Revista Digital Universitaria
10 de julio 2005 • Volumen 6 Número 7 • ISSN: 1067-6079

ARTICULO

DESCRIPCIÓN E IDENTIFICACIÓN DE RECURSOS EN INTERNET: METADATA

Dr. Félix del Valle Gastaminza
Profesor Titular de la Universidad Complutense de Madrid.
Departamento de Biblioteconomía y Documentación.
fvalle@ccinf.ucm.es
http://www.ucm.es/info/multidoc/profesores/fvalle/

D. Rodrigo Sánchez Jiménez
Ayudante de la Universidad Complutense de Madrid.
Departamento de Biblioteconomía y Documentación.
rsanchezj@ccinf.ucm.es
http://multidoc.rediris.es/rodrigosj/

D. José Ramón Pérez Agüera
Ayudante de la Universidad Complutense de Madrid.
Departamento de Sistemas Informáticos y Programación
jpereza@ccinf.ucm.es
http://multidoc.rediris.es/joseramon/

RESUMEN

Presentamos una perspectiva general e introductoria de los metadatos y su utilización en Internet, la relación de los metadatos con la solución de algunos de los problemas propios de Internet, así como con las apuestas de futuro en el ámbito de la Web Semántica. Se presta una especial atención a una de las iniciativas más importantes en el ámbito de los metadatos, DCMI y su conexión con las tecnologías del W3C para el desarrollo de la Web Semántica.

Palabras clave: Metadatos, Dublin Core, Web Semántica, RDF, Recuperación de Información

Resource Description and Identification on the Internet: Metadata

ABSTRACT

This paper tries to show a wide scope and introductory overview on metadata and its use on the Internet, the relation between metadata and the solution of some of the problems inherent to the Internet, as well as its relation with the Semantic Web. We also focus on one of the most important metadata initiatives, the DCMI, and its connections with the W3C technologies for the Semantic Web Development.

Keywords: Metadata, Dublin Core, Semantic Web, RDF, Information Retrieval

10 de julio 2005 • Volumen 6 Número 7 • ISSN: 1067-6079

INTRODUCCIÓN

El concepto Metadatos es difícil de definir con claridad. El prefijo "meta" significa "sobre", "junto a", "después", lo que sugiere que los metadatos son compañeros de viaje de los datos, sean estos un artículo científico, una fotografía o un sitio web. Los metadatos describen un recurso de información o facilitan el acceso a un recurso de información y es inherente a los metadatos el hecho de hay una asociación de algún tipo entre los elementos de metadatos y el recurso de información al que se refieren. En principio es el mismo tipo de relación que existe entre una ficha catalográfica de una biblioteca y el libro correspondiente. La información almacenada en el campo "META" de una página web HTML son metadatos y están asociados de forma intrínseca al recurso de información al que se refieren. Pero también los datos de indización obtenidos por los robots de búsqueda son metadatos relacionados con el recurso de información por medio de la URL. En cualquier caso podemos comprender que los metadatos son información o documentos secundarios[1].

Se puede establecer la siguiente tipología de metadatos:

Metadatos descriptivos: Relacionados con el contenido. Recogen información sobre lo que el documento contiene o trata. Su utilidad es sobre todo localizadora pues aportan información que los robots de búsqueda procesan para el usuario final.

Metadatos administrativos: Relacionados con el contexto. Explican el quién, el qué, el cuándo, el dónde y el porqué de un determinado documento o recurso. Su objetivo es facilitar a los gestores de colecciones a manejar los documentos, archivarlos, distinguirlos, gestionar sus derechos, preservarlos, etc.

Metadatos técnicos: Relacionados con los procesos necesarios para la reproducción y almacenamiento de los materiales, su calidad, o datos sobre su digitalización.

Metadatos estructurales. Establecen las relaciones intrínsecas o extrínsecas de un documento o conjunto. Por ejemplo, las relaciones entre distintos capítulos de un libro, artículos de una revista, o el número de una revista dentro de una colección.

Esta distinción es útil para describir el conjunto de aspectos que suele ser el objeto de los metadatos, aunque cualquier iniciativa combinará varios o todos los tipos mencionados con anterioridad. Existen conjuntos de metadatos para un número muy elevado de propósitos específicos de complejidad diversa, como la descripción de fotografías para el ámbito de las bibliotecas y los museos[2], el procesamiento de fotografía[3], el tratamiento de documentación de video en la industria audiovisual[4] o su conservación[5]. Nosotros haremos un breve recorrido por una iniciativa de alcance mucho más general, el Conjunto de Metadatos Dublin Core.

---

[1] Para una definición / categorización alternativa y también muy clarificadora se puede consultar Gilliland-Steward, Anne. Setting the Stage. Publicado en la web de metadatos del Trust J. Paul Getty en 2000. http://www.getty.edu/research/conducting_research/standards/intrometadata/2_articles/index.html
Para una visión más exhaustiva recomendamos el trabajo de Eva Méndez, sobre todo el primer capítulo pp. 29-72. cf. Bibliografía.

[2] Véase la iniciativa VRA Core 3.0 http://www.vraweb.org/vracore3.htm

[3] La mayor parte de las cámaras digitales generan precisamente Exif, un conjunto de metadatos fuertemente implicado en la esfera técnica. http://www.exif.org/

[4] SMPTE es sin duda el diccionario de metadatos más importante del ámbito del audiovisual. La dirección de su sitio web es: http://www.smpte-ra.org/mdd/index.html

[5] El proyecto NEDLIB, http://www.kb.nl/coop/nedlib/

LA NORMA DUBLIN CORE

El conjunto de elementos de metadatos del Dublin Core[6] fue creado a partir de 1995 como una respuesta para mejorar la eficacia de la recuperación en la WWW. El Dublin Core ha sido desarrollado como una norma genérica de metadatos para el uso de bibliotecas, archivos, gobierno y otros editores de información electrónica y ha adquirido la consideración de Standard por parte de la ISO, que lo denomina Standard 15836-2003. En cualquier caso es con mucho la iniciativa más extendida en Internet.

Tal y como fue pensado originalmente era bastante circunspecto en sus objetivos y pretendía servir para tratar "documentos objeto", como páginas HTML, archivos PDF e imágenes (GIF, JPEG, etc.) La norma del Dublin Core trataba de ser descriptiva, más que evaluadora. Se limitaba deliberadamente a un pequeño conjunto de elementos (quince) que podían aplicarse a un amplio tipo de recursos de información.

La norma ha ido desarrollando distintas controversias en lo que se refiere a la semántica de los metadatos (reglas para el contenido de los quince campos) como para la sintaxis (reglas para estructurar y expresar los campos mismos). En ambas áreas se han producido cambios y no son definitivas. Las características de la norma DC pueden sintetizarse en cinco categorías:

Simplicidad. Está pensado para que lo utilicen tanto personas no familiarizadas con la catalogación como profesionales de la descripción. Los elementos son muy claros semánticamente y carecen de la complejidad y aridez de una ficha de catalogación.

Interoperabilidad semántica. En Internet, la disparidad de modelos de descripción interfiere en la capacidad de buscar por encima de las barreras entre disciplinas. Si se promueve un conjunto común de descriptores que ayuden a unificar otros datos de contenido se incrementa la interoperabilidad semántica entre disciplinas.

Consenso internacional. La aceptación de DC en el ámbito internacional facilitaría un Internet más democrático, menos controlado por materiales de una sola procedencia.

Extensibilidad. Es una alternativa económica a otros modelos de descripción más complejos como el MARC pero tiene suficiente flexibilidad y extensibilidad para aceptar la estructura y semántica más elaborada inherente a otras normas más complejas.

Modularidad de los Metadatos en la Web. La diversidad de metadatos existente en la red requiere una infraestructura que soporte la existencia de paquetes de metadatos independientes y complementarios.

Veamos los elementos que recoge DC, tal y como aparecen enunciados en la versión española del documento original de la DCMI[7]:

Título. Etiqueta: TITLE. Nombre dado a un recurso por su CREADOR o EDITOR.

Autor o Creador. Etiqueta: CREATOR. La persona u organización responsable del contenido intelectual del recurso. Por ejemplo, autores en el caso de los documentos escritos, o artistas, ilustradores o fotógrafos para recursos visuales.

---

[6] http://dublincore.org/ También se puede visitar http://es.dublincore.org/ para la versión castellana.
[7] http://www.rediris.es/metadata/

10 de julio 2005 • Volumen 6 Número 7 • ISSN: 1067-6079

Tema y descriptores. Etiqueta: SUBJECT La materia del recurso. Debe ser expresado mediante palabras o frases que describan el contenido o materia del recurso. Se debe favorecer el uso de vocabularios controlados y de esquemas formales de clasificación.

Descripción. Etiqueta: DESCRIPTION. Descripción textual del contenido de un recurso, incluyendo resúmenes documentales textuales o descripciones de contenido de recursos visuales.

Editor. Etiqueta: PUBLISHER Entidad responsable de la presentación del recurso, como una editorial, departamento de universidad o entidad.

Otros colaboradores. Etiqueta: CONTRIBUTOR Persona u organización, no específicamente autor, que ha realizado una contribución significativa a un recurso aunque secundaria respecto al elemento CREATOR.

Fecha. Etiqueta: DATE La fecha en la que el recurso es presentado en la forma actual. Se recomienda seguir la forma AAAA-MM-DD tal como está definido en http://www.w3.org/TR/NOTE-datetime, aplicación de la norma ISO 8601.

Tipo de Recurso. Etiqueta: TYPE Categoría de recurso, como Home Page, Novela, Poema, Ponencia, Informe técnico, Ensayo, Diccionario. Con objeto de facilitar la interoperabilidad el elemento TYPE debería seleccionarse de una lista normalizada en la que se está trabajando[8].

Formato. Etiqueta: FORMAT. Formato de datos del recurso, con objeto de identificar el software y hardware necesario para visualizar u operar el recurso. Como en el caso anterior, se está trabajando en listas normalizadas de formatos.

Identificador de Recurso. Etiqueta: IDENTIFIER Cadena de caracteres o números utilizados para identificar inequívocamente el recurso. Como ejemplos pueden citarse URL o URN. Otros identificadores internacionales, como DOI, pueden también utilizarse y para recursos off-line el ISBN.

Fuente. Etiqueta: SOURCE Cadena de caracteres o números utilizados para identificar inequívocamente la obra de la que el recurso deriva. Por ejemplo una versión en PDF de una novela puede llevar un elemento SOURCE con el ISBN de la edición impresa del libro.

Idioma Etiqueta: LANGUAGE Idioma(s) del contenido intelectual del recurso.

Relación Etiqueta: RELATION. Relaciones del recurso con otros recursos. El objetivo de este elemento es facilitar un medio de expresar relaciones entre recursos que tengan relaciones formales con otros aunque tengan entidad propia. Por ejemplo, imágenes en un documento, capítulos de un libro, artículos de una revista.

Cobertura Etiqueta: COVERAGE Características espaciales y temporales del recurso.

Gestión de derechos. Etiqueta: RIGHTS. Enlace a una nota de copyright, a una mención de derechos de autor, o a un servicio que facilite información sobre los términos de acceso a un recurso.

---

[8] La última propuesta está en http://sunsite.berkeley.edu/Metadata/types.html

Los metadatos se pueden utilizar localmente como campos de una base de datos o como parte de un sistema de intercambio de información de ámbito restringido. Ahora bien, ¿cómo hacemos para darle utilidad a nuestros metadatos en el contexto de Internet? Los metadatos se pueden utilizar en el contexto de Internet con la esperanza de ser recuperados por algún buscador o a través de redes de clientes y servidores de metadatos, como las propuestas por el modelo OAI[9]. Nosotros nos centraremos en el primero de los casos, ya que creemos que los problemas están bastante bien resueltos en el segundo, al menos mientras "llega la Web Semántica"[10].

METADATOS Y RECUPERACIÓN DE INFORMACIÓN

El objetivo fundamental del uso de metadatos en Internet es optimizar el acceso y recuperación de la información. Documentalistas y bibliotecólogos miden la recuperación de información en términos de relevancia y precisión. Si se pierde información relevante la tasa de relevancia es pobre (hay un alto nivel de silencio). Si aparece mucha información irrelevante hay una tasa de precisión baja (hay un nivel de ruido alto). En algunos caso (por ejemplo, búsqueda de normas, de patentes, de leyes) es esencial una relevancia muy alta pero, en muchos casos, los usuarios se contentan con tasas muy bajas, con un pequeño número de documentos relevantes.

El usuario tiene también la opción de los directorios y de los portales, sitios donde se establecen clasificaciones apriorísticas en las que se van incluyendo sitios que pueden tener información potencialmente útil sobre el tema propuesto. Dentro de estas clasificaciones destacan las realizadas por instituciones educativas y científicas que acaban siendo muy clarificadoras en lo referente a las páginas de referencia sobre un determinado ámbito pero adolecen de un problema de exhaustividad bastante evidente, ya que el porcentaje de los recursos procesables manualmente no llega a ser una ínfima parte de los potencialmente relevantes[11].

En la práctica los navegantes lo prueban todo: tienen sus buscadores favoritos, utilizan metabuscadores, conocen sitios donde hay enlaces de gran valor hacia cosas que interesan pero siempre son conscientes de la existencia de altísimas tasas de ruido y de silencio. Son conscientes de que no se encuentra todo lo que hay pero, sobre todo, se ve muy claramente que la mayor parte de lo que se obtiene no interesa. Este es un problema que se podría solucionar en gran parte mediante la utilización de metadatos, pero su utilización, aunque está bastante extendida, no es ni mucho menos uniforme.

Los buscadores y la indización tradicional

Al realizar una búsqueda simple en muchos de los buscadores utilizados en Internet utilizando una única palabra observamos que el número de respuestas es elevadísimo. En Google con la palabra FOTOGRAFIA se obtienen 4.550.000 páginas web: la tercera referencia que ofrece es, por cierto, una página en catalán sobre microfotografía de insectos. El ejemplo ilustra que los buscadores pueden devolver un enorme caudal

[9] http://www.openarchives.org/

[10] La iniciativa OAI resuelve de forma muy solvente el problema de intercambiar metadatos y aprovecharlos en beneficio de la recuperación de información, sin embargo abandona expresamente la posibilidad de ofrecer modelos de descripción más complejos con el objeto de facilitar la implementación de software OAI y la extensión del modelo, que intenta ser fundamentalmente interoperable. Por este motivo creemos más interesante dedicar estas líneas al ámbito de los buscadores y la Web Semántica.

[11] El Open Directory Project (http://dmoz.org/) es probablemente la iniciativa más interesante en este sentido, llegando a atesorar mediante la mediación de agentes humanos más de 4.500.000 sitios web. Sin embargo Google indizaba en las mismas fechas (Mayo de 2005) más de 8.000.000.000 páginas web.

10 de julio 2005 • Volumen 6 Número 7 • ISSN: 1067-6079

de información irrelevante porque cuentan con pocos medios para distinguir entre palabras importantes y palabras incidentales en los documentos de texto. Si la búsqueda se puede dirigir a palabras que realmente se están utilizando como términos significantes habrá una mejora en la precisión. Si podemos dirigir la búsqueda hacia palabras y frases cuyo rol esté correctamente identificado igualmente reduciremos el ruido: es decir, si puedo precisar que "Blanco" es el apellido del autor cuyos documentos busco no me aparecerá la ciudad de Blanco (Texas) o la página web de El Caballo Blanco, una especie de parque de vacaciones con cocodrilos en Australia. Este es el papel que pueden jugar a la perfección los metadatos que son capaces de especificar e identificar la información clave de un recurso de información: el autor, el título, la materia, el editor, etc. En otras palabras, la estructura propia de los metadatos es fundamental para una adecuada recuperación.

Los buscadores poseen una importante capacidad para decidir la importancia de los temas que componen un documento, lo que podríamos asimilar a la indización tradicional, y tecnológicamente es factible obtener la materia de los mismos documentos mediante técnicas de clasificación automática. Sin embargo existen muchas otras formas de ofrecer información sobre un documento (lo que en el ámbito de la documentación entenderíamos por catalogación) que ningún sistema parece hoy día ser capaz de ofrecer. En palabras de Eva Méndez:

...a diferencia de los indizadores humanos, estas aplicaciones, en general, no identifican características de un documento como la materia de la que trata, el autor, la fecha de publicación, el tipo de documento o las condiciones de acceso (no pueden distinguir, por ejemplo, si un documento es un poema o un informe científico)[12].

Esta "información catalográfica" se podría ofrecer como parámetros de búsqueda decisivos a la hora de recuperar un documento. Esta información requiere de su inserción manual (salvo contadas y parciales excepciones[13]) por parte de los autores, ya que el proceso en conjunto no parece ser automatizable el día de hoy. Pero al margen de su obtención manual o automática, la forma de hacer útiles estos metadatos es explicitándolos de forma estructurada y legible por máquinas, un asunto que trataremos algo después.

Obstáculos a vencer

El problema fundamental de los metadatos en el ámbito de Internet es que no todos los buscadores los utilizan como referencia válida a la hora de ponderar los documentos. El caso más llamativo es quizá el de Google, que declara públicamente su desconfianza en la información que los autores de Webs generan sobre sus contenidos[14]. Sin embargo el desarrollo de tecnologías que analicen el significado de los documentos es bastante difícil, y además sólo funciona con los documentos textuales, mientras que la recuperación de documentos fotográficos o audiovisuales es mucho más deficiente. Para comprobar esto basta con hacer algunas búsquedas aleatorias en Google y Altavista, lo que deja a la vista que la información utilizada para indizar las imágenes se extrae del contexto o de las etiquetas *alt* de descripción de imágenes.

[12] Méndez, Eva; Metadatos y Recuperación de información en Internet. Pp. 240 y 241.
[13] DC-Dot ofrece por ejemplo un sistema de edición de metadatos asistido que trata de "adivinar" alguno de los valores de Dublin Core. Accesible en: http://www.ukoln.ac.uk/cgi-bin/dcdot.pl
[14] Literalmente se puede leer acerca del contenido de las páginas que "los editores del sitio pueden manipular mediante metacódigos": http://www.google.es/intl/es/corporate/tech.html

La consecuencia práctica que podemos obtener de todo esto es que independientemente de que se haga manual o automáticamente sigue siendo muy deseable para mejorar la recuperación de información utilizar metadatos. Otro problema distinto es qué cosas están capacitados para hacer después los programas que reciben esa información, o qué capacidad de comprensión sobre diferentes conjuntos de metadatos pueden llegar a tener.

Los documentólogos perciben además algunos de los problemas más acuciantes de Internet como problemas de carácter lingüístico, problemas de representación. Se sabe que en un sistema de recuperación de información sobre una colección de documentos, estos deben estar representados de alguna manera dentro del sistema. Una estrategia ideal de representación debe capturar el contenido informativo de los documentos de la forma más exacta posible así como identificar documentos de contenido similar y diferenciar documentos de contenido distinto. Los sistemas de recuperación de información carecen de la perspicacia del proceso de la mente humana que es capaz de discernir el significado de un texto y deben solventar esto mediante estrategias de representación basadas en el análisis de las características de la aparición de las palabras (frecuencia, relación, posición). Muchos de los enfoques de representación asumen la capacidad de las palabras para transmitir y discriminar significado y presuponen que la frecuencia de aparición de una determinada palabra es indicio de su importancia significante respecto al texto[15].

Los buscadores de Internet, como los sistemas tradicionales de recuperación de información, representan las colecciones de documentos por medio de índices inversos. Estos índices los construyen con las palabras extraídas de cada documento (página web, documento pdf...etc) por un sistema de indización automática que aplica a cada sitio unos determinados algoritmos: analiza la frecuencia de aparición de las palabras, elimina palabras vacías, agrupa palabras con la misma raíz, analiza y valora el lugar de aparición de cada palabra (título, resumen, metadatos) y propone, finalmente, unos términos ponderados, es decir que tienen, cada uno de ellos, unos valores positivos. Estas ponderaciones determinarán en parte el orden de recuperación de los documentos, ya que la idea final es la de proporcionar al usuario un listado ordenado de recursos pertinentes, un *ranking*. También tendrán en cuenta la importancia relacional del documento y los documentos más citados, o más enlazados, serán considerados más importantes[16].

Falta de normalización en los conceptos

Queda claro no obstante que independientemente del paradigma en el que se base el algoritmo de indización, o las heurísticas y parámetros propios del indizador, todo reposa en la representación de los documentos, que siempre se hace sobre la base del lenguaje natural. Ahora bien, el lenguaje no es en absoluto una herramienta de comunicación precisa. A menudo, un concepto puede ser expresado de muchas formas distintas y salvo que haya algún mecanismo que las reconozca sinónimas para el sistema de recuperación son diferentes. También hay muchos términos que expresan distintos significados, a veces muy distantes y esto tampoco es detectado por el sistema.

Para resolver esta ambigüedad en la interpretación algunos sistemas pueden utilizar un vocabulario controlado - un conjunto de elementos léxicos que representan conceptos específicos- y así se reduce el número de significados posibles. Se puede utilizar un vocabulario controlado para indizar manualmente o se puede implementar un *thesaurus* unido a un sistema de normalización automática que transforme las palabras incontroladas en términos controlados y válidos. Esta posibilidad, no obstante, no se utiliza (hablando en los términos más optimistas) de forma habitual en Internet. En cualquier caso sería necesario utilizar alguna fuente unificada de normalización para que ésta pudiera ser tenida en cuenta por cualesquiera sistemas de recuperación.

[15] Esta es una simplificación de los modelos existentes, que pueden ser mucho más complejos. Véase Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier. Modern Information Retrieval. New York : ACM; Harlow, Essex : Addison-Wesley Longman, 1999

[16] El algoritmo que mejor representa esta tendencia es el pionero PageRank de Google, cuya descripción general se puede obtener en: http://www.google.es/intl/es/corporate/tech.html

10 de julio 2005 • Volumen 6 Número 7 • ISSN: 1067-6079

Tampoco existe siquiera la posibilidad de desarrollar un lenguaje documental atómico que resuelva todas las posibles necesidades de normalización que se plantearían en las distintas iniciativas de metadatos. Esto supone que la representación de documentos adolece siempre de problemas de normalización, ya que dado el hecho de que no existe ningún estándar aceptado ni usado en este sentido, ningún buscador llega siquiera a tener en cuenta este factor a la hora de indizar los documentos sobre los que tiene cobertura.

Otro aspecto susceptible de mejora es el hecho de que la mayor parte de los ingenios de búsqueda no indizan los sitios Web de forma completa sino que se limitan a recorrer dos o tres niveles jerárquicos. Así pierden documentos muy significativos que en sitios muy grandes o complejos (organismos públicos, universidades) pueden estar situados en un nivel de jerarquía bajo. Se podría crear localmente un depósito de metadatos accesible en un nivel jerárquico alto[17], aunque por otra parte este es un problema fundamentalmente vinculado a las incompatibilidades de las arañas de búsqueda con los sitios Web dinámicos. Por tanto podríamos hablar de dos problemas claramente distinguibles desde el punto de vista de la documentación y ciencias afines:

- La falta de descripciones explícitas y formalizadas de los recursos (los metadatos no se aplican de forma generalizada y el tratamiento que los buscadores hacen de los recursos no parece poder equipararse en riqueza semántica al de los metadatos).

- Falta de sistemas de normalización homogéneos y de amplia utilización.

Seguramente estos problemas sólo se resolverían si la Red cambiara radicalmente, cosa que por otra parte está en proyecto.

WEB SEMÁNTICA[18]

Tim Berners-Lee define la Web Semántica como "… una extensión de la Web actual en la que se proporciona a la información un significado bien definido, lo que permite a la gente y las computadoras trabajar en cooperación."[19] Por tanto se trata de una nueva concepción de la WWW en la que el significado de las cosas y la capacidad de hacer inteligible este significado a máquinas juegan un papel esencial. Desde un punto de vista tecnológico la Web Semántica gira entorno a una serie de pilares[20]:

- Resource Description Framework (RDF[21]). RDF proporciona un modelo de datos y una sintaxis para expresarlos que pretenden posibilitar la codificación de modelos complejos de metadatos en forma legible por máquina y maximizar al mismo tiempo la interoperabilidad de dichos metadatos. RDF es la base de la Web Semántica, y está basado en XML, aunque no es propiamente XML[22].

[17] Esta es por ejemplo una de las ideas presentes en el modelo de diseminación de metadatos de OAI que ya mencionamos.

[18] El principal foco de la actividad entorno a la Web Semántica es el mantenido por el W3C http://www.w3.org/2001/sw/Activity

[19] Traducido de: Berners-Lee, Tim; Hendler, James; Lassila, Ora. The Semantic Web. Scientific American. 2001. http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2

[20] No haremos mención explícita de otras iniciativas importantes como SPARQL http://www.w3.org/TR/2005/WD-rdf-sparql-query-20050419/ o SKOS-CORE http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/

[21] Los documentos normativos sobre RDF se encuentran en: http://www.w3.org/RDF/ Existe una traducción al Castellano de Eva Méndez en: http://rayuela.uc3m.es/~mendez/RDF/syntax/REC-rdf-syntax.htm Este documento cubre el modelo y la sintaxis RDF.

[22] La forma de serializar RDF, XML/RDF, sí está expresada en XML, pero RDF es más amplio que la sintaxis en que se materializa. Una perspectiva interesante y clarificadora acerca de este particular la encontramos en: Mazzochi, Stefano. A No-Nonsense Guide to Semantic Web Specs for XML People. http://www.betaversion.org/~stefano/linotype/news/57/

[23] La traducción al Castellano de Eva Méndez está disponible en: http://rayuela.uc3m.es/~mendez/RDF/schema/CR-rdf-schema.htm La versión oficial del W3C se puede encontrar en: http://www.w3.org/TR/rdf-schema/

- RDF*Schema*[23]. RDF es sólo un marco general para la descripción de recursos, pero es necesario especificar y restringir las descripciones para que se adapten a las necesidades concretas de aplicación de metadatos propios de diversos dominios. Los *Schemata* RDF permiten la creación de vocabularios específicos para campos de aplicación concretos.

- OWL Ontology Web Language[24]. OWL es un lenguaje para construir ontologías. En breves palabras una ontología se puede definir como un conjunto de especificaciones de conceptos que son utilizables por máquinas. Se podría caer en la tentación de compararlas con los familiares *thesaurus*, pero no lo haremos, porque estructuralmente son diferentes, a pesar de poseer relaciones y jerarquías, y porque lo que hace de las ontologías (al menos de las generadas con OWL) una herramienta excelente para el ámbito de la Web Semántica es la posibilidad de utilizarlas para hacer inferencias lógicas y "mapear" dominios conceptuales distintos para poder reutilizar descripciones fuera del ámbito en que fueron previstas.

Como se puede apreciar la Web Semántica tiene una importante relación con los metadatos. El consorcio W3C comenzó sus actividades sobre los metadatos en el ámbito del "Technology and Society Domain"[25], a partir del cuál se vehicularon un conjunto de actividades y grupos de trabajo que seguirían su posterior desarrollo en el marco de la actividad sobre la Web Semántica (Semantic Web Activity[26]) del Consorcio Web.

En el marco de este primer ámbito de actuación (el de los metadatos), se declara la necesidad de completar una parte importante de la tecnología Web, aquella que hace referencia al etiquetado, catalogación e información descriptiva en una forma que permita que las páginas Web sean buscadas y procesadas adecuadamente, fundamentalmente por computadoras[27].

De esta forma las preocupaciones del consorcio acerca de los metadatos se dirigieron rápidamente a procurar la mejor forma de modelar y codificar los metadatos para que pudieran servir para el fin antes mencionado, su utilización por máquinas. De entre los primeros proyectos encabezados por el W3c en este sentido destacan PICS[28], y RDF[29]. PICS estaba mucho más cerca de una iniciativa de metadatos común, que se preocupaba por etiquetar adecuadamente los recursos para permitir el control de contenidos a terceros, pero RDF iba bastante más allá, y se convirtió, como ya hemos visto en la espina dorsal de la Web Semántica. RDF es bastante complejo, pero en origen surge para modelar metadatos de forma que estos sean legibles e intercambiables por máquinas.

Habíamos planteado algo más arriba, dos problemas que nos parecían fundamentales en lo referente a la recuperación de información en Internet, y por lo que se puede apreciar ambos caen naturalmente entre los problemas que la Web Semántica podría solucionar.

La utilización de cualquier modelo de metadatos es viable, y su reutilización por parte de buscadores es posible gracias a su codificación en XML/RDF y la utilización de los lenguajes de consulta adecuados, singularmente SPARQL. De hecho la posibilidad de expresar mediante RDF*Schema* conjuntos de metadatos específicos también trae implícita la posibilidad de combinarlos de forma que se adapten a una aplicación concreta sin que estos pierdan su significado original.

---

[24] Para una introducción a OWL se puede visitar el documento oficial del W3C en: http://www.w3.org/TR/owl-features/ El sitio Web del grupo de trabajo en Ontologías Web se puede visitar en: http://www.w3.org/2001/sw/WebOnt/#Current

[25] El trabajo sobre metadatos del W3c se puede consultar en: http://www.w3.org/Metadata/

[26] Página del Semantic Web Activity Statement en : http://www.w3.org/2001/sw/Activity

[27] Metadata Activity Statement, documento en el que se declaran los objetivos del consorcio en este ámbito y el estado de la cuestión en el momento de su publicación. http://www.w3.org/Metadata/Activity.html

[28] Sitio web de PICS: http://www.w3.org/PICS/

[29] Sitio web de RDF: http://www.w3.org/RDF/

[30] http://wordnet.princeton.edu/

10 de julio 2005 • Volumen 6 Número 7 • ISSN: 1067-6079

Por otra parte podemos utilizar Ontologías para "mapear" el significado de diferentes conjuntos terminológicos, de forma que a largo plazo se podría solucionar el problema de la normalización desde un punto de vista descentralizado, en lugar de los intentos centralizados precedentes como WordNet[30]. Esto nos proporcionaría la capacidad de eliminar gran parte de la ambigüedad semántica existente en los documentos Web.

METADATOS Y WEB SEMÁNTICA: ALGUNOS EJEMPLOS CONCRETOS

Uno de los frutos visibles de la interacción entre la actividad en Web Semántica y la actividad en Metadatos es la relación existente entre RDF y Dublin Core. Actualmente el conjunto de elementos DC se encuentra codificado en RDF y RDF*Schema*[31]. Esto hace posible, por ejemplo que Dublin Core sea reutilizable en el ámbito de photoRDF, una iniciativa para el desarrollo de metadatos de descripción de fotografía que combina el *squema* DC con otros dos *squemata* que complementan las capacidades descriptivas necesarias para la fotografía[32]. Otra de las aplicaciones posibles es la de mezclar DC y MPEG7 para el tratamiento de documentos audiovisuales, área en la que también se está investigando[33]. Como se puede apreciar la capacidad de reaprovechamiento del trabajo es bastante importante.

Lo expuesto anteriormente perfila DC como una buena elección para la descripción de recursos en Internet, aunque DC tiene inicialmente un problema importante, y es que su interoperabilidad es posible por la sencillez y generalidad de las descripciones propuestas. Esto hace que el conjunto inicial de elementos DC sea de difícil aplicación en ámbitos que requieran mayor precisión y riqueza descriptiva. Sin embargo DC prevé además un mecanismo para solucionar estos problemas, los calificadores. El uso de Dublin Core Calificado[34] en documentos RDF le dota de una expresividad mucho mayor, a través de los *refinadores de elementos y esquemas de codificación*, respectivamente calificadores que especifican todavía más el significado del elemento (como título y título alternativo) y calificadores que permiten limitar los valores de los elementos, sus contenidos.

Estos últimos proporcionan una vía de expresión sencilla para los lenguajes documentales tradicionales como los sistemas de clasificación y los tesauros que pueden solucionar los problemas de normalización en ámbitos restringidos, aunque serán completamente ignorados por los buscadores. Esto nos lleva a una característica muy interesante de Dublin Core, lo que se denomina el principio Dumb-Down, por el cuál cualquier elemento DC cualificado debe poder ser interpretado como uno no cualificado. Esto facilita la interoperabilidad del conjunto, pero además nos permite (ignorando los calificadores recomendados) crear calificadores propios que se adapten a nuestras necesidades específicas, de forma que desde el punto de vista del régimen interno tengamos riqueza descriptiva, y desde el punto de vista externo capacidad de interoperar.

En conjunto Dublin Core nos permitiría el paso a la Web Semántica de forma poco traumática y es probablemente la iniciativa de metadatos más prometedora hoy por hoy si nuestras necesidades descriptivas no son demasiado exigentes o no tenemos problemas para mezclar *squemata* o crear nuestros propios calificadores.

---

[31] Ver http://dublincore.org/documents/dcq-rdf-xml/index.shtml para una guía completa sobre cómo codificar DC en RDF. Las referencias a los *Squema* se pueden encontrar en: http://es.dublincore.org/schemas/rdfs/index.shtml

[32] PhotoRDF es actualmente un proyecto en evolución, pero los resultados preliminares parecen muy prometedores. http://www.w3.org/TR/photo-rdf/

[33] Véase el interesante trabajo de Hunter, Jane. An Application Profile which combines Dublin Core and MPEG-7 Metadata Terms for Simple Video Description. http://www.metadata.net/harmony/video_appln_profile.html

[34] http://es.dublincore.org/documents/2002/05/15/dcq-rdf-xml/index.shtml

CONCLUSIONES

La utilización de metadatos puede contribuir a resolver algunos de los problemas más evidentes de la recuperación de información en Internet, pues introduce elementos de valoración contextual y de significado y reduce significativamente la ambigüedad de las demandas de información y si el ámbito de expresión es lo suficientemente rico e interoperable su utilización será fundamental para el futuro de Internet.

BIBLIOGRAFÍA

Antoniou, Grigoris; Harmelen, Frank van. *A Semantic Web Primer*. Cambridge: The MIT Press. 2004.

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier. *Modern Information Retrieval*. New York: ACM; Harlow, Essex : Addison-Wesley Longman, 1999.

De Jong, Annemieke. *Los metadatos en el entorno de la producción audiovisua*l. 2ª ed. México: Radio Educación. 2003.

Méndez Rodríguez, Eva María. *Metadatos y recuperacion de información : estándares, problemas y aplicabilidad en bibliotecas digitales.* Gijón : Ediciones Trea, D.L. 2002