y Feather (1998). Estos autores distinguen entre metadato y super-metadato. Este último se define como el dato asignado a cada DDO[**] que puede contener información del tipo:

• Nombre del sistema de metadatos utilizado en el DDO que permite su lectura.

• Cualquier información necesaria para gestionar el DDO que pueda no estar contenida dentro del metadato del DDO (expectativas de vida, cuándo será reemplazado por una próxima versión, etc.).

El esquema sería el siguiente:



En realidad no se trata más que de una estructura elaborada a partir de los actuales sistemas de repositorios de datos.

Hasta ahora, ninguna de las definiciones citadas ha entrado a describir objetivos o fines del uso de los metadatos. Cathro (1997) fue uno de los primeros en hacerlo al considerar que el metadato no sólo sirve para describir un recurso sino que, además, ayuda a acceder a un recurso informativo. Es muy importante esta aportación, ya que nos sirve para retomar la idea expuesta en el primer capítulo de este trabajo: la utilización de metadatos para mejorar la recuperación de la información en Internet.

Kerhevé y Gerbé (1997), que también comparten esta idea, afirman además que la utilización de estos sistemas facilita la gestión y el compartir grandes conjuntos de datos.

De todo lo expuesto hasta ahora podemos extraer varios puntos cruciales (dato sobre el dato, concepto de objeto, recuperación de información) que nos pueden ser útiles para la realización de una nueva definición que aglutine a todas las publicadas hasta la fecha, de tal forma que resulte posible concluir que metadato es *toda aquella información descriptiva sobre el contexto, calidad, condición o características de un recurso, dato u objeto que tiene la finalidad de facilitar su recuperación, autentificación, evaluación, preservación o interoperatividad*.

De esta forma, son ejemplos de metadatos:

• El encabezamiento de un fichero multimedia (imagen, vídeo o audio).

• El resumen de un documento.

• El catálogo de una base de datos.

• Los términos asignados haciendo uso de un tesauro.

• Las palabras extraídas de un texto.

[*] DLO (Document Like Object).

[**] Chilvers define Digital Data Object como cualquier recurso informático (páginas Web o revistas electrónicas) cuya información pueda ser almacenada y localizable independientemente de la forma en la que fue originalmente creada.

• Las fichas catalográficas en cualquier formato (ISBD, MARC...).

• Las páginas amarillas.

• Etc.

En Internet podemos encontrarlos también en multitud de formas:

• PICS

• Índices de documentos contenidos en una Intranet

• Direcciones IP o DNS

• Directorios X-500

• Encabezamiento de mensajes de correo electrónico

• Descripción de los archivos accesibles vía FTP

• Términos extraídos por los motores de indización/ búsqueda

• Etc.

De todos estos sistemas, sólo vamos a centrarnos en el estudio de aquellos que facilitan la:

• Identificación de documentos en un entorno distribuido

• Descripción de su contenido

• Localización y accesibilidad

• Gestión de derechos: copyright, reproducción, restricciones de acceso...,

ya que son los que más se acercan a la posible solución del problema planteado a lo largo del apartado 1 de este texto: el exceso de información en Internet y la dificultad de su localización y posterior recuperación.

Importancia de los metadatos

Tras lo expuesto podemos destacar varias razones que resaltan la importancia de los sistemas de metadatos:

• **Incrementan la accesibilidad**: la existencia de un conjunto de metadatos que describa correctamente uno o varios objetos aumenta la posibilidad de acceder a ellos (Gilliland-Swetland, 1998). Por otro lado, los metadatos hacen posible la búsqueda de información en múltiples colecciones a la vez. Por medio del mapeo entre sistemas heterogéneos es posible consultar, con una única ecuación de búsqueda, bases de datos que utilicen diferentes sistemas de metadatos para describir sus objetos.

• **Disminución del tráfico en la Red**: al indizar la representación del objeto, y no el objeto en sí, no requiere demasiado ancho de banda para hacer las búsquedas o generar los índices (Ortiz-Repiso, 1999).

• **Expandir el uso de la información**: ya que facilitan la difusión de versiones digitales de un único objeto.

• **Control de versiones**: no sólo en lo que se refiere a gestionar la vida de un objeto, sino también en lo que tiene que ver con su difusión, es decir: generar diferentes metadatos con distintas cantidades de información sobre un mismo objeto con el fin de distribuirla a un público heterogéneo.

• **Aspectos legales**: los metadatos permiten establecer claramente las restricciones de explotación, informar sobre los derechos de autor, control del uso de todo, o una parte, del objeto, método de pago por su disfrute, controlar el acceso a información restringida...

• **Preservación del objeto original**.

Tal y como afirman Milstead y Feldman (1999), las búsquedas a través del Web son, en la actualidad, un proceso de equiparación (matching) entre los términos de la consulta y los del documento. Si esa equiparación no se produce (bien sea por un problema en la forma de definir la petición, bien porque esa información sí se encuentra pero bajo otro concepto que lo describe), el documento no se recuperará. Para estas autoras la utilización de metadatos junto al uso de lenguajes controlados permitiría aumentar la precisión en la mayoría de búsquedas en Internet.

Instituciones implicadas en la introducción de metadatos

Siempre que se habla de metadatos, tarde o temprano, aparece la pregunta: ¿quién debería ser el responsable de introducirlos en los documentos electrónicos? Este interrogante nos lo encontramos especialmente en el entorno Internet. Si bien es cierto que el objetivo principal de nuestro trabajo no es el de aclarar esta cuestión no es menos cierto que el éxito de lo que se

propone aquí tiene una estrecha relación con una respuesta clara para esta pregunta.

A decir verdad, han sido pocos los autores que hayan propuesto soluciones suficientemente eficaces para este problema. Entre otras cosas, y muy probablemente, porque no se trata de un aspecto técnico o de un problema tecnológico sino de concienciación sobre la necesidad de las cosas. A estas alturas poca gente cuestiona que sean los bibliotecarios los que realicen las descripciones bibliográficas de los ejemplares que componen la colección en sus centros y, en realidad, lo que hacen es describir el contenido de fondos, es decir, introducir metadatos. Con esto no pretendemos señalar al bibliotecario como eje fundamental para que funcione un sistema de recuperación de información distribuida. Tan sólo apuntamos a la posibilidad de que la respuesta sea mucho más sencilla de lo que parece.

Tal y como señala Heery (1996), las organizaciones involucradas en la creación, mantenimiento y actualización de metadatos se pueden categorizar en:

1. Editores: del campo de la edición clásica y del mundo de la edición electrónica.

2. Servicios de información: agencias bibliográficas nacionales (en el entorno británico el British Library National Bibliographic Service), agencias bibliográficas comerciales (Whitaker), agencias que sirven resúmenes o servicios de indización (INSPEC), bases de datos de publicaciones periódicas (Blackwells, CARL).

3. Proveedores: de monografías (Dawson) o de publicaciones periódicas o seriadas (Swets, EBSCO).

4. Bibliotecas: por medio de agencias generadoras o gestoras de catálogos colectivos (OCLC – precisamente Dublin Core nace como iniciativa de este consorcio bibliotecario) y de cada una de las bibliotecas que preste especial atención a los recursos electrónicos (bibliotecas digitales).

Un aspecto que resulta evidente es que con el aumento de productos en formato electrónico, la generalización de la edición y la aparición constante de nuevas herramientas, la descripción de estos recursos por medio de metadatos es un hecho que debe tender a globalizarse.

Para Heery (1996), las organizaciones involucradas en esta propuesta se pueden considerar como "clásicas" dentro del mundo de la edición – ya sea ésta electrónica o no –. Con el aumento de productos en formato electrónico, la introducción de metadatos es fruto del trabajo de varias organizaciones. Además de estas categorías existe un nuevo grupo de entidades implicadas dentro del proceso de edición:

1. Autores: el ejemplo más claro lo tenemos en las páginas Web;

2. Servicios de búsqueda en Internet;

3. Servicios de archivos electrónicos: colecciones de materiales electrónicos como Oxford Text Archive, Essex Data Archive, Electronic Text Centre de la University of Virginia, Cervantes Virtual…;

4. Depósitos (repositorios) de colecciones de documentos: algo muy común dentro del mundo académico norteamericano (Los Alamos National Psysics Pre-print Archive);

5. Bibliotecas digitales.

De estos cinco grupos, quizá los que más interesen para los objetivos de este trabajo sean los dos primeros. En principio porque utilizan sistemas sencillos de metadatos centrados en técnicas muy básicas de descripción de contenido (Dublin Core, IAFA, PICS...) y, segundo, porque son los que más posibilidades presentan de interactuar con metalenguajes (SGML, XML...).

Metalenguajes

Varias tecnologías aplicadas al Web han expandido recientemente las posibilidades y capacidades de los metadatos, aumentando su riqueza en la descripción y facilitando el acceso al documento objeto. Estas herramientas suministran una mayor semántica y estructuración de los documentos, permitiendo más opciones de trabajo con los objetos (datos) y los metadatos (Hudgins; Agnew; Brown, 1999). Estas tecnologías son el SGML (Standard Generalized Markup Language) y XML (eXtensible Markup Language).

Si bien es cierto que SGML no se puede considerar como algo reciente – con sus orígenes en la década de los setenta, en 1986 se convirtió en ISO con el número 8879 – su utilización como sistema "incubador" de metadatos sí resulta novedoso. SGML es un metalenguaje que permite la creación de diferentes lenguajes de etiquetado a partir de una DTD (Document Type Definition). Las DTDs pueden convertirse en estándares para diferentes comunidades de usuarios. Esto es lo que ha sucedido con sistemas como TEI (Text Encoding

Initiative) para las humanidades y arte o EAD (Encoded Archival Description) para archivos. En el fondo ambos sistemas, de manera más o menos directa, se pueden considerar conjuntos de metadatos.

Algo parecido está sucediendo con XML. Este lenguaje, de reciente creación, es una versión abreviada de SGML. Su objetivo se centra en la posibilidad de intercambiar documentos (referenciales o a texto completo) estructurados a través del Web (Rosa; Senso, 1999). En realidad, lo que XML añade a HTML es la estructuración del documento sin detenerse sólo en la presentación. Con XML es posible establecer una estructura arbórea con todos los elementos que constituyen un documento para discriminar, rápidamente, los aspectos genéricos de los específicos. Este sistema de representación se ha revelado como vital para la generación automática de metadatos en diversos sistemas compatibles, como, por ejemplo, RDF (Resocurce Descritpion Format).

Tal es la integración que existe entre estos lenguajes de etiquetado con los sistemas de metadatos que los últimos modelos han sido elaborados utilizando su misma filosofía de trabajo:

• Al basarse en DTDs, la creación, modificación y gestión de metadatos es muy sencilla (especialmente si la comparamos con la lenta evolución que sufren sistemas más complejos como MARC).

• Como están integrados con lenguajes que permiten el tratamiento de cadenas de caracteres, es fácil automatizar procesos (incluido la introducción automática de metadatos en documentos).

• Aportan más posibilidades de trabajo. En la actualidad los metadatos pueden estar incluidos dentro del propio objeto (por ejemplo, dentro de la etiqueta HEAD del código HTML), en un documento aparte (e incluyendo una llamada tipo LINK del objeto al metadato y viceversa) o almacenados en repositorios con enlaces al objeto.

• Permiten, utilizando una DTD de puente, el intercambio de información entre bases de datos que estén elaboradas utilizando diferentes formatos.

A la par que estas tecnologías, se están desarrollando diferentes modelos de trabajo que permiten, utilizando estándares ampliamente reconocidos y valorados por los profesionales de la información – como puede ser la norma Z39.50 – junto a sistemas de metadatos – Dublin Core especialmente –, una organización más eficaz de las colecciones así como mayor efectividad en la recuperación de la información.

Tipología de los metadatos

Existe una gran variedad de formatos en la actualidad. Además, nos encontramos con que la mayoría de las bibliotecas digitales que utilizan metadatos para identificar sus objetos (bien mediante repositorios, bien dentro del mismo objeto) tienden a generar sus propios modelos (Hípola; Vargas-Quesada; Senso, 2000). Esto crea serios problemas a la hora de integrar estos sistemas dentro de un criterio común. Por este motivo, en el presente trabajo nos centraremos en aquellos formatos que son de dominio público y que más están siendo descritos por la comunidad científica.

Utilizando como base el criterio utilizado por Gilliland (1998), podemos considerar los siguientes tipos (Cuadro 1):

CUADRO 1
Tipologia de los metadatos

Tipo	Definición	Ejemplos
Administrativo	Usados en la gestión y administración de recursos de información	• Adquisición de información • Derechos y reproducción • Requerimientos legales para el acceso • Localización de información • Criterios de selección para la digitalización • Control de la versión
Descriptivo	Utilizados para representar recursos de información	• Registros catalográficos • Proporcionar ayuda en la búsqueda • Índices especializados • Hiperenlazar relaciones entre recursos • Anotaciones de los usuarios
Preservación	Para salvaguardar los recursos de información	• Informar sobre las condiciones de uso de los recursos físicos • Informar sobre las acciones llevadas a cabo para preservar versiones físicas y digitales de recursos
Técnico	Relativos a cómo funcionan los sistemas o el comportamiento de los metadatos	• Documentación de hardware y software • Digitalización de la información (formato, ratio de compresión...) • Autentificación y datos de seguridad (encriptación, passwords...) • Control de tiempo de respuesta de sistemas
Uso	Relativos al nivel y tipo de uso que se hace con los recursos informativos	• Información sobre versiones • Reutilización del contenido del recurso

Esta clasificación, a la que por supuesto no se le pueden asignar formatos definidos por ser demasiado general, no es excluyente. Es decir, que un sistema puede pertenecer a más de un tipo. Lo que tiene de importante es que, además, permite obtener una visión global de las diferentes acciones para las que se puede orientar el uso de metadatos.

Echamos en falta en esta clasificación una serie de criterios que consideramos clave, como pueden ser el método de creación o asignación (manual o automático), los protocolos con los que está asociado el metadato, o la complejidad y la riqueza en la descripción del recurso.

Un sistema clasificatorio que se acerca bastante a este modelo es el propuesto por Heery en el proyecto Biblink (la primera fila expresa la evolución en cuanto a complejidad de creación, y va del más sencillo al más difícil. La segunda se refiere al objetivo) (1996), como se presenta en el Cuadro 2.

CUADRO 2
Sistema classificatorio en el proyecto Biblink

Sencillo ----------> Complejo			
Localización	Selección	Evaluación	Análisis
Generados por un robot	Robot con ayuda del hombre	Introducción manual	Introducción y análisis de contenido manual
No estruturados	Valores acompañados de atributos	Cualificadores	Lenguaje de etiquetado estructurado (SGML)
Interface que conecta el formulario con el servicio http mediante CGI	Servicio de directorio (whois++) con enrutamiento mediante CIP (Common Indexing Protocol)	Z39.50	Nuevas versiones de Z39.50 (con SQL, texto completo, browsing...)
Propietario	Apareciendo en la actualidad	Normas genéricas utilizadas actualmente en el mundo bibliotecario	Normas utilizadas en temáticas muy especializadas

CUADRO 3
Esquema propuesto por Dempsey y Heery

	Columna uno	Columna dos	Columna tres
Características de los registros	Formatos simples	Formatos estructurados	Formatos ricos
	Propietarios	Nuevos formatos	Normas internacionales
	Indexación de texto completo	Estructura de campos	Etiquetado elaborado
Formatos	Lycos Altavista Yahoo, etc.	Dublin Core IAFA RFC 1807 SOIF NetFirst	ICPSR CIDI EAD TEI MARC

Debemos tener en cuenta que esta estructuración se engloba dentro de un proyecto europeo que tiene como finalidad la búsqueda del modelo a seguir por Bibliotecas Nacionales con vistas al tratamiento electrónico de registros. No obstante, este sistema avanza un poco más en los puntos antes mencionados. Al mismo tiempo, presenta el problema de ser excesivamente ambiguo y, por lo tanto, mezclar sistemas de metadatos simples con otros complejos.

Basándose en este esquema clasificatorio, Smits (1996) realizó una modificación para crear una tipología de metadatos para cartografía e información espacial. Por este motivo no hemos considerado oportuno el incluirlo en el presente apartado. No obstante, la clasificación descrita presenta diferentes deficiencias como, por ejemplo, considerar Dublin Core como un modelo únicamente válido para Internet.

A pesar de ser sustancialmente más simple, el esquema propuesto por Dempsey y Heery (1997) es el que más se acerca al sistema clasificatorio necesario para este trabajo (Cuadro 3).

• La Columna uno incluye datos relativamente estructurados cuya recuperación suele ser automática. En la mayoría de los casos, se trata de información con una semántica reducida y que no permite la búsqueda por campos, es decir, todos aquellos datos que son generados por robots (sistemas actuales). El hecho de que los recursos no estén indexados de forma apropiada hace que el usuario pueda perder información relevante (como de hecho así ocurre).

• El segundo grupo está compuesto por todos aquellos sistemas que contienen una descripción lo suficientemente clarificadora como para que el usuario pueda acceder fácilmente al recurso. Además, el hecho de almacenar la información en campos agiliza la búsqueda. Una de las características clave de este segmento es que su introducción no tiene por qué

corresponder a especialistas (salvo, claro está, el formato SOIF que es generado automáticamente por el gather dentro de Harvest).

• Para finalizar, el tercer conjunto está formado por todos aquellos formatos que contienen un alto grado de descripción y, por tanto, de complejidad en lo que se refiere a su creación. Es tal su nivel de especificidad que en la mayoría de los casos no sólo son válidos para la localización y recuperación de información sino que, además, son el complemento ideal para la descripción total de conjuntos de objetos.

Junto a esta clasificación, podemos observar los siguientes atributos y características propias de los metadatos (Baca, 1998) (Cuadro 4):

CUADRO 4
Atributos y características de los metadatos

Atributo	Características	Ejemplos
Fuente	Metadatos internos generados por el agente creador con el propósito de informar sobre el momento de su creación	Nombres de ficheros Estructuras de directorios Formatos de ficheros y algoritmos de compresión
	Metadatos externos relativos a una información que se modifica después de su creación	Registros catalográficos Información sobre derechos de autor
Método de creación	Metadato generado automáticamente por un ordenador	Índices de palabras clave Logs Weblogs y bitácoras
	Metadatos creados manualmente	Herramientas descriptivas
Naturaleza	Creados por el autor del documento objeto	Los utilizados en páginas HTML
	Generados por profesionales de la información, independientemente de quién sea el autor del documento objeto	Registros MARC Encabezamientos de materia
Estado	Estático: no cambian desde su creación	Título, fecha de creación
	Dinámico: varía con el uso del documento objeto	Estructuras de directorios Logs
	A largo plazo: necesario para asegurarse de que el documento objeto será accesible en todo momento	Información de los derechos (de autor, de uso, de difusión...)
	A corto plazo: con clara vocación transaccional	Información sobre el uso
Estructura	Con estructura basada en estándares	MARC TEI AACR2
	Sin estructura predecible	Metadatos ad hoc (la mayoría de los generados en y para bibliotecas digitales)
Semántica	Normalizados por medio de un vocabulario controlado	MARC AACR2
	No controlados	Etiquetas HTML
Nivel	Colecciones de metadatos relativos a colecciones de documentos objeto	MARC Índices especializados
	Un metadato relativo a un documento objeto individual, fuera de cualquier colección	Información sobre el formato Leyenda de una imagen

Sistemas de metadatos

En la actualidad existen numerosos sistemas que se están implementando en gran cantidad de proyectos. Dado que es prácticamente imposible recogerlos todos, nos centraremos en aquellos que afectan directamente al procesamiento de la información, los que tengan un uso más extendido y los que, además, satisfacen los siguientes requerimientos:

• Identificación de documentos en un entorno distribuido.

• Descripción de su contenido.

• Localización y accesibilidad.

• Gestión de derechos: copyright, reproducción, restricciones de acceso...

Entre ellos destacan:

Los aceptados por la norma HTML.

DC o DCMI (Dublin Core Metadata Initiative).

RDF (Resource Description Framework).

TEI (Text Encoding Initiative).

MARC DTD (Machine Readable Cataloging Document Type Definition).

EAD (Encoded Archival Description).

PICS (Platform for Internet Content Selection).

MCF (Meta Content Format).

IAFA (Internet Anonymous FTP Archive).

SOIF (Summary Object Interchange Format).

La mayoría de estos sistemas se utilizan de forma aislada, ya que su objetivo es, fundamentalmente, satisfacer unos requerimientos muy específicos (EAD para descripción de documentos de archivos, SOIF e IAFA como ficheros de intercambio en sistemas de indización distribuida, PICS para permitir o no el acceso a determinados contenidos...).

Mención aparte merece RDF que, gracias a su orientación, permite la inclusión de otros sistemas de metadatos para favorecer el intercambio de información entre bases de datos heterogéneas.

CONCLUSIONES

Resulta evidente que las estructuras de metadatos están adquiriendo una posición preponderante en lo que se refiere a la descripción de recursos electrónicos entendidos como objetos. Cada vez son más numerosos los proyectos, sitios Web o sistemas de consulta que se valen de ellos para lograr mejores prestaciones a la hora de la representación, localización y recuperación de recursos electrónicos.

Al contrario de lo que sucede con formatos más complejos y menos flexibles como TEI, el sistema más extendido en la actualidad – Dublin Core Metadata Initiative – pone más énfasis en facilitar al máximo el acceso al recurso y menos en proporcionar una descripción exhaustiva del mismo. Esto resulta vital, ya que ha sido un fallo tradicional en los catálogos bibliotecarios en los que, por el contrario, se hace más hincapié en la descripción que en dotar a los registros de más y mejores elementos de recuperación.

La mayoría de sistemas de metadatos ofrecen la solución técnica necesaria para realizar una descripción homogénea y estricta de los recursos sin necesidad de limitar las opciones de localización y recuperación. Al mismo tiempo es posible utilizar la mayoría de conjuntos de metadatos actuales junto a cualquier lenguaje de marcas derivado del SGML, lo que les aporta la característica de multiplataforma que los convierte en la herramienta ideal para crear un entorno de información integrada en el que el catálogo proporcione acceso tanto a los documentos tradicionales como a la información electrónica.

Junto a esto, la posibilidad de incluir información referida a la calidad, condición o características del recurso aporta un valor añadido inestimable que los actuales sistemas no pueden ofrecer.

El uso de conjuntos de metadatos que faciliten la interoperatividad entre diversas bases de datos (como puede ser el uso de RDF y DC), la utilización de lenguajes de etiquetado más manejables que el SGML y menos simples que el HTML (XML), la aplicación de protocolos pensados para la recuperación de información (Z39.50), la aplicación de técnicas de recuperación de información para generar servicios determinados (DSI), así como el desarrollo del Web, hacen vislumbrar un futuro halagüeño a los metadatos.

En palabras de Duval (Chen y Chang, 1998), *los metadatos son parte de la infraestructura de la información necesaria para ayudar a crear orden en el caos del Web, proporcionando descripción, clasificación y organización.*

Artigo recebido em 17-03-2003 e aceito para publicação em 02-04-2003

REFERÊNCIAS

META attributes by count. Disponível em: <http://vancouver-webpages.com/META/bycount.shtml>. Acesso em: fev. 2003.

BACA, Murtha. *Introduction to metadata*: pathways to digital information. Los Angeles : Getty Information Institute, 1998.

BERNERS-LEE, Tim. *Metadata architecture*: documents, metadata and link. Disponível em: <http://www.w3.org/DesignIssues/Metadata.html>. Acesso em: fev. 2003..

CAPLAN, P. You call it corn, we call it syntax-independent metadata for document-like objects. *The Public Access Computer Systems Review*, v. 4, n. 6, 1995.

CATHRO, W. *Metadata*: an overview. Disponível em: <http://www.nla.gov.au/nla/staffpaper/cathro3.html>. Acesso em: fev. 2003.

CHEN, H. H.; CHANG, Y. S. The role of metadata in national taiwan university digital library / museum project. *Journal of Library and Information Science*, v. 23, n. 2, p. 51-65, 1998.

CHILVERS, A.; FEATHER, J. The management of digital data: a metadata approach. *Electronic Library*, v. 16, n. 5, p. 335-371, 1998.

DEMPSEY, L.; HEERY, R. Desire: development of an European Service for Information on Research and Education. Unión Europea, 1997.

DREWRY, M. *et al.* Metadata: quality vs. quantity. *In*: *IEEE METADATA CONFERENCE, 2., 1997*. [S. l.] : IEEE, 1997.

ELSEN, J. Portals will open web's doors to masses. *New York Post*, 18 enero 1998.

ERCEGOVAC, Z. Introduction. *Journal of the American Society for Information Science*, v. 50, n. 13, p. 1165-1168, 1999.

GILL, T. Metadata and the World Wide Web. *In*:_____. *Introduction to metadata*: pathways to digital information. Los Angeles : Getty Information Institute, 1998, p. 9-18.

GILLILAND-SWETLAND, A. J. Defining metadata. *In*:_____. Introduction to metadata: pathways to digital information. Los Angeles: Getty Information Institute, 1998, p. 1-8.

HAKALA, J. Internet metadata and library cataloguing. *ICBC*, v. 28, n. 1, p. 21-5, 1999.

HARVEST. 2000. Disponível em: <http://www.searchtools.com/tools/harvest.html>. Acesso em: fev. 2003.

HEERY, R. *Biblink*: LB4034 D1.1 metadata formats. [S. l.] : Biblink, 1996.

HÍPOLA, P.; VARGAS-QUESADA, B.; MONTES HERNÁNDEZ, A. Descripción y evaluación de agentes multibuscadores. *El Profesional de la Información*, v. 8, n. 11, p. 15-26, 1999.

_____;_____; SENSO, J. Bibliotecas digitales: situación actual y problemas. *El Profesional de la Información*, v. 9, n. 4, p. 4-13, 2000.

HOWE, D. Free on-line dictionary of computing. Disponível em: <http://wombat.doc.ic.ac.uk/foldoc/>. Acesso em: fev. 2003.

HUDGINS, J.; AGNEW, G.; BROWN, E. *Library and Information Technology Association*: getting mileage out of metadata applications for the library. Chicago : American Library Association, 1999.

HUSBY, O. *Metadata*: Elag'97. [S. l. : s. n.], 1997.

IANNELA, R.; WAUGH, A. *Metadata*: enabling the Internet. Disponível em: <http://archive.dstc.edu.au/RDU/reports/CAUSE97>. Acesso em: fev. 2003.

KERHERVÉ, B.; GERBÉ, O. Models for metadata or metamodels for data? *In*: *IEEE METADATA CONFERENCE, 2.*, 1997. [S. l.], 1997.

KOSTER, M. *Robots in the web*. Disponível em: <http://www.robotstxt.org/wc/threat-or-treat.html>. Acesso em: fev. 2003.

LANGE, H R.; WINKLER, B J. Taming the Internet: metadata, a work in progress. *Advances in Librarianship*, v. 21, p. 47-72, 1997.

MILLER, P. Metadata for the mases. *Ariadne*, n. 5, 1996.

MILSTEAD, J.; FELDMAN, S. Metadata: cataloguing by any other name. *Online*, n. 1, p. 25-31, 1999.

MONTES HERNÁNDEZ, A. Posibilidades de consulta de los buscadores. *El Profesional de la Información*, v. 8, n. 3, p. 8-14, 1999.

NETGAMBIT. Search engines generate traffic. Disponível em: <http://www.nua.net/surveys/?f=VS&art_id=868880518&rel=true>. Acesso em: fev. 2003.

OLVERA LOBO, M. D. Rendimiento de los sistemas de recuperación de información en la Wold Wide Web: revisión metodológica. *Revista Española de Documentación Científica*, v. 23, n. 1, p. 63-77, 2000.

ORTIZ-REPISO, JIMÉNEZ, V. Nuevas perspectivas para la catalogación: metadatos versus MARC. *Revista Española de Documentación Científica*, v. 22, n. 2, p. 198-219, 1999.

PASQUINELLI, A. I*nformation technology directions in libraries*: a sun microsystems white paper. Disponível em: <http://wwwsun.com/products-n-solutions/edu/libraries/libtechdirection.html>. Acesso em: fev. 2003.

ROSA, A. XML orientado a objetos. *El Profesional de la Información*, v. 8, n. 9, p. 4-23, 1999.

_____; SENSO, J. XML como medio de normalización y desarrollo documental. *Revista Española de Documentación Científica*, v. 22, n. 4, p. 488-504, 1999.

SMITS, J. Digital metadata, standards for communication and preservation. *European Research Libraries Cooperation*, v. 6, n. 4, p. 83-406, 1996.

THE SEARCH ENGINE REPORT. The new meta tag are coming - or are they? Disponível em: <http://searchenginewatch.internet.com/sereport/97/12-metatags.html>. Acesso em: fev. 2003.

VARGAS-QUESADA, B.; HÍPOLA, P. Agentes inteligentes: definición y tipologia. Los agentes de información. *El Profesional de la Información*, v. 8, n. 3, p. 13-21, 1999.

WEBTOP. Disponível em: <http://www.webtop.com/search/vanilla/press190800.htm>. Acesso em: fev. 2003.