

## Artículo de revisión

### Organización y recuperación de información en Internet: teoría de los metadatos

MSc. Isabel Daudinot Founier<sup>1</sup>

#### RESUMEN

El desarrollo de la teoría de los metadatos alcanza mayor auge con el avance de Internet y la necesidad de describir los recursos de información digitales con vista a su localización y posterior recuperación. Se abordan aspectos relacionados con los recursos digitales y sus características, los metadatos, su definición, tipos y funciones; así como los tipos de metadatos más frecuentes en HTML y su relación con el lenguaje controlado.

*Palabras clave:* Recursos digitales, metadatos, vocabularios controlados, HTML.

#### ABSTRACT

The development of the metadata theory speeds up point with the advancement of Internet and the need to describe the resources of digital information in order to facilitate their location and further retrieval. Aspects related to the digital resources and their characteristics, the metadata, their definition, types and functions; as well as the most frequent types of metadata in HTML and their relation with the controlled language, are dealt with.

*Key words:* Digital resources, metadata, controlled vocabularies, HTML.

Copyright: © ECIMED. Contribución de acceso abierto, distribuida bajo los términos de la Licencia Creative Commons Reconocimiento-No Comercial-Compartir Igual 2.0, que permite consultar, reproducir, distribuir, comunicar públicamente y utilizar los resultados del trabajo en la práctica, así como todos sus derivados, sin propósitos comerciales y con licencia idéntica, siempre que se cite adecuadamente el autor o los autores y su fuente original.

Cita (Vancouver): Daudinot Founier I. Organización y recuperación de información en Internet: teoría de los metadatos. Acimed 2006;14(5). Disponible en: [http://bvs.sld.cu/revistas/aci/vol14\\_5\\_06/aci02506.htm](http://bvs.sld.cu/revistas/aci/vol14_5_06/aci02506.htm) [Consultado: día/mes/año].

El surgimiento y desarrollo de Internet ha modificado totalmente la experiencia acumulada durante años de trabajo en los aspectos más importantes de la creación, distribución, almacenamiento, modo de acceso y representación de los recursos de información. El desarrollo de las tecnologías de la información y las comunicaciones ha hecho posible que cualquiera pueda crear y publicar en la red, lo que ha provocado un crecimiento incontrolable de los recursos de información existentes en Internet, y cuando paradójicamente se habla de *acceso global a la información* o de *la era de la información*, los usuarios se enfrentan a una red caótica, carente del orden necesario para poder recuperar la información pertinente. Entonces, ellos se ven en la necesidad de utilizar los llamados buscadores de información existentes en Internet, los cuales poco pueden también hacer para ayudar a los usuarios a localizar información relevante como consecuencia, en gran medida, de la falta de tratamiento documental existente en la red.

Los recursos de información se almacenan en la red por sus creadores sin ningún tipo de descripción que permita su organización y posterior recuperación, y esto es posiblemente uno de los puntos más débiles de Internet. Millones de recursos de

información no pueden recuperarse al no contar con una adecuada descripción que permita a los buscadores de información su localización y recuperación.

El uso normalizado de metadatos se presenta como una solución para ayudar a los creadores de los recursos de información a asociar a ellos descripciones susceptibles de procesarse por los buscadores de información.

### **Los recursos digitales y sus características**

El desarrollo precipitado de las tecnologías de la información y las comunicaciones ha generado cambios revolucionarios en casi todas las esferas del quehacer humano. La bibliotecología y todas las ramas relacionadas con el manejo y uso de la información, por supuesto, no escaparon a esta realidad. En los últimos años, las bibliotecas experimentan una necesaria etapa de transformación y adaptación a las nuevas formas de procesamiento y manejo en sentido general de la información.

Durante siglos, las bibliotecas conformaron sus fondos a partir de la colección de objetos físicos, que debían adquirirse de alguna manera, y sobre los que la institución ejercía luego derechos de propiedad. Las colecciones eran entonces un conjunto de objetos físicos portadores de información, albergados en una localización física precisa, y patrimonio exclusivo de una institución. Los bibliotecarios desarrollaron, a lo largo del tiempo, criterios y políticas para la selección y procesamiento de aquellos materiales que se correspondían con los objetivos de su institución.<sup>1</sup>

Internet produjo un cambio total de paradigma en relación con la selección, procesamiento y recuperación de la información. Muchas bibliotecas han dejado de ser una institución social cuya función principal era la conservación o preservación de los documentos almacenados en sus fondos, y no se trata solo del hecho de que en Internet los recursos son digitales, se trata también de que muchos son de acceso gratuito, y eso elimina uno de los aspectos que influía de manera significativa en el desarrollo de las colecciones de las bibliotecas tradicionales: la disponibilidad de recursos para la adquisición. En Internet, los recursos de información están al alcance de todos, siempre y cuando seamos capaces de desarrollar mecanismos que permitan su localización; el problema radica en cómo organizar la información en Internet para garantizar su localización y recuperación exitosa.

La facilidad para crear y publicar recursos de información que han permitido las nuevas tecnologías ha generado un crecimiento sin precedente en la cantidad y variedad de recursos de información disponibles en la red; esto, unido al hecho de que las redes de comunicación han permitido que los contenidos informativos puedan transmitirse en formato digital sin límites geográficos para el acceso, produce un cambio total en el quehacer de las bibliotecas y los profesionales de la información.

Algunos han llegado a vaticinar la desaparición de las funciones de las bibliotecas y de sus profesionales, pero como expresa *Bermello Crespo*, las bibliotecas han utilizado eficientemente los desarrollos tecnológicos de su época para elevar la calidad de sus servicios. El trabajo de los profesionales de la información ha estado marcado por los mecanismos y dispositivos disponibles para garantizar sus funciones bibliotecarias, pero nunca la tecnología ha determinado los objetivos que persigue la actividad bibliotecaria.<sup>2</sup>

Las bibliotecas sobrevivirán y seguirán con sus nombres y funciones tradicionales: seleccionar, organizar, conservar y proporcionar acceso a los registros del conocimiento humano en todas las formas, los bibliotecarios deben introducir los cambios tecnológicos que acepte la comunidad a la que sirven y a las que necesariamente hay que educar y dirigir por nuevos rumbos, pero esto tiene que hacerse de manera razonable, tanto en su realidad como en la percepción de las cosas. Las bibliotecas sirven a la enseñanza y a la investigación, en la medida en que estas funciones cambien, las bibliotecas cambiarán también.<sup>3</sup>

Uno de los mayores retos a los que se enfrentan los bibliotecarios actualmente es el procesamiento de la información en la red. Para asumir este reto, es necesario que

conozcan las características o particularidades de este tipo de información en formato digital.

*Heery y Burnett* agrupan las características o particularidades de la información en formato digital en los siguientes aspectos:<sup>4,5</sup>

- *Organización*: En los recursos digitales, la organización no es centralizada al estilo de un catálogo de biblioteca ni se organizan en tablas por atributos al estilo de los sistemas más conocidos en gestión de datos informáticos; tampoco se seleccionan ni se coleccionan. Mientras los recursos bibliotecarios se seleccionan y adquieren, los archivos de datos se solicitan y adquieren.
- *Ubicación*: Poseen localización remota, frecuentemente de ninguna manera asociada a la institución, como es el caso de un registro de catálogo. Deben considerarse detalles como el modo de acceso disponible; por ejemplo, FTP, HTTP y restricciones a este, como contraseñas. Un mismo recurso puede residir en varios sitios diferentes.
- *Versiónes*: El mismo recurso puede existir en diferentes formatos, por ejemplo, Postscript, ASCII, etcétera.
- *Inestabilidad*: Los datos tienen, a menudo, vida corta en Internet. Los archivos sufren traslados constantes entre distintos servidores y la dirección electrónica (URL) original deviene obsoleta. Los creadores cambian y desarrollan documentos en una URL existente, de forma que muchos recursos en Internet se encuentran bajo elaboración. Son móviles e inestables comparados con los recursos de naturaleza fija, estables, propios de la biblioteca e inseguros, a diferencia de los archivos de datos.
- *Redundancia*: Las versiones viejas de documentos no se eliminan, o la información deviene obsoleta. Algunos recursos contienen datos de mantenimiento y revisión que brindan al lector información sobre su actualización.
- *Complejidad*: Un sitio en Internet ¿es un documento o pueden ser varios? La existencia de interrelación entre páginas Web es otro signo a considerar en este punto. Una de las características de Internet es la información hipertextual. Los documentos se relacionan estrechamente mediante enlaces y muchas veces es difícil determinar los límites de un documento o de un sitio.

Los procedimientos y herramientas que tradicionalmente han utilizado las bibliotecas para el procesamiento de sus colecciones en formato impreso, hoy no son adecuados para el tratamiento de los recursos de información en formato digital en Internet. Es preciso considerar que las dimensiones del conjunto de recursos de información en Internet son tan grandes, que impide a cualquier biblioteca abarcarlo todo y menos aún poder realizar su tratamiento documental.

Diariamente, en Internet aparecen miles de sitios, otros desaparecen, se mudan de dirección o modifican su contenido, para cualquier biblioteca tratar de controlar este fenómeno sería un reto inalcanzable. Las bibliotecas en Internet tienen necesariamente que seleccionar los recursos de información para sus usuarios en dependencia de sus necesidades en este sentido, tarea que no es fácil, por el caos que reina en la red.

Todas las organizaciones involucradas en la generación, recuperación y uso de los documentos digitales reconocen la necesidad de establecer normas que permitan estructurar su contenido y contribuyan a facilitar el proceso de búsqueda y recuperación de información. Esto, por supuesto, justifica la necesidad de que la información alojada en Internet disponga de alguna manera de un tipo de procesamiento documental que garantice su localización y recuperación exitosa; de lo contrario, se corre el riesgo de no poder acceder a mucha información valiosa y lo que es peor, desconocer totalmente su existencia.

Cuando se habla de procesamiento documental en Internet, se habla de metadatos, que surgen como respuesta a la necesidad de contar con una estructura adecuada para la descripción normalizada de documentos digitales a fin de posibilitar la localización y recuperación selectiva de la información en la red.

## **Metadatos**

*Definición:* El término metadatos está compuesto por el prefijo inseparable de origen griego *meta* que, según el Diccionario de la Lengua Española significa "junto a", "después de", "entre" o "con" y el término de origen latino *datum* (lo que se da), que según esta misma fuente tiene dos acepciones significativas: la primera: "antecedente necesario para llegar al conocimiento exacto de algo o para deducir las consecuencias legítimas de un hecho"; y la segunda: "información dispuesta de manera adecuada para su tratamiento por un ordenador".<sup>6</sup>

Si se analiza etimológicamente la palabra metadatos, puede afirmarse que significa algo más que datos, es decir, son datos preliminares sobre los recursos de información. En términos precisos, puede establecerse que los metadatos constituyen información sobre los datos o como frecuentemente se le denomina: datos sobre datos.

Independientemente de que los bibliotecarios describen fuentes de información desde que las bibliotecas comenzaron el registro de los datos de sus colecciones, el término metadatos comenzó a aparecer con cierta frecuencia en la década de los años 80 en la literatura sobre sistemas de gestión de bases de datos. Al respecto Méndez precisa que la primera publicación en la que encontró el término metadata en el título, en la base de datos LISA, es de 1982, y define los metadatos, en el contexto de los sistemas de gestión de bases de datos, de la siguiente manera:<sup>7</sup>

"Los metadatos constituyen información sobre datos numéricos [...] un sistema de metadatos funciona como un equivalente paralelo a los sistemas de datos numéricos de forma que los analistas, los que toman decisiones, los que resuelven problemas y los gestores de sistemas, aprendan lo suficiente sobre los datos numéricos para conocer su probabilidad de uso válido y apropiado".<sup>8</sup>

A mediados de la década de los años 90, el término metadatos comienza a aceptarse por todas las disciplinas relacionadas con la organización de la información en Internet. Al principio, el término se relacionaba con la gestión e interoperabilidad de datos geoespaciales y el mantenimiento de sistemas de gestión de datos; se refería a un conjunto de normas industriales con una documentación adicional interna y externa, y datos necesarios para la identificación, representación, interoperabilidad, gestión técnica y uso de los datos alojados en los sistemas de información.

En la medida que Internet comenzó a crecer vertiginosamente, los metadatos adquirieron mayor relevancia y se convirtieron en un término presente para todos en la red como consecuencia de la necesidad de normalizar el procesamiento de la información en formato digital dispersa.

Los metadatos son datos secundarios como pueden ser el autor, el título, las palabras clave, el resumen, la fecha, u otros que describen los datos primarios o recursos de información, es decir, se emplean para suministrar información sobre datos producidos, ellos describen el contenido y otras características de los datos primarios para posibilitar a una persona o máquina ubicar y entender los datos.

Uno de los ejemplos más ilustrativos para comprender el significado del término metadatos son precisamente los conocidos catálogos de las bibliotecas, esta es la razón por la que algunos autores han llegado a plantear que metadatos es exactamente lo que hasta el momento se denominaba registros bibliográficos o descripciones bibliográficas. Estas opiniones son, en parte, ciertas porque si los metadatos son datos secundarios que sirven para describir los recursos de información, a todos los productos de una actividad catalográfica podría llamársele metadatos. Sin embargo, es necesario aclarar que el concepto estricto de metadatos surge en la red y presenta muchas más connotaciones.

Las descripciones bibliográficas pueden calificarse como un tipo de metadatos si se analizan como un producto de la actividad catalográfica de las obras impresas en sentido general, que cumplen una función determinada en un contexto determinado. *Gradmann*, al referirse a los metadatos plantea: "el contexto general de producción y uso de esta información es substancialmente diferente y está impulsada para ir más allá del paradigma tradicional de la catalogación. Considerar que el proceso de

creación de metadatos es un tipo de catalogación simplificada, sería probablemente un error importante".<sup>9</sup>

Independientemente de que tanto los registros catalográficos como los metadatos presentan la función común de describir los recursos de información para garantizar su posterior recuperación, las diferencias entre uno y otro son evidentes y significativas. *Graciela Spedalieri* señala algunas diferencias básicas entre el registro bibliográfico tradicional y los metadatos:<sup>1</sup>

- El registro bibliográfico se crea generalmente con posterioridad a la creación del documento, y es físicamente independiente, mientras que un registro de metadatos se crea junto con el documento y forma parte de este. De esta manera, el recurso en Internet puede llevar consigo una estructura de datos que facilite su recuperación.
- El registro bibliográfico se crea según normas bastante complejas, y por eso requiere personal altamente calificado (catalogadores), mientras que los metadatos se han diseñado con la idea de que el creador del documento pueda crear el registro de metadatos sin los conocimientos especializados que precisa la catalogación para crear un registro.
- Por la complejidad de las normas que rigen la creación de los registros bibliográficos, la tarea requiere mucho tiempo, mientras que la simplicidad de los metadatos lleva a que la creación de registros sea mucho más rápida.
- La complejidad de las normas tradicionales proviene en parte de la existencia de lo que se denomina control de autoridad: el proceso de asegurar la consistencia de aquellos datos del registro que se utilizan como puntos de acceso. Los metadatos carecen de este elemento. Los registros bibliográficos tradicionales se apoyan, además, en normas de descripción y codificación bastante detalladas que aseguran una mayor uniformidad en los datos.

Los metadatos también difieren de las tradicionales fichas catalográficas en que la información sobre la localización se sitúa en el propio registro para permitir la recuperación directa de la información a partir de la aplicación de un software apropiado.<sup>4</sup>

Para *Rosa San Segundo*, a diferencia de las fichas catalográficas, una de las características más importantes de los metadatos es su capacidad de relación o de establecer enlaces. De esta forma, se han hecho imprescindibles en la recuperación global de la información en Internet, porque se trata de indizar y clasificar inmensas cantidades de información de diversos tipos.<sup>10</sup>

Los metadatos no solo se relacionan con la descripción de un objeto de información, Aunque muchos especialistas asocien este término preferentemente con la descripción y catalogación, ellos pueden también referirse al contexto, la gestión, la conservación y el uso de los recursos de información. Por esta razón, puede afirmarse que los metadatos pueden ser algo más que simples elementos descriptivos y es la razón por la que, cada vez, se incorporan nuevas esferas a su conceptualización.

A continuación, se exponen algunas de las definiciones de metadatos que ofrecen diferentes autores:

- Es el conjunto de elementos que pueden utilizarse para describir y representar objetos de información.<sup>11</sup>
- Los metadatos son datos que se asocian con objetos que liberan a sus usuarios potenciales [personas o programas] de la necesidad de tener un conocimiento avanzado y completo sobre su existencia o características [...]. Los metadatos constituyen un conocimiento que permite a los usuarios, humanos o automatizados, comportarse de manera inteligente.<sup>12</sup>
- Los metadatos en el contexto de Internet, son informaciones estructuradas sobre la información distribuida, todos aquellos datos comprensibles por la computadora, estructurados de forma que sirven para localizar, identificar y describir el contenido de un documento web, y por tanto, para recuperar

información en la red.<sup>13</sup>

- Los metadatos son información documentada por medio de herramientas de tecnologías de la información que mejoran la comprensión, tanto técnica como comercial, de los datos y de los procesos relacionados con ellos.<sup>14</sup>
- Metadatos es una información que entiende la computadora sobre recursos Web y otras cosas.<sup>15</sup>
- Esta definición se contextualiza en la información manipulada por productores de software, diferentes de los usuarios tradicionales de bibliotecas. Su objetivo es la eficacia y celeridad de su funcionamiento, en comparación con la catalogación tradicional.<sup>9</sup>

*Burnett* describe los metadatos desde el punto de vista bibliotecológico como:<sup>5</sup>

- Cualquier información que registra la caracterización y relaciones de los datos fuente, o el conjunto de elementos de datos que pueden utilizarse para describir y representar objetos de información.

Desde el punto de vista informático, esta misma autora los define como “Cualquier información que soporta la efectiva utilización de datos, incluida la información que pueda facilitar su gestión, acceso y análisis”.

Es importante señalar que para diferentes autores y en dependencia del uso que realicen de estos, los metadatos se definen de formas diferentes, incluso un mismo autor los define desde diferentes puntos de vista. Resulta muy difícil encontrar una definición única de metadatos, cada definición hasta el momento está en dependencia de varios factores: quién lo utiliza, para qué lo utiliza y en qué contexto lo utiliza.

La mayoría de las definiciones de metadatos en el contexto de la red consideran tres aspectos fundamentales:<sup>7</sup>

- La funcionalidad, la función de describir e identificar datos primarios o recursos de información.
- El contexto, la red.
- El interlocutor, las personas o máquina.

Entonces, puede asumirse que el término metadatos en el contexto de Internet, se refiere a cualquier dato que ayuda, tanto a las personas como a las máquinas o computadoras, a la identificación, descripción y localización de los recursos electrónicos existentes en la red.

Los metadatos son datos secundarios debidamente estructurados, que sirven para describir los recursos de información existentes en Internet, cuyo objetivo es ayudar en su identificación y ulterior localización, tanto por parte de las personas como de las computadoras.

Tipos:

Son múltiples los tipos y funciones de los metadatos existentes actualmente, porque estos dependen de muchos factores, entre los cuales pueden citarse: el tipo de información que describen, el nivel de estructuración de esta información, el lugar donde se encuentren los metadatos, su ámbito de aplicación, el tipo de usuarios que los utiliza y sus finalidades, entre otros. Por tanto, son múltiples también las clasificaciones de metadatos realizadas hasta ahora por diferentes autores.

Con fines prácticos, los tipos y funciones de los metadatos pueden clasificarse en tres amplias categorías: descriptivos, estructurales y administrativos. Estas categorías no siempre tienen límites bien definidos y con frecuencia presentan un significativo nivel de superposición. Por ejemplo, los metadatos administrativos pueden incluir una amplia gama de información que podría considerarse como metadatos descriptivos y estructurales.<sup>16</sup>

- Los metadatos descriptivos son aquellos que sirven para la descripción e identificación de los recursos de información en el nivel local (sistema) para permitir la búsqueda y recuperación de información.
- Los metadatos estructurales facilitan la navegación y presentación de recursos electrónicos, proporcionan información sobre la estructura interna de los recursos, incluyen página, sección, capítulo, numeración, índices y tabla de contenidos; describen la relación entre los materiales, por ejemplo: la fotografía B se incluyó en el manuscrito A; además, unen los archivos y los textos relacionados.
- Los metadatos administrativos facilitan la gestión y procesamiento de las colecciones digitales, tanto a corto como a largo plazos e incluyen datos técnicos sobre la creación y el control de calidad; comprenden la gestión de derechos y requisitos de control de acceso y utilización, además de información sobre la acción de preservación.

*Bill Prothman* distingue dos tipos de metadatos: los que describen la entidad de datos o el objeto de información y aquellos que describen su almacenamiento y uso. No obstante, señala cinco categorías más específicas:<sup>17</sup>

- *Metadatos de acceso*: permiten la interrogación, navegación y recuperación de información; describen, entre otros aspectos, cómo se estructuran lógicamente los datos.
- *Metadatos semánticos* : sirven para dotar a la información almacenada de un significado o propósito específico.
- *Metadatos de calidad* : posibilitan un análisis cualitativo de los datos.
- *Metadatos de transferencia* : se refieren a cómo los datos pueden transferirse entre aplicaciones.
- *Metadatos de almacenamiento* : revelan cómo y dónde se almacenan los datos en un sistema.

Al analizar los metadatos desde la perspectiva del contenido, sobre la base del contenido informativo que engloban e identifican al mismo tiempo y su función clave en la interoperabilidad semántica, *Kashyap* y *Sheth* distinguen dos tipos de metadatos:<sup>18</sup>

- *Metadatos independientes del contenido*: aquellos que recogen la información que no depende del contenido del documento, como la localización, la fecha de creación, modificación, etcétera.
- *Metadatos dependientes del contenido*: aquellos que, por supuesto, dependen del contenido asociado al objeto de información que describen. Este tipo de metadatos recoge normalmente información sobre la representación y estructura y facilita la interoperabilidad. Se divide en:
- *Metadatos basados en el contenido directo*: aquellos que dependen directamente del contenido del objeto de información. Por ejemplo, los índices de un documento a texto completo o el color y la forma en una imagen digital.
- *Metadatos descriptivos del contenido* : que describen la información de un documento sin utilizar expresamente su contenido; por ejemplo, las anotaciones textuales (descriptores, identificadores, etc) que especifican el contenido de una imagen.

A propósito de esta división o clasificación de *Kashyap* y *Sheth* que involucra la función que desempeñan los metadatos en la interoperabilidad semántica, es oportuno referirse al hecho de que uno de los conceptos clave en la recuperación de información basada en metadatos es el de interoperabilidad, que puede definirse, en el marco de los metadatos, como la habilidad de dos o más sistemas o componentes para intercambiar información y utilizarla intercambiada sin un esfuerzo especial en su sistema.<sup>19</sup> La interoperabilidad semántica es la posibilidad de que se pueda entender cada elemento dentro de un esquema o sistema de metadatos; por ejemplo: en el formato de metadatos Dublin Core, el elemento <creator> se define como la persona u organización responsable del contenido intelectual de un recurso; sin embargo, en el formato TEI se define el elemento <autor> en los términos siguientes: contiene el nombre del autor o de los autores, sean personales o corporativos de una obra; constituye la mención de responsabilidad para toda una unidad bibliográfica (TEI Header <<http://www.tei-c.org/P4X/HD.htm>>). En estas definiciones, los dos

elementos, según los diferentes formatos son semánticamente equivalentes, pero son tipos de propiedades diferentes. Es indispensable que exista interacción semántica para una efectiva recuperación entre diferentes formatos de metadatos.

Al comparar los atributos de distintos esquemas, *Burnett*, *Bor Ng* y *Park* distinguen dos tipos de metadatos:<sup>20</sup>

- *Metadatos intrínsecos* : incluyen atributos como: materia, título, autor, editor, lugar de publicación, otro agente, fecha, tipo de objeto de información, forma del identificador (URN [Uniform Resource Name], ISBN), relación, fuente, idioma, cobertura, resumen, versión, notas, firma, clasificación, nivel de seguridad y descriptores.
- *Metadatos extrínsecos* : abarcan información como: requerimientos del sistema, modo de acceso, accesibilidad, coste, control, extensión o tamaño del documento, descripción codificada y descriptores de la revisión.

Si se consideran los diferentes aspectos que influyen en la asignación de metadatos, puede desarrollarse una clasificación basada en la idiosincrasia del propio proceso de creación de metadatos, como se muestra a continuación:<sup>7</sup>

En dependencia de la fuente y momento de creación, se pueden distinguir:

- *Metadatos internos* : generados en el momento de creación o digitalización del documento.
- *Metadatos externos* : relacionados con un documento, pero creado por alguien distinto al agente que crea la información.

Según el método concreto de creación se distinguen:

- Metadatos generados automáticamente por el ordenador.
- Metadatos creados manualmente en la cabecera del documento.

Podrían señalarse también los creados manualmente pero por medio de una plantilla o cualquier software que facilite su creación. En dependencia del nivel de especialización de los responsables de la asignación de metadatos, se distinguen:

- Los creados por personas que no son expertas en las temáticas ni en documentación.
- Los creados por especialistas, por expertos.
- Por el control semántico de los datos:
- Metadatos controlados que responden a un vocabulario estándar, lista de autoridades o tesoro.
- Metadatos no controlados que no responden a ningunas de estas herramientas de control terminológico.

Según la forma de almacenamiento, pueden distinguirse:

- Metadatos embebidos en el propio documento.
- Metadatos separados del documento.

Por el nivel de estructuración, se clasifican en:

- Metadatos altamente estructurados (MARC, TEI, etc.).
- Metadatos no estructurados (anotaciones).

Según la perdurabilidad, existen:

- Metadatos estáticos, que no suelen cambiar una vez creados (título, fecha de creación, etc.)
- Metadatos dinámicos, que pueden cambiar con el uso y la manipulación del



documento -por ejemplo, los datos relativos a la resolución de una imagen.

- Metadatos de larga duración, que aseguran que el objeto de información digital sea accesible con el transcurso del tiempo.
- Metadatos de corta duración, son normalmente los de tipo operacional; datos relativos a la administración del documento.

Actualmente, existen en la red éstas y muchas otras clasificaciones de metadatos realizadas por diferentes autores; esta proliferación de clasificaciones responde a que los metadatos aún están en fase de construcción técnica y por eso no existe un consenso generalizado en su conceptualización o sobre los tipos de metadatos existentes.

Funciones:

Las funciones de los metadatos pueden analizarse desde el nivel del sistema y desde el nivel del usuario final. En el primero, facilitan la interoperabilidad y la capacidad de compartir (*shareability*) datos entre las herramientas de descubrimiento de recursos y esto acelera la concreción de proyectos, mejora la utilidad de las investigaciones y de la toma de decisiones, así como reduce costos al minimizar la duplicación de esfuerzos. Desde la perspectiva del usuario, pueden facilitar la capacidad de determinar: qué datos están disponibles (¿existen los objetos de información?, ¿dónde están?); si satisfacen necesidades específicas (¿es auténtico?, ¿es bueno?, ¿cómo puede determinarse si es útil o no?); cómo adquirirlo y cómo transferirlo a un sistema local.<sup>5</sup>

Aunque los tipos y las funciones de los metadatos están muy relacionados, *Steele* distingue tácitamente dos funciones principales de los metadatos:<sup>21</sup>

- Proveer un medio para descubrir qué datos existen y cómo pueden obtenerse o accederse.
- Proveer un mecanismo de búsqueda para reunir metadatos.

Por su parte, *Martínez Arellano* afirma que los metadatos tienen tres funciones básicas:<sup>22</sup>

- Proporcionar una descripción de una entidad de información junto con otra información necesaria para su manejo y preservación.
- Proporcionar los puntos de acceso a esa descripción.
- Codificar esa descripción.

Los metadatos, aún cuando cumplen con las funciones tradicionales de los catálogos —identificación y descripción de la información, búsqueda, recuperación y ubicación de la información— también desempeñan otras funciones inherentes a la información en formato digital como objeto de información, entre las que pueden citarse: limitaciones de uso, valoración del contenido, formas de acceso a la información, autoría y propiedad intelectual, actualización de la información, accesibilidad de los contenidos, preservación y conservación, visibilidad de la información y formas de acceso a la información.

Según nuestra opinión, entre las funciones fundamentales de los metadatos está el describir los objetos de información, informar a los usuarios la existencia, el contenido, la calidad y los objetivos de los recursos de información descritos, con el fin de que ellos puedan determinar si un recurso de información es potencialmente útil, sin necesidad de acceder a este. A su vez, los metadatos deben garantizar el punto de acceso a la información que los usuarios desean consultar.

Los metadatos también son útiles para preservar en forma documentada la información que poseen las instituciones, esto evita que las instituciones sean vulnerables a perder todo el conocimiento sobre sus datos, si por algún motivo desaparece la persona que los originó.

## Metadatos y HTML

El lenguaje de marcado de hipertextos HTML (*HyperText Markup Language*) es una herramienta sencilla para especificar la estructura de los documentos en la red, a partir de un conjunto de etiquetas que sirven para definir la forma de presentar el texto y otros elementos de las páginas. La definición de tipo de documento HTML, con su estructura elemental, se orienta principalmente a describir los elementos estructurales (cabecera y cuerpo del documento) que aparecen en el hipertexto.

El HTML, por tanto, se comporta de forma superficial en la representación de la información e indica solo a los navegadores cómo deben representar la información en lugar del significado de la información.<sup>7</sup> Para describir la información con el objetivo de recuperarla posteriormente, el HTML posibilita a los autores de las páginas Web especificar metadatos o información sobre los documentos.

En HTML, los elementos META o metatag se utilizan para que el propio autor pueda especificar información sobre el documento y se insertan en la cabecera del documento. Las etiquetas <META> son de aquellas que se llaman "elementos vacíos", porque no llevan etiqueta de cierre. Su estructura es la de una declaración propiedad: valor (de una propiedad y de un valor para esta propiedad) que puede realizarse de dos formas diferentes:

- Desde dentro de un documento, por medio del elemento META.
- Desde fuera de un documento, vinculando los metadatos por medio del elemento LINK.

Aunque la forma común de insertar metadatos en un documento HTML es por medio de la etiqueta <META> no existe ninguna regla sobre la información que se puede o no incluir en los metatags. El standard HTML define la estructura de una etiqueta META, pero no sus valores.

Por ejemplo:

```
<meta name="Author" content="Isabel Daudinot Founier" lang="es">
```

Esta declaración dice que el autor del documento es Isabel Daudinot Founier, e indica que el valor del atributo content está en español.

A continuación se describirán, con ejemplos, las etiquetas META de HTML más comunes: (Metadatos en html):

### 1. *Description:*

```
<meta name="description" content="Servicio de páginas amarilla que ofrece información sobre la localización de instituciones cubanas o extranjeras radicadas en territorio nacional, así como de los servicios que brindan".
```

 Mediante esta etiqueta puede incluirse como metadato una descripción de la página o documento. Esta descripción puede leerse y utilizarse por algunos buscadores. La etiqueta <meta> admite el atributo lang, que permite ofrecer información sobre el idioma.

### 2. *Keywords:*

```
<meta name="keywords" content="Páginas amarillas, Cuba, Empresas cubanas, Negocios, Instituciones cubanas, Información de Cuba".
```

 Las palabras clave son aquellas que resumen de forma significativa el contenido de la página.

Estas dos etiquetas META (description y keywords) son las más importantes. Algunos buscadores leen estas etiquetas y juzgan sobre la importancia de la página, según ellas. Una buena descripción y un grupo de palabras clave correctamente asignadas, puede ayudar mucho a aumentar la visibilidad de una página Web.

### 3. *HTTP-EQUIV*:

Las páginas Web se distribuyen por Internet mediante el protocolo http, que regula el diálogo entre el servidor y el navegador. Con estas etiquetas, el autor de la página puede predeterminar aspectos de este diálogo. Su sintaxis general es:

```
<meta http-equiv="content-type" content="text/html; charset=ISO-8859-1"/>
```

Con esta etiqueta META, se indica al navegador el tipo de documento: si es un documento html o un documento de texto plano o cualquier otra cosa que se desee indicar. En el ejemplo anterior, significaría que es un documento texto-html y que usa un juego de caracteres ISO-8859-1. esta información es útil para que el navegador de los visitantes conozca qué juego de caracteres debe emplear.

### 4. *Refresh*:

```
<meta http-equiv="refresh" content="5"/>
```

```
<meta http-equiv="refresh" content="5; url=http://www.cuba.cu/mapa.php"/>
```

Mediante esta etiqueta y el valor refresh se indica al navegador que debe recargar la página una vez transcurrido el tiempo especificado. En la primera línea, un refresh simple, con orden de recarga a los cinco segundos. En la segunda línea, un refresh con redirección. El navegador debe refrescar la página a los cinco segundos, recargando en su lugar la especificada.

### 5. *Expires*:

```
<META HTTP-EQUIV="expires" CONTENT="Tue, 15 Dec 2002 21:29:02 GMT">
```

Mediante la orden expires, se informa sobre la vigencia del documento. Su utilidad principal es la de evitar que el navegador guarde indefinidamente en caché una página y forzarlo a cargar las nuevas versiones. La fecha debe presentarse necesariamente en formato RFC840 (GMT), es decir:

día\_de\_semana, DD-Mon-YY HH:MM:SS TIMEZONE(Greenwich) todo eso en inglés.  
Por ejemplo Fri, 01 Jan 2003 07:20:55 GMT

Si la fecha está expresada en un formato distinto, el valor de expires se interpretará como now, es decir, se fija una caducidad inmediata.

### 6. *Pragma*:

```
<meta http-equiv="pragma" CONTENT="no-cache"/>
```

Ordena al navegador que no almacene una copia cache de la página. Resulta útil cuando se trata de páginas dinámicas cuyo contenido cambia frecuentemente. Su único valor aceptable es no-cache.

### 7. *Set-cookie*:

```
<META HTTP-EQUIV="Set-Cookie"  
CONTENT="nombre_cookie;expires=fecha_valor;path=" />
```

Mediante este metatag, puede enviarse un cookie al navegador del usuario. Si contiene fecha de vencimiento, el navegador la conservará en el disco duro hasta dicha fecha; en caso contrario, se borrará al terminar la sesión de navegación.

### 8. *Window-target*:

Sirve para especificar el nombre de la frame donde la página debe cargarse. Por

ejemplo:

```
<META HTTP-EQUIV="window-target" CONTENT="-top">
```

Estas son las más corrientes de entre las dirigidas a tratar con los servidores. Las menos utilizadas son:

```
<META HTTP-EQUIV="content-Script-Type" CONTENT="text/javascript">
```

```
<META HTTP-EQUIV="content-Style-Type" CONTENT="text/css">
```

Ellas informan respectivamente el tipo de lenguaje script utilizado y el uso de CSS

#### 9. *Robots:*

Se pueden ofrecer instrucciones a los robots de los buscadores acerca de qué páginas debe incluir en sus bases de datos y cuáles no. Sus valores son index, que indica la página que debe procesarse; noindex es justo lo contrario; follow indica al robot que siga los links que existan en la página y procese las páginas resultantes; nofollow lo contrario. Los valores por defecto se entiende que son index, follow (separados por coma).

```
<META NAME="robot" CONTENT="index,follow">
```

Estas son las etiquetas META de mayor uso que contempla el HTML y que no están sujetas a ningún formato en particular. El autor de un recurso puede utilizarlas con los propósitos anteriormente descritos. Con el objetivo de dar mayor flexibilidad y precisión a la descripción de los documentos, pueden utilizarse en HTML, formatos normalizados de metadatos como el Dublin Core, TEI, entre otros. Estos formatos de metadatos se incrustarían igualmente en la cabecera del documento.

Aún cuando estas etiquetas META sirven para describir la información de los documentos en HTML, éstas no garantizan la flexibilidad y la semántica necesaria, el HTML tiene un conjunto de etiquetas prefijadas y finitas que resultan insuficientes, en muchos casos, para una descripción cabal de determinados recursos de Información.

La superficialidad e inflexibilidad del lenguaje HTML para la representación de la información obliga a la utilización de otro lenguaje más flexible que asegure una mayor consistencia en el procesamiento de la información en Internet y que conduzca hacia una Web más semántica, que asegure la interoperabilidad que la red necesita. Este nuevo lenguaje es, sin lugar a dudas, el XML, que representa una arquitectura de la información más abierta y flexible. Para la representación de la información en el nuevo escenario de XML, se utilizará el RDF (Resource Description Framework) cuyas posibilidades para estructurar, intercambiar y recuperar información son infinitas.

#### **Metadatos y vocabularios controlados**

La recuperación de información es parte de un complejo sistema de comunicación entre los autores de la información registrada en los documentos y los usuarios que la solicitan. Por tanto, la función principal de los trabajadores de la información es planificar, diseñar, crear, desarrollar y mantener buenos canales o puentes para esta comunicación.

En cualquier sistema documental, cuando los documentos ingresan es necesario realizar una labor de análisis del contenido del documento en cuestión y una traducción al lenguaje de indización empleado. De la misma manera, cuando se efectúa una consulta hay que realizar también estas dos operaciones para traducir la petición de información del usuario en una consulta en los términos del lenguaje de indización.<sup>23</sup>

En la cadena documental, el análisis documental consiste del resumen en texto libre y

la indización en lenguaje controlado de los documentos.<sup>24</sup>

Como alternativa al lenguaje documental, puede recurrirse al lenguaje empleado en el propio documento para realizar dicha representación; en este caso, se hablará de indización y recuperación en lenguaje libre.

Algunos autores, entre los cuales se encuentra *Angós Ullate*, plantean que, a la hora de elaborar las descripciones de los recursos de información, es muy importante el uso de lenguajes normalizados y de índices para facilitar la recuperación de información.<sup>25</sup>

Sin embargo, desde que la World Wide Web se ha generalizado como medio de difusión y publicación de información, los índices se elaboran, por lo general, con términos extraídos del lenguaje natural, y se echa en falta la descripción bibliográfica de los documentos.

La falta de normalización en la creación de los índices y la descripción de los documentos tiene consecuencias negativas, imposibilitan la búsqueda de información pertinente y hacen que el proceso de recuperación de la información sea deficiente.

Por esta razón, la etapa de análisis documental se ha convertido nuevamente en la clave de este proceso: el uso de tesauros, encabezamientos de materias u otro tipo de vocabularios controlados para la descripción de los diferentes elementos de un documento en las etiquetas META, son aún muy importantes en la descripción de los documentos en Internet, y se debe seguir realizando.

*Méndez*, por su parte, opina que los tesauros en el contexto de la información distribuida de la intranet, estructurada por metadatos, proporcionarán un soporte a la recuperación de información basado en el conocimiento y facilitará la combinación de múltiples bases de datos o la unificación del acceso a diversos contenidos. Los tesauros en estos sistemas de información serán, entonces, algo más que una mera herramienta para la indización: son el soporte semántico de la metainformación.<sup>13</sup>

En esta misma línea, *Blain* afirma que el más importante aspecto [...] concierne a cómo están representados los documentos en un sistema. La computadora más rápida o las más sofisticadas técnicas de búsquedas no pueden superar los problemas derivados de una pobre representación [de naturaleza lingüística] de los documentos.<sup>26</sup>

Mientras, *Gill Urdiciain* plantea que el lenguaje libre ofrece la posibilidad de recuperar información muy especializada y actualizada, cuya terminología o no está incluida en el vocabulario controlado, o no está representada de forma suficientemente específica. El lenguaje libre neutraliza las deficiencias del lenguaje controlado y viceversa, y por esto ambos sistemas no solamente no son antagonistas sino que se complementan el uno al otro.<sup>27</sup>

En ocasiones, los especialistas en información se enfrentan al problema de que no existen entre los términos finitos de un lenguaje controlado (tesauro) determinado, algunos de los descriptores necesarios para indizar un documento. En estos casos, es muy socorrido recurrir al lenguaje natural y más aún en el contexto de Internet, donde un mismo sitio puede tener información de diferentes disciplinas.

En los sitios de carácter general, deben emplearse tesauros también de carácter general; por eso, puede ser muy común no encontrar los términos adecuados para la descripción de la información concerniente a cada página Web. Recurrir, en estos casos, al lenguaje natural y combinarlo con el lenguaje controlado es una solución. Ahora bien, no debe perderse de vista que siempre que sea posible la utilización de un lenguaje controlado se estructurará mejor la información y, por tanto, su recuperación será mas efectiva.

Sin embargo, no se debe perder de vista que el uso del lenguaje natural o libre conduce a la ambigüedad y ella incrementa significativamente la cantidad de recursos

de información que se recupera, pero reduce la precisión o relevancia de la información recuperada. Por esta razón, siempre que sea posible es muy importante la utilización de un vocabulario o lenguaje controlado para ofrecer una mayor consistencia a las palabras utilizadas y a la indización en sentido general, y eliminar los problemas de polisemia y sinonimia que tanto afectan la recuperación de la información.

## REFERENCIAS BIBLIOGRÁFICAS

1. Spedalieri G. Selección y registro de recursos electrónicos. 2001. Disponible en URL: <http://abgra.sisbi.uba.ar/documentos/catalogacion.pdf> [Consultado: 6 de enero del 2006].
2. Bermello Crespo L. Bibliotecas digitales y actividad bibliotecaria. *Ciencias de la Información* 2001;32(1):57-68.
3. De Gennaro R. Bibliotecas, la tecnología y el mercado de la información. México DF: Iberoamérica, 1993. p.3-5.
4. Heery R. Review of metadata formats. *Program* 1996; 30(4):345-73. Disponible en URL: <http://www.ukoln.ac.uk/metadata/review.html> [Consultado: 10 de enero del 2006].
5. Burnett K. Control or management: a comparison of the two approaches for establishing metadata schemes in the digital environment Rutgers University ( New Jersey ), Florida State University (Florida). 1997. Disponible en URL: <http://www.scils.rutgers.edu/~sympark/asis.html> [Consultado: 4 de enero del 2006].
6. Real Academia Española. Diccionario de la Lengua Española. 22 ed. Madrid: Real Academia Española; 2001.
7. Méndez Rodríguez EM. Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: Trea; 2002. p.429.
8. Liston MD, Dolby JL. Metadata System for Integrated Access to Numeric Data File', *Drexel Library Quarterly* 1982;18(3-4):147-60.
9. Gradmann S. Catalogación Versus Metadata: Vino Viejo en Odres Nuevos? 1998. Disponible en URL: <http://www.ifla.org/IV/ifla64/007-126s.htm> [Consultado: 14 de enero del 2006].
10. San Segundo MR. Organización del conocimiento en Internet: metadatos bibliotecarios DUBLIN CORE'. Disponible en URL: [http://fesabid98.florida-uni.es/Comunicaciones/r\\_sansegundo.htm](http://fesabid98.florida-uni.es/Comunicaciones/r_sansegundo.htm) [Consultado: 14 de enero del 2006].
11. Xu A. Accessing Information on the Internet. Disponible en URL: <http://www.oclc.org/oclc/man/colloq/xu.htm> [Consultado: 9 de enero del 2006].
12. Dempsey L, Heery R. Metadata: A Current View of Practice and Issues. *Journal of Documentation* 1998;54(2):149.
13. Méndez Rodríguez EM, Merlo Vega JA. Localización, identificación y descripción de recursos Web: tentativas hacia la normalización. En: VII Jornadas Españolas de Documentación. Bilbao: Universidad del País Vasco; 2000. p.221-31.
14. Seiner RS. Questions metadata can answer. Disponible en URL:
15. [http://www.cai.com/products/decisionbase/questions\\_metadata\\_can\\_answer.pdf](http://www.cai.com/products/decisionbase/questions_metadata_can_answer.pdf) [Consultado: 18 de enero del 2006].
16. Berners-Lee T. Tejiendo la Red: el inventor del World Wide Web nos descubre su origen. Madrid: Siglo XXI de España; 2000.
17. Biblioteca de la Universidad de Cornell. Departamento de Investigación. Tutorial de digitalización de imágenes. Llevando la teoría a la práctica. 2001. Disponible en URL: <http://www.library.cornell.edu/preservation/tutorial-spanish/metadata/metadata-01.html> [Consultado: 16 de enero del 2006].
18. Prothman B. Meta Data: Managing Needles in the Proverbial Haystacks. 2000. Disponible en URL: <http://www.ieee.org/membership/students/potentials/febmar2000/pdf/pages20.pdf> [Consultado: 20 de enero del 2006].
19. Kashyap V, Sheth A.. Information Brokering Across Heterogeneous Digital data: A Metadata –based approach. Boston : Kluwer Academic Publishers, 2000.
20. ALA. Committee on Cataloguing. Description and access. Task Force on Metadata: Summary Report, june 1999. Disponible en: <http://www.libraries.psu.edu/tas/jca/ccda/tf-meta3.html> [Consultado: 2 de febrero del 2006].
21. Burnett K, Kwong Bor Ng, Soyeon P. A Comparison of the two traditions of metadata development. *JASIS (Special Topic Issue: Integrating Multiple*

- Overlapping Metadata Standarts) 1999;50(13):1209-17.
22. Steele C. 1996. Organization of knowledge of the Internet. Information Management Report 1996:1-6.
  23. Martínez Arellano FF. 2000, Metadatos y organización de recursos electrónicos. México DF: Colegio de Bibliotecología. Facultad de Filosofía y Letras. Universidad Nacional Autónoma de México, 2000. Disponible en URL: <http://cuib.unam.mx/~felipe/metadata2000/indice.htm#contenido> [Consultado: 24 de enero del 2006].
  24. López Ferrer M. Concepción, construcción y mantenimiento de una herramienta de indización para un centro de documentación sobre gestión de la innovación y el conocimiento: una experiencia. En: VII Jornadas Españolas de Documentación. Bilbao: Universidad del País Vasco, 2000.p.165.
  25. Valle Gastaminza F. Nuevo paradigma del análisis documental de programas y materiales de televisión. 2000. Disponible en URL: <http://www.ucm.es/info/multidoc/prof/fvalle/pradig.htm> [Consultado: 24 de enero del 2006].
  26. Angós Ullate JM., Fernández Ruíz MJ., Salvador Oliván JA. El impacto de WWW en la evolución y futuro de los centros de documentación. Disponible en URL: <http://www.cobdc.org/09jornades/7es/25.pdf> [Consultado 12 de febrero del 2006].
  27. Blain DC. Language and representation in information retrieval. Amsterdam : Elsevier, 1990. p.155.
  28. Urddiciain BG. Evaluación del rendimiento de tesauros españoles en sistemas de recuperación de información. Revista Española de Documentación Científica 1998;21(3):301.

Recibido: 4 de mayo de 2006. Aprobado: 26 de mayo de 2006.

MSc. *Isabel Daudinot Founier*. Empresa de Tecnologías de la Información y Servicios Telemáticos Avanzados. Ave. 47 e/ 18 y 20, Miramar, Ciudad de La Habana. Cuba. Correo electrónico: [isabel@citmatel.inf.cu](mailto:isabel@citmatel.inf.cu)

**<sup>1</sup>Máster en Ciencia de la Información y Bibliotecología. Empresa de Tecnologías de la Información y Servicios Telemáticos Avanzados.**

Ficha de procesamiento

Clasificación: Artículo de revisión

Términos sugeridos para la indización

Según DeCs<sup>1</sup>

INTERNET; ALMACENAMIENTO Y RECUPERACIÓN DE LA INFORMACIÓN.

INTERNET; INFORMATION STORAGE AND RETRIEVAL.

Según DeCI<sup>2</sup>

INTERNET; WWW; RECUPERACIÓN DE LA INFORMACIÓN; METADATOS; TECNOLOGÍA DE LA INFORMACIÓN.

INTERNET; WWW; INFORMATION RETRIEVAL; METADATOS; INFORMATION TECHNOLOGY.

<sup>1</sup>BIREME. Descriptores en Ciencias de la Salud (DeCS). Sao Paulo: BIREME, 2004.

Disponible en URL: <http://decs.bvs.br/E/homepagee.htm>

<sup>2</sup>Díaz del Campo S. Propuesta de términos para la indización en Ciencias de la Información. Descriptores en Ciencias de la Información (DeCI). Disponible en URL:

<http://cis.sld.cu/E/tesauro.pdf>

**[Índice Anterior](#) [Siguiete](#)**