Aplicación de tesauros, taxonomías y ontologías en los sistemas de gestión de contenidos mediante tecnologías de la Web Semántica

Applications of thesauri, taxonomies and ontologies in content management systems with Semantic Web technologies

Juan Antonio PASTOR SÁNCHEZ (1) y Francisco Javier MARTÍNEZ MÉNDEZ (2)

(1) (2) Facultad de Comunicación y Documentación de la Universidad de Murcia, Campus Universitario de Espinardo-30100 Murcia, pastor@um.es (2) javima@um.es

Resumen

Los Sistemas de Gestión de Contenidos (CMS) representan el paradigma actual en el diseño y desarrollo de sistemas y servicios de información basados en la Web. Uno de sus principales retos es poder gestionar información al mismo tiempo que se implantan estructuras coherentes de acceso a la misma, con el fin de no poner trabas al acceso a la información y evitar la desubicación de los usuarios en sus procesos de navegación para la recuperación de información. Esto precisa del uso de herramientas próximas a la organización del conocimiento, desarrolladas generalmente mediante la implementación de taxonomías o, de un modo más flexible, de tesauros. Ambos tipos de sistemas han sido utilizados para la clasificación y descripción del contenido de los documentos. Sin embargo, también resulta muy interesante emplearlos en la construcción automática de sistemas de navegación dinámicos, procedimiento que puede llevarse a cabo por medio del uso de ontologías, que, adicionalmente, podrían emplearse para interrelacionar o inferir los contenidos gestionados en un CMS con aquellos ubicados en los sistemas de información corporativos. Se muestran diversas formas en las que pueden combinarse algunas tecnologías surgidas asociadas a la representación de tesauros y ontologías en el ámbito de la Web Semántica (como SKOS y OWL), con la finalidad de exponer una serie de planteamientos que orientarían el desarrollo futuro de sistemas CMS, persiguiendo la obtención de esa mayor coherencia de la estructura informativa.

Palabras clave: Tesauros. Taxonomías. Web Semántica. Arquitectura de la Información. Gestión de la Información. Recuperación de la Información

Abstract

The Content Management Systems (CMS) represent the current paradigm in the design and development of information systems and services based on the Web. One of the major challenges facing this kind of applications is to combine, at the same time, the information management processes with the implementation of a coherent structure for access to the data. This requires the use of tools near to the Knowledge Organization, usually by implementing taxonomies or, more flexibly, thesauri. Both of them have been traditionally used for classification and content description of documents. However, it is also very interesting to use them into the automatic construction of dynamic navigation systems, a procedure that can be done through the use of ontologies that, additionally, could be used to interrelate or infer the content managed in a CMS with that content located in corporate information systems (intranets). Diifferent ways to combine these emerging technologies are shown, associated with the representation of thesauri and ontologies in the Semantic Web (such as OWL and SKOS), with the aim to present a series of approaches that could guide future developments of CMS systems in order to obtain greater consistency in information structures.

Keywords: Thesauri. Taxonomies. Semantic Web. Information Architecture. Information Management. Information Retrieval.

1. Introducción

El inicio del nuevo milenio trajo consigo un nuevo paradigma en los procesos de elaboración de páginas Web: los Sistemas de Gestión de Contenidos (CMS). Anteriormente se utilizaban de forma exclusiva editores HTML que incorporaban (algunos de ellos) características de gestión de sitios Web y sus ficheros asociados. Comparada con la actual, era una forma prácticamente artesanal de elaborar contenidos que exigía (y exige) un conocimiento de diferentes lenguajes para el marcado de documentos o la definición de estilos visuales. En el caso de elaborar pági-

nas dinámicas a partir de información estructurada en base de datos se hacía necesario incluso manejar con cierta soltura algún tipo de lenguaje de programación.

Otra de las dificultades que plantea la gestión de sitios Web de cierto tamaño es la gestión colaborativa del mismo. Desde su origen hasta la actualidad, la Web se basa en un modelo donde los roles de lector y autor están claramente separados. Esta situación obliga a que la edición de una página Web se realice en el entorno "local" de la propia estación de trabajo del usuario y posteriormente se transfiera al servidor para su publicación. Algunas tecnologías permiten definir unidades de red o sistemas de ficheros que apuntan al sistema de ficheros del servidor dentro de la estación del usuario, editando directamente sobre el servidor de forma transparente al usuario. Dicha solución requiere el uso de software adicional o la configuración de conexiones para el acceso a sistemas de ficheros remotos. Se trata, en definitiva, de un modelo de edición y publicación que implica la instalación y configuración de herramientas de edición y publicación de contenidos, limitando la movilidad de editores por la disponibilidad de determinadas aplicaciones para determinados sistemas operativos.

La simplificación de los procesos indicados anteriormente motivó la aparición de los CMS (Tramullas y Garrido, 2006). Estos sistemas permiten, por un lado, editar de un modo más sencillo los contenidos del sitio Web, y por otro facilitan la ubicuidad de las herramientas de edición. La estructura de permisos o roles resulta fundamental en los CMS. De esta forma, se establecen diferentes niveles de gestión del sitio Web coordinados con la ayuda de herramientas de flujo de trabajo (Serrano-Cobos, 2007, p. 214). El más básico permite al usuario concentrar su esfuerzo en las tareas de organización, categorización, descripción, edición y creación de contenidos y documentos. Otros niveles más avanzados se dirigen a la configuración del sistema, diseño de estilos visuales, tipos de contenido, servicios adicionales, etc. Desde el punto de vista del usuario, la proliferación y uso de los CMS hacen más asequibles las tareas asociadas a los procesos de diseño y mantenimiento de sitios Web (Ambite et al., 2006).

La aportación de los CMS no se detiene en los aspectos más cercanos al usuario. La mayor parte de estos sistemas utilizan bases de datos y permiten diseñar estructuras de información para los contenidos. Además, implementan mecanismos para la organización de los mismos y la integración de recursos externos. Se han creado múltiples tecnologías (ASP, PHP o JAVA entre otras) para el desarrollo de aplicaciones Web. Una de las claves de su gran avance ha sido su integración con sistemas gestores de bases de datos (MySQL o PostgreSQL principalmente). El acceso a estas tecnologías se ha producido en el ámbito del software libre, siendo precisamente este hecho lo que ha favorecido su uso y creación de herramientas con un grado de elaboración cada vez mayor. Los CMS no han sido ajenos a esto, de hecho, muchos de los contenidos visibles actualmente en la Web son el resultado de páginas dinámicas soportadas en bases de datos.

Este trabajo intentará clarificar el modo en el que los CMS pueden integrarse dentro de un sistema de información corporativo, así como el papel que las tecnologías de la Web Semántica pueden desempeñar para ello.

2. El paradigma de Gestión de Información en los CMS

Durante los últimos diez años hemos asistido a cambios importantes en los modelos funcionales de los CMS. El almacenamiento estructurado de HTML y la gestión de ficheros está dando paso a otro tipo de técnicas basadas en procesos propios de gestión de la información. Esto parece lógico puesto que la gestión colaborativa de sitios Web corporativos de cierto tamaño está más cerca de dicho proceso que del simple diseño visual. Resulta de gran importancia la incorporación a los CMS de herramientas de definición de diferentes tipos de contenidos, la aplicación de metadatos o la creación y aplicación de taxonomías y clasificaciones. Todo ello ha permitido un salto cualitativo trascendental en el planteamiento de los CMS. De esta forma se han reforzado las tesis que apuntan a la organización de los contenidos como uno de los aspectos clave en estos sistemas. También se ha superado la visión de los contenidos Web como un conjunto heterogéneo de páginas enlazadas entre sí. Ahora se tiende a la definición de estructuras que establecen una tipología de contenidos con una semántica más explícita y cercana al concepto de metadatos (Horrocks et al., 2003).

Se observa claramente una tendencia en la estructuración y organización de contenidos que permitiría un desarrollo más eficaz de la Web Semántica. Precisamente los CMS están conformándose como un marco para la creación y mantenimiento de servicios y productos de información basados en Web. Los desarrollos tecnológicos basados en "portlets" o mensajes SOAP proporcionan un medio para aunar diferentes servicios Web (Alonso et al., 2004 y Bo-

oth et al., 2004). Un CMS que incorpore estas funcionalidades permite aglutinar y definir servicios y portales de información. Dicha definición adopta la forma de un conjunto integrado y estructurado de recursos de información cuyo contenido variará según ciertos parámetros.

El escenario descrito anteriormente se completa planteando un alcance funcional de los CMS más allá de la gestión de contenidos Web. Estas herramientas extienden los sistemas de información integrales (Baldomero, 2006). Surge así una nueva línea de trabajo que permite cubrir la separación existente entre la gestión de información, gestión documental y explotación de bases de datos corporativas por un lado y la gestión de contenidos Web, de otro. En definitiva, la gestión de información corporativa se reutiliza para los contenidos Web a partir de la definición de servicios y productos mediante el marco de diseño y desarrollo que proporcionan los CMS.

2.1. Definición e integración estructural de contenidos

Han quedado atrás los tiempos en los que los CMS únicamente permitían la gestión de contenidos sin ningún tipo de estructura. La superación de paradigma de página Web ha sido posible gracias a los CMS y la aplicación de esquemas de organización similares a los de las bases de datos. A pesar de que algunos tipos de estos sistemas, como los Wikis —donde el esquema de acceso a la información es el hipertexto, utilizado de forma natural en el propio contenido de un documento— siguen utilizando este paradigma, es cada vez más frecuente encontrarse con herramientas de este tipo que estructuran el contenido en varios campos. Incluso Wordpress, desarrollado a partir de un modelo para el mantenimiento de weblogs, incorpora en sus últimas versiones la posibilidad de estructurar las entradas añadiendo campos descriptivos.

El objetivo final es alcanzar una caracterización más detallada del contenido que la proporcionada por la clásica dualidad “título-cuerpo del documento”. Algunos CMS poseen una perspectiva más avanzada a través de la definición de diferentes tipos estructurados de contenido con una amplia variedad en la tipología de campos.

Los elementos gestionados en el sistema de información corporativo pueden reutilizarse para su consulta en la Web y desempeñar una función esencial. Para ello, es necesario definir una correspondencia entre los tipos de contenidos definidos por medio de un CMS y aquellos almacenados en bases de datos relacionados y documentales.

Es aquí donde entran en juego las especificaciones de metadatos, porque aportan un modelo para la descripción de contenidos y su posterior intercambio. Desde este punto de vista hay que indicar que las tecnologías de la Web Semántica están incorporándose poco a poco a la arquitectura funcional de los CMS. Un ejemplo de ello es la sindicación RSS como solución a la optimización en la reutilización de contenidos.

Por otro lado, es factible unir los procesos de gestión de información y documentación y de gestión de contenidos Web. El planteamiento anterior precisa de mecanismos para definir los ámbitos de publicación porque no toda la información gestionada por la organización, ya sea documental o de otro tipo, es susceptible de ser consultada en la Web. La explotación de las bases de datos corporativas para la publicación o integración de contenidos Web requiere la identificación automática de aquella información destinada a ser reutilizada a tal efecto.

2.2. Organización de contenidos

Los contenidos además de ser estructurados internamente han de ubicarse en una macroestructura conceptual. Para ello, se emplean instrumentos tales como las taxonomías y clasificaciones. Los CMS utilizan generalmente esquemas jerárquicos para la organización de contenidos. La capacidad de estas estructuras para definir el grado de generalización o especialización de unos contenidos con respecto a otros resulta de gran utilidad en la Web (Gillrich, 2003). Además, es posible aplicar los tesauros para esta función, ya que aportan estructuras asociativas que complementan las jerarquías de conceptos.

Se trata por tanto de aplicar un proceso similar a la indización, asignando los elementos de un esquema conceptual a contenidos o elementos informativos, si bien su utilidad no se limita a la posterior recuperación de información. También es posible utilizar las estructuras jerárquicas o asociativas de estos instrumentos para deducir relaciones entre contenidos. De este modo es posible inferir que un documento es más específico o genérico que otro a partir de su indización y la estructura del lenguaje documental. Esto aporta información de valor añadido a los propios contenidos.

Además de organizar los contenidos para su localización durante la consulta por parte de los usuarios, estos sistemas también constituyen

una valiosa herramienta para la gestión de los mismos. La mayoría de los CMS aplican el modelo autor-lector de forma que el mantenimiento de los contenidos puede realizarse simultáneamente a su consulta. El uso combinado de varios esquemas conceptuales permite caracterizar diversos aspectos conceptuales de un documento. Es posible utilizar un vocabulario para la descripción del contenido semántico de un elemento y combinarlo con otros que describan la tipología, formato o los niveles de acceso, difusión y gestión. Por tanto, se definen diferentes facetas de descripción de un elemento, ampliando el alcance de la indización conceptual y proponiéndose estructuras alternativas de acceso a la información.

Las estructuras de organización no se emplean únicamente a los contenidos Web. Lo ideal sería que tanto el sistema de información corporativo como el CMS utilizado apliquen los mismos esquemas conceptuales para la organización y recuperación de información. Esto aporta mayor coherencia e integración entre los contenidos publicados en la Web y los del resto del sistema de información (Cobo Romaní, 2005, p. 186).

2.3. Sistemas de navegación dinámicos

Uno de los aspectos más problemáticos en la gestión de contenidos con CMS es la necesidad de implementar sistemas de navegación. La localización de información en entornos de bases de datos relacionales o documentales se basa en la selección terminológica y la construcción de consultas para la ejecución del proceso de búsqueda. Sin embargo, el acceso a la información en la Web mediante hipervínculos precisa de sistemas de navegación bien diseñados y útiles.

La eficiencia de un sistema de navegación está ligada su capacidad de ofrecer una panorámica general de los contenidos a los que da acceso. Esta doble función de los sistemas de navegación se ve dificultada por la actualización de los contenidos. En efecto, el incremento del volumen de información supone un descenso progresivo en la capacidad de adaptación del sistema de navegación para dar acceso a los contenidos. Las complicaciones de gestión inherentes al rediseño de los sistemas de navegación y la reubicación de contenidos suponen un serio handicap para el mantenimiento de un sitio Web.

Resulta lógico que un sistema de navegación tenga un carácter dinámico, dependiendo del contenido global del sistema. Esto no quiere decir que deba modificarse cada vez que se inserte un nuevo documento. Nuestra propuesta parte de la aplicación de esquemas conceptuales en la indización de los contenidos. Un tesauro constituye un marco conceptual para la gestión de un sistema de información, pero pueden emplearse de forma conjunta con taxonomías para la caracterización facetada de recursos. Estos instrumentos tienen el inconveniente de no poder ser utilizados directamente para la construcción de estructuras de navegación, debido a su complejidad para el usuario final. Existe pues, un cierto vacío funcional entre las clasificaciones o tesauros utilizados para indizar los contenidos y los menús de navegación globales o locales de un sitio Web. Para cubrir esta separación pueden definirse ontologías que definan reglas para construir la estructura y el ámbito de aplicación de los sistemas de navegación. Este tipo de ontologías parte de una definición en la que intervendrían los siguientes elementos o propiedades:

- Propiedades generales del sistema de navegación: denominación, ámbito o ámbitos de alcance (global a todo el sitio o local en un subsitio o micrositio)
- Elementos (conceptos) de uno o varios esquemas que intervienen en la ontología
- Correspondencia entre el etiquetado del sistema de navegación y el etiquetado léxico de los conceptos utilizados en la ontología.
- Definición de jerarquías y relaciones entre sistemas de navegación. Mediante este mecanismo es posible construir (y mantener) la estructura completa de los diversos sistemas de navegación de un sitio Web.

Por tanto, es evidente que una ontología permite, a partir de esquemas conceptuales comunes, definir la correspondencia entre las estructuras conceptuales de gestión y las de difusión de los contenidos informativos. Es imprescindible contar con un elenco de tecnologías capaces de implementar y llevar a la praxis las propuestas esbozadas anteriormente. En consecuencia, mostraremos el papel que pueden desempeñar en este trabajo las tecnologías más representativas de la Web Semántica.

3. Aplicación multinivel de tecnologías de la Web Semántica

En contra de la opinión mayoritaria, la Web Semántica constituye una realidad. El éxito de las diversas tecnologías que proponen esta nueva funcionalidad de la Web reside en la transparencia de su aplicación (es un hecho cierto que la mayoría de los usuarios desconocen que la Web conforma un medio de gran

eficacia para el intercambio de datos entre aplicaciones).

En la introducción se hacía mención a la importancia que han tenido las bases de datos en el desarrollo de los CMS. Esto ha beneficiado también el desarrollo de la Web Semántica. El alto grado de estructuración, característico de las bases de datos, ha supuesto una abundancia de materia prima para el intercambio de información con un grado de carga semántica muy elaborado. Hasta el planteamiento del concepto de Web Semántica, la Web (e Internet en general) se centraba en la transmisión de "cajas negras", es decir, ficheros sin ningún tipo de descripción formal del contenido. Lo importante era que los archivos llegaran a los destinatarios o solicitantes para su uso o interpretación.

Un breve análisis de las nuevas tecnologías Web muestra la importancia de los metadatos para la descripción de recursos. La clave de los metadatos en la Red es la adopción de modelos compartidos y su codificación mediante especificaciones formales pudiendo intervenir en los procesos de descripción conceptual de recursos de e incluso en la formalización de características lógicas de los mismos (Méndez, 2007, p. 7, p. 63).

3.1. El largo camino de los metadatos: de XML a RDF

Desde la aparición en 1998 de XML (siglas de "Lenguaje de Etiquetado Extensible" propuesto por el W3C (1)) mucho se ha hablado de la importancia de los metadatos. Quizás los factores críticos en su desarrollo y uso sean la elección de un modelo de metadatos y su integración en un marco conceptual de desarrollo adecuado a la Web Semántica. Dublin Core (Méndez, 2007, p. 61) se constituyó como un modelo para la descripción de metadatos (a un nivel básico) siendo adoptado rápidamente para su aplicación en la Web. A las evidentes ventajas de sencillez, simplicidad e independencia sintáctica (puesto que Dublin Core propone un modelo de descripción) hay que sumar su modularidad y su versatilidad más allá de la propia Web.

Este modelo de metadatos se incorpora perfectamente en la Web Semántica a través de su serialización RDF. Éstas y otras especificaciones como FOAF, SIOC o RSS, aportan mayor potencia de descripción semántica a la información gestionada mediante CMS. Las estructuras de bases de datos con los que suelen operar estos sistemas encuentran un camino para la comunicación de información entre aplicaciones.

RDF (siglas de "Marco de Descripción de Recursos") ofrece un modelo para la representación de información estructurada en forma de metadatos, para describir recursos disponibles en Internet (o de otro tipo). La descripción de recursos se realiza en forma de tripletas del tipo "sujeto-predicado-objeto". De esta forma el sujeto es el recurso a describir, el predicado es una propiedad o relación del recurso y el objeto es el valor asignado a esta propiedad o el recurso con el que establece la relación. Sujeto, predicado y objeto hacen referencia a un recurso identificado con una URI. El objeto puede hacer referencia a una URI o tener asignado un valor de cadena literal o un literal tipificado cuyo dominio es definido en otro recurso.

Esto permite aprovechar otros vocabularios XML e integrarlos dentro de declaraciones RDF. Adicionalmente puede aplicarse RDFS, extensión semántica de RDF que permite describir esquemas sencillos usando clases, subclases, propiedades, dominio de aplicación y rango de valores. La integración anteriormente mencionada permite la incorporación de las ventajas de RDF y RDFS en especificaciones que utilicen este marco descriptivo e incide en el incremento de su capacidad. Por otra parte, las experiencias de Kabish y Neiling (2005) han demostrado que esta tecnología, junto con el uso de RDQL (un lenguaje de interrogación para RDF) permite redefinir las estrategias de integración de datos, abriendo nuevas perspectivas a los sistemas de almacenes de datos y a los procesos de minería de datos.

La sencillez del planteamiento de RDF y RDFS dota a la Web Semántica de un marco descriptivo fácil de implementar en la arquitectura de los CMS. A modo de ejemplo, es común el uso de RSS con el objetivo para la sindicación de contenidos, propiciando el intercambio de información entre sistemas. De este modo, diversas bases de datos pueden combinarse mediante RDF, ofreciendo resultados que alcanzan un alto grado de interoperabilidad semántica. Consiguientemente los CMS disponen de una herramienta para el mapeado de contenidos hacia una especificación formal, tomando como fuentes las bases de datos de contenidos.

Si bien el primer objetivo de los metadatos es la descripción formal de recursos de información, desde una perspectiva más amplia, observamos que la arquitectura técnica empleada para ello puede ser aplicada en el desarrollo de soluciones de mayor alcance. La definición y representación de aspectos conceptuales y de atributos de la lógica de las relaciones entre recursos constituye un paso adelante que explota los metadatos en un nivel de mayor abstracción.

3.2. Esquemas conceptuales con SKOS: simplificando lo complejo

Tal como se ha visto en el apartado 2.2., la mayoría de CMS suelen emplear algún tipo de esquema conceptual en los procesos de clasificación de documentos y en los de búsqueda de información. Pese a ello, es muy reveladora la reflexión de Pérez Agüera (2004) según la cual los tesauros se han integrado en entornos de gestión y recuperación de información documental, pero no siempre se han utilizado en el ámbito de la recuperación de información automática. Uno de los obstáculos que ha encontrado esta aplicación es la representación de los mismos con un modelo adecuado, aplicando una tecnología que permita la creación, mantenimiento, interoperabilidad, reutilización e incluso la integración de diferentes esquemas de forma sencilla y ágil.

Desde la aparición de XML se ha venido utilizando este lenguaje para desarrollar múltiples trabajos de representación de tesauros y esquemas conceptuales (iniciativas como Zthes, MeSH o Topic Maps han tenido continuidad a lo largo del tiempo). Pero desde hace algunos años la tendencia es usar el modelo propuesto por RDF/RDFS y su correspondiente codificación en XML para este fin. Es una evolución lógica en las líneas de trabajo dentro de la representación de tesauros en la Web Semántica.

La representación de tesauros con RDF/RDFS implica superar el modelo propuesto por las normas ISO 2788:1986 o ANSI/NISO Z39.19 para la construcción y el mantenimiento de tesauros. Estas normas describen un tesauro como un conjunto de términos de diverso tipo que entre los que se establecen relaciones semánticas. Pero la Web precisa de una visión más amplia y flexible que supere la idea de término como elemento central del tesauro y que amplíe el número, tipo y significado de las relaciones (Tudhope, Harith y Jones, 2001). Los tesauros conceptuales, multilingües, con estructuras de agrupación y relaciones semánticas y de correspondencia entre tesauros amplia–bles suponen una necesidad que ayuda a una aplicación eficaz de estos instrumentos.

Es cierto que los trabajos de Otman (1996) o Cruse (2004) aportan análisis muy exhaustivos de las propiedades matemáticas de las relaciones, tanto semántica como terminológicamente. Dichas aportaciones podrían superar la funcionalidad y potencia de los tesauros o taxonomías, a costa de sacrificar aspectos pragmáticos como la sencillez de diseño y mantenimiento.

Por este motivo, con el paso del tiempo las diferentes propuestas de representación de tesauros mediante RDF/RDFS (ILRT, LIMBER, CERES, GEM o AGROVOC, entre otras) han confluido en SKOS (siglas de “Simple Knowledge Organization System”). Se trata de una iniciativa del W3C, aún en desarrollo (2), en forma de aplicación de RDF que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales como listas encabezamientos de materia, taxonomías, esquemas de clasificación, tesauros y cualquier tipo de vocabulario controlado.

Los elementos del modelo SKOS (Miles y Brickley, 2005, Isaac y Summers, 2008, Miles y Bechhofer, 2009), incluidos en la Figura 1 son esencialmente clases y propiedades cuya estructura e integridad son definidas por las características lógicas y las relaciones entre las mismas. En SKOS, un sistema de organización del conocimiento se expresa en términos de conceptos estructurados a través de relaciones que conforman esquemas. Tanto los conceptos como los esquemas se identifican mediante URIs.

Un concepto puede tener asociadas múltiples etiquetas, pero sólo una de ellas por cada idioma puede ser asociada como etiqueta preferente. El resto se denominan etiquetas alternativas, aunque también pueden definirse etiquetas ocultas con la finalidad de ser utilizadas en los procesos de búsqueda e indización sin que sean visibles para los usuarios. Es posible asignar a los conceptos notaciones en forma de códigos de clasificación o de identificación dentro de un esquema conceptual determinado. También pueden documentarse los conceptos con notas de diferente naturaleza como definiciones, notas de alcance o notas de edición entre otras.

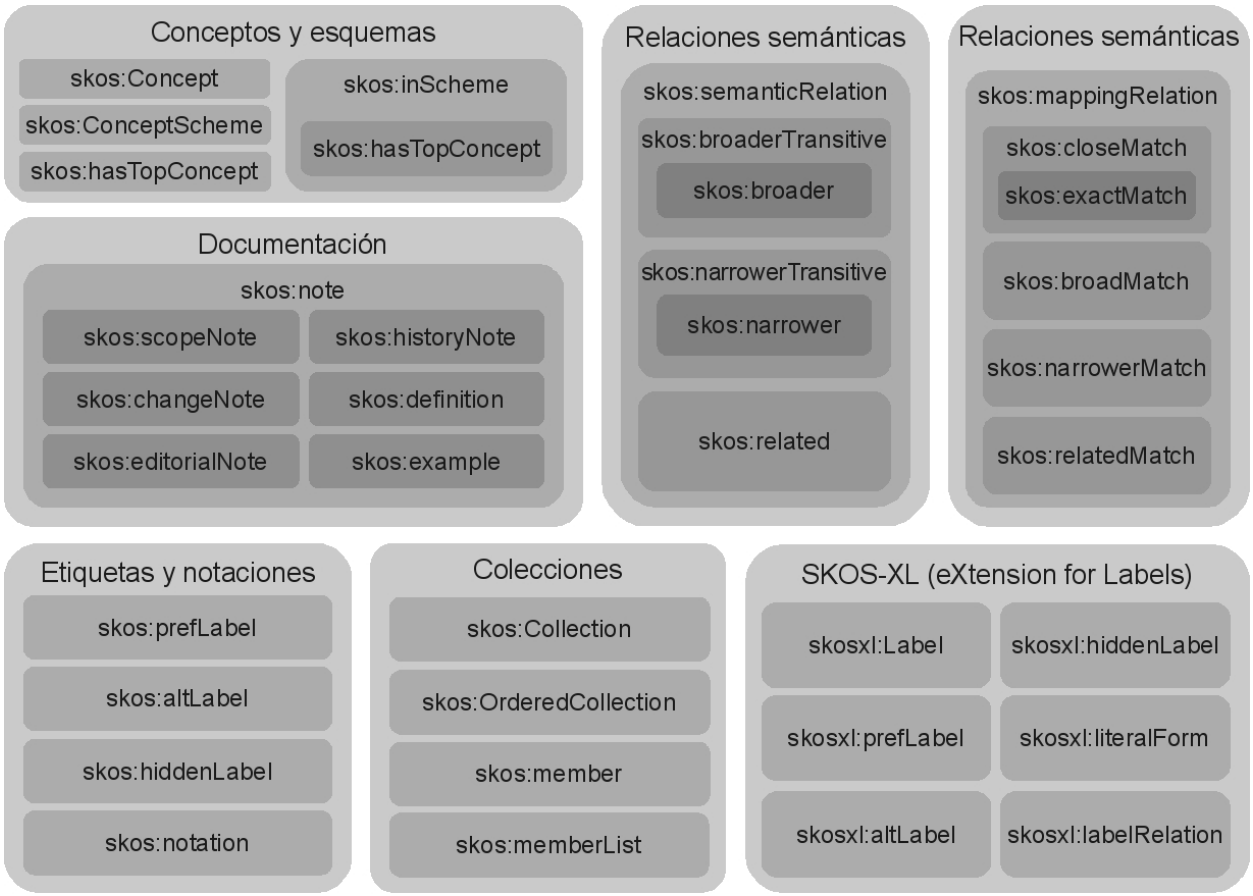

Figura 1

El modelo SKOS contempla el establecimiento de relaciones semánticas entre conceptos de un mismo esquema o de relaciones de correspondencia (mapeado) entre conceptos de diferentes esquemas. Los conceptos también pueden agruparse en colecciones que a su vez pueden etiquetarse y ordenarse. Es posible definir tipologías de relaciones entre etiquetas mediante una extensión denominada SKOS-XL. Para ello es necesario definir las unidades léxicas (etiquetas) a relacionar como recursos accesibles desde sendas URIs, al igual que ocurre con los conceptos, esquemas y colecciones. De esta forma se define un tipo especial de entidad léxica al que se le asigna una cadena literal que puede repetirse para distintas unidades.

El conocimiento descrito de manera explícita como una ontología formal se expresa como un conjunto de axiomas y hechos. Pero un tesauro o cualquier tipo de esquema de clasificación no incluye este tipo de afirmaciones, sino que identifica y describe (con el lenguaje natural o expresiones no formales) ideas o significados a los que nos referimos como conceptos (Arano, 2005) que pueden organizarse en estructuras que carecen de una semántica formal y que no pueden considerarse como axiomas o hechos.

Es decir, un tesauro únicamente proporciona un mapa intuitivo de la organización de temas en los procesos de clasificación y búsqueda de objetos (generalmente documentos) relevantes a un dominio específico. Para convertir un tesauro o esquema de clasificación en conocimiento formal, debe transformarse en una ontología, proceso que resulta muy costoso porque una ontología no proporciona un modelo de datos que se pueda aplicar fácilmente. Esto sucede porque los tesauros se han desarrollado sin una semántica formal, fundamentalmente como herramientas que ayudan en la navegación o en la recuperación de información. No obstante, es factible aplicar OWL (3) para construir un modelo de datos apropiado al nivel de formalización exigido por un tesauro. De hecho, SKOS es en realidad una ontología definida con OWL Full.

El uso de RDF en el desarrollo de SKOS permite obtener documentos preparados para su lectura por aplicaciones informáticas, así como su intercambio y su publicación en la Web. Además su uso conjunto con el elemento dc:subject de Dublin Core permite la indización de recursos de información mediante conceptos.

SKOS se ha diseñado para crear nuevos sistemas de organización o para migrar los ya existentes, adaptándolos a su uso en la Web Semántica de forma fácil y rápida. Proporciona un vocabulario muy sencillo y un modelo intuitivo que puede ser utilizado independientemente o de forma conjunta con OWL. Esto último puede ser útil para expresar formalmente estructuras de conocimiento sobre un dominio concreto ya que SKOS no puede realizar esta función al no tratarse de un lenguaje para la representación de conocimiento formal. Por todo ello, SKOS se considera como un paso intermedio, un puente entre el caos resultante del bajo nivel de estructuración de la Web actual y el riguroso formalismo descriptivo de las ontologías.

3.3. Cerrando el círculo: descripción de aspectos lógicos

En el apartado anterior se ha insistido en las posibilidades de SKOS usado conjuntamente con OWL. Mientras que RDF conforma un modelo general descriptivo, SKOS supone una solución para la caracterización conceptual de los mismos. Es decir, se dispone de una tecnología capaz de representar mediante un modelo compartido aspectos asociados al análisis formal y al análisis del contenido de recursos de información.

Sin embargo este trayecto no concluye aquí. La Web Semántica nos propone un nivel, que supone un grado de evolución mayor, para explicitar las relaciones lógicas entre recursos y elementos de forma estructurada y sistémica.

Para el desarrollo de esta propuesta no se parte de cero. Se vuelve a utilizar RDF como modelo de representación incrementado su capacidad expresiva. Ciertamente, SKOS supone una aproximación muy simple a las propiedades lógicas de las relaciones entre recursos, incluso existen alternativas. Pese a ello, merece la pena explotar al máximo las posibilidades que ofrece la representación de esquemas conceptuales en la Web Semántica y la indización de recursos de información con estos instrumentos.

Las ontologías permiten la obtención y representación de conocimiento a partir de modelos compartidos para su reutilización y compartición (Gómez Pérez et al., 2004, p. 8-9). Este tipo de conceptualizaciones explícitas, tal y como están definidas por Gruber (1993, p. 199) son susceptibles de ser aplicadas en el desarrollo de una capa lógica en los CMS.

Por este motivo, el análisis de la estructura jerárquica o asociativa de un esquema conceptual y la asignación de conceptos a recursos de información permiten realizar inferencias de gran utilidad. A partir de la ubicación en la estructura del esquema de aquellos conceptos utilizados en la indización es posible formular un criterio para la ordenación o agrupación de documentos. Así pues, resulta posible incluir información de valor añadido a un documento o recurso consultado acerca de aquellos recursos con los que se encuentra relacionado jerárquica o asociativamente.

El uso de SKOS en los procesos de recuperación de información puede extender las capacidades de motores de búsqueda genéricos o de las herramientas de búsqueda disponibles en los CMS. Las dificultades de selección terminológica durante la realización de consultas podrían paliarse con el uso de esquemas conceptuales a través de una interfaz gráfica adecuada. En ocasiones dicha selección se desarrolla en un proceso de “ensayo y error” por parte del usuario ya que desconoce las peculiaridades del lenguaje de consulta y el entorno de trabajo. Esta situación se agrava con las peculiaridades léxicas y terminológicas del propio lenguaje natural, tales como sinonimia, polisemia o variaciones lingüísticas entre otras. Resulta evidente que la aplicación del tesauro, o algún otro tipo de esquema conceptual, como herramienta de control terminológico, incrementaría las posibilidades de éxito en las operaciones de búsqueda (Pastor, 2009, p. 129).

Los tesauros y taxonomías pueden aplicarse en los CMS para definir privilegios de acceso y mantenimiento de contenidos. Asignando grupos de usuarios a la consulta o gestión de recursos indizados con determinados conceptos es posible plantear una política de administración más coherente y con un alcance global. De hecho algunos gestores de contenidos (como Drupal) incorporan esta dinámica de funcionamiento. El salto que supondría el empleo de SKOS permitiría compartir y reutilizar estos instrumentos dentro de Sistemas de Información Corporativos.

Como puede verse SKOS plantea unas líneas de trabajo aún por utilizar plenamente, aunque su límite se presupone bastante claro cuando se desea profundizar más allá de la organización y descripción conceptual. Los aspectos lógicos que nos ofrece, aunque útiles, no dejan de ser limitados. Esto resulta obvio si pensamos que uno de los pilares de desarrollo del modelo se fundamenta en su sencillez.

No obstante, SKOS no supone un callejón sin salida, puesto que al final del límite de sus posibilidades se encuentra OWL. No hay que olvidar que el desarrollo y mantenimiento de ontologías

es una tarea bastante compleja. Este es el motivo por el que no resulta factible su uso directamente por parte de cualquier usuario. Por tanto, siguiendo los trabajos de otros autores como Arano (2005) y García Jiménez (2004), pensamos que es posible el desarrollo de técnicas que combinen tesauros y ontologías y su incardinación en la arquitectura funcional de los CMS puede resultar fructífera.

OWL haría posible compartir sistemas de navegación, dotándolos de una estructura lógica formal. Por tanto, se abren nuevas expectativas para que aplicaciones informáticas puedan hacer uso del nivel navegacional de un sitio Web. Relacionar un sistema de navegación elaborado a partir de un esquema conceptual tendría su utilidad en un análisis automático de sitios Web. Los motores de búsqueda (como Google o Yahoo!) analizan los hipervínculos existentes en las páginas Web cuando estas páginas son indizadas y esta información es tomada en cuenta en la ponderación de los resultados de una búsqueda. La explotación de estructuras lógicas y conceptuales de los contenidos de un sitio Web podría ofrecer resultados más eficientes.

También podrían utilizarse para especificar las condiciones de visualización de determinados contenidos. Estas definiciones pueden simplificar los procesos de gestión y actualización de contenidos y la construcción y mantenimiento automáticos de índices de enlaces. Los contenidos podrían adaptarse a partir de determinadas propiedades lógicas definidas a través de ontologías. Dichos mecanismos serían de especial utilidad en la adaptación de los contenidos a perfiles o dispositivos móviles.

Obviamente, este tipo de procesos serían definidos utilizando asistentes que harían transparente al usuario el manejo de ontologías. Los CMS deberían incorporar asistentes para facilitar la definición de estructuras lógicas, asociadas a diferentes tareas. Además de la oportuna reducción de la carga de trabajo de los administradores se formalizarían servicios Web adicionales, superando la simple sindicación de contenidos con RSS. De este modo, se brindaría la oportunidad de aprovechar y reutilizar contenidos utilizando coordinadamente estructuras de navegación y esquemas conceptuales.

4. Reflexiones y conclusiones

Uno de los retos a los que nos enfrentamos hoy día en la consecución de un modelo de gestión de contenidos, más allá de la metáfora de "página Web" con la que funcionan muchos de los CMS actuales es el conseguir incorporar a los tesauros, taxonomías o clasificaciones para describir la descripción del contenido, aproximando un poco más la gestión de contenidos Web a la organización del conocimiento.

Desde el punto de vista corporativo, las organizaciones suelen contar son sistemas de información en donde se opera con bases de datos relacionales o documentales. Actualmente estas aplicaciones suelen diseñarse para su funcionamiento sobre tecnología Web y comparten tecnología con los CMS y, en ocasiones, el mismo objeto de trabajo. Es razonable pensar que los CMS deben formar parte del conjunto de aplicaciones para la gestión de información en el seno de una organización. La reflexión que subyace a esta idea es la de diseñar o mejorar los flujos de información utilizados para la elaboración de contenidos. Es frecuente que la información que se publica en la Web haya seguido un proceso de gestión y edición específico, aunque sorprendentemente, suele estar previamente almacenada digitalmente en otros medios o aplicaciones de la organización. Esta situación está motivada por la carencia de una visión más amplia de la gestión de contenidos Web, un punto de vista que la conciba como parte integral de un sistema de información más extenso y general.

Los CMS ya no se conciben exclusivamente como aplicaciones para la publicación de contenidos. Por el contrario, se consideran como marcos integradores de servicios y de contenidos, capaces de reutilizar y agregar los datos y objetos documentales gestionados por otras herramientas. Las ventajas son evidentes, aunque no debemos olvidar que se trata de una vía de doble sentido y que aún deben establecerse las correspondientes sinergias entre los CMS y el resto de aplicaciones. Ciertas características de unos y otros deben incorporarse mutuamente para favorecer la integración de las arquitecturas funcionales.

En consecuencia, asistimos a un cambio en la dinámica operativa de los CMS, que está permitiendo incorporar las tecnologías de la Web Semántica en el núcleo de estos sistemas. Dichas tecnologías pueden garantizar la viabilidad de los procesos de intercambio de información dentro de un mismo sistema o entre sistemas diferentes, tal y como se muestra en la Figura 2.

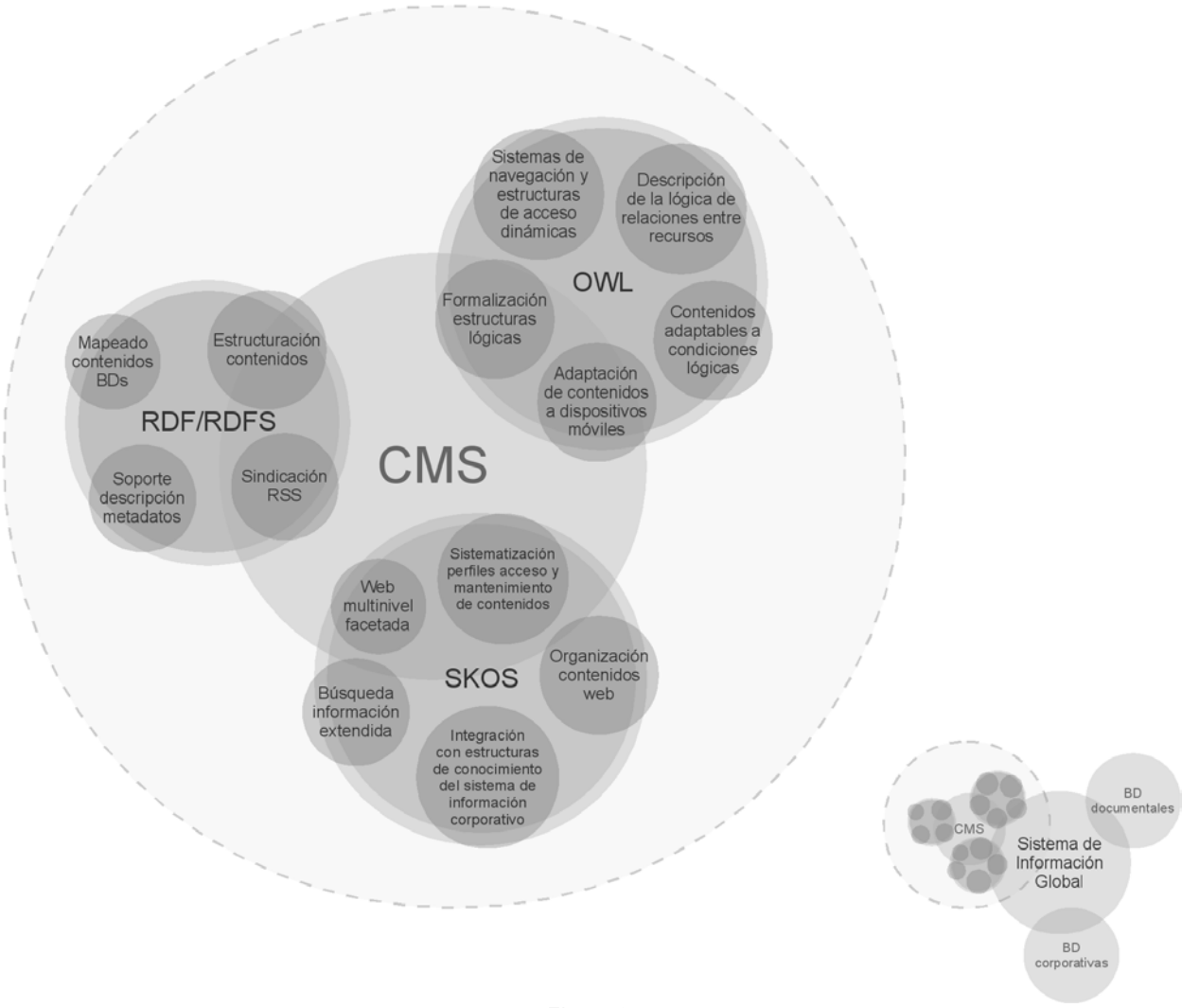

Figura 2

Nuestra propuesta incorpora el concepto de estructuras de metadatos en la gestión de contenidos Web y su almacenamiento y reutilización a través de RDF, así como el uso de este modelo para la interrelación con otras bases de datos y como herramienta de gran valor para el intercambio de información. Se trata de una perspectiva donde cobra una gran importancia la incorporación de esquemas conceptuales representados con SKOS en los procesos de descripción, localización y acceso a la información. Su utilidad puede verse ampliada mediante el uso de OWL para la definición de ontologías. Finalmente, este camino nos ha de llevar a una visión multinivel e integral de los CMS en los sistemas de gestión de información y al replanteamiento de determinados procedimientos con soluciones conceptuales que cuentan con una base tecnológica de desarrollo.

Notas

(1) W3C son las siglas del World Wide Web Consortium, consorcio internacional que produce recomendaciones para el desarrollo de la Web (http://www.w3.org/).

(2) El sitio Web de desarrollo de SKOS es http://www.w3.org/2004/02/skos/

(3) OWL es el acrónimo de "Ontology Web Language", un lenguaje de marcado para publicar y compartir datos usando ontologías en la Web. Una de sus variantes es OWL Full.

Referencias

Alonso, G., Casati, F., Kuno, H., y Machiraju, V. (2004). Web Services. Concepts, Architectures and Applications. // Berlin: Springer Verlag, 2004.

Ambite Arnal, A.; Díaz Gavilanes, R.; Naya Altuna, R.; Ruíz Luna, L.R. (2006). Gestores de contenido CMS y C-CMS (Groupware). http://cv.uoc.edu/~cv052_75_000_01_r06/pf_myc_0606/proyectos/uocms_pf.doc (2009-03-27).

Arano, S. (2005). Los tesauros y las ontologías en la Biblioteconomía y la Documentación. // Hipertext.net- 3. http://www.hipertext.net/web/pag260.htm (2009-03-27).

Baldomero Martínez, J. M. (2006). Los sistemas de gestión de contenidos en el ámbito de la Gestión Integral de Información. Facultad de Comunicación y Documentación, Universidad de Murcia, Proyecto Fin de Carrera.

Becket, D. (2004). RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. World Wide Web Consortium. http://www.w3.org/TR/rdf-syntax-grammar/ (2009-03-27).

Brickley, D.; Guha, R. V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. World Wide Web Consortium. http://www.w3.org/TR/rdf-schema/ (2009-03-27).

Booth, D.; Haas, H.; McCabe, F.; Newcomer, E.; Champion, M.; Ferris, C.; Orchard, D. (2004). Web Services Architecture. World Wide Web Consortium. http://www.w3.org/TR/ws-arch (2009-03-27).

Cruse, D. A. (2004). Meaning in Language: An introduction to Semantics and Pragmatics. Oxford: Oxford University Press, 2004.

Dean, M. y Schreiber, G. (2004). OWL Web Ontology Language Reference. W3C Recommendation 10 Feb 2004. // World Wide Web Consortium. http://www.w3.org/TR/owl-ref/ (2009-03-27).

García Jiménez, A. (2004). Instrumentos de representación del conocimiento: Tesauros versus Ontologías. // Anales de Documentación. 7 (2004) 79-95. http://www.um.es/ojs/index.php/analesdoc/article/viewFile/1691/1741 (2009-03-27).

Gómez Pérez, A.; Fernández López, M.; Corcho, O. (2004). Ontological Engineering. London: Springer Verlag, 2004.

Gruber, T. R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Palo Alto: Stanford Knowledge Systems Laboratory, 1993.

Horrocks, I.; Patel-Schneider, P. F.; Harmelen, F. van. (2003). From SHIQ and RDF to OWL: The making of a web ontology language // Journal of Web Semantics. http://www.cs.man.ac.uk/~horrocks/Publications/download/2003/HoPH03a.pdf (2009-03-27).

Isaac, A.; Summers, E. (2008). SKOS Primer. W3C Working Draft 29 August 2008. Editores: A. Isaac y E. Summers. http://www.w3.org/TR/2008/WD-skos-primer-20080829 (2009-03-27).

Kabisch, T.; Neiling, M. (2005). Wrapping of web sources with restricted query interfaces by query tunneling. // Proceedings of InterDB 2005. International Workshop on Database Interoperability. http://www.cedis.fu-berlin.de/mitarbeiter/mneiling/publications/Kabisch_Neiling_QueryTunneling_interdb05.pdf (2009-03-27).

Manola, F.; Miller, E. (2004). RDF Primer. W3C Recommendation 10 February 2004. World Wide Web Consortium. Recuperado el 15 de junio de 2008, de http://www.w3.org/TR/rdf-primer/ (2009-03-27).

Matthews, B.; Miles, A. (2001). Review of RDF Thesaurus Work: A review and discussion of RDF schemas for thesauri. Public DRAFT for discussion. // World Wide Web Consortium. http://www.w3.org/2001/sw/Europe/ reports/thes/8.2/ (2009-03-27).

Méndez Rodríguez, E. (2007). Dublin core, metadatos y vocabularios. // Anuario ThinkEPI. 2007:1 (2007) 61-64.

Miles, A.; Brickley, D. (eds.) (2005). SKOS Core Guide. 2nd W3C Public Working Draft 2 November 2005. http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102 (2009-03-27).

Miles, A.; Bechhofer, S. (2009). SKOS Reference. W3C Candidate Recommendation 17 March 2009. World Wide Web Consortium. http://www.w3.org/TR/2009/CR-skos-reference-20090317 (2009-03-27).

Otman, G. (1996). Les représentations sémantiques en terminologie. Paris: Masson, 1996.

Pastor Sánchez, J. A. (2009). Diseño de un sistema colaborativo para la creación y gestión de tesauros en Internet basado en SKOS. Universidad de Murcia, Tesis Doctoral. Publicado en TDR (Tesis Doctorales en Red) http://www.tesisenred.net/TDR-0403109-113737/index_cs.html (2009-03-27).

Patel-Schneider, P. F.; Hayes, P.; Horrocks, I. (2004). OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation 10 Feb 2004. // World Wide Web Consortium. http://www.w3.org/TR/owl-semantics/ (2009-03-27).

Tramullas Saz, J.; Garrido Picazo, P. (2006). Sistemas de Gestión de Contenidos. // Tramullas Saz, J. (coord.) Tendencias en documentación digital. Gijón: Trea, 2006, 135-161.

Tudhope, D.; Harith, A.; Jones, C. (2001). Augmenting Thesaurus Relationships: Possibilities for Retrieval. // Journal of Digital Information. 1:8 (2001) Article 41. http://jodi.tamu.edu/Articles/v01/i08/Tudhope/ (2009-03-27).

Serrano-Cobos, J. (2007). Evolución de los sistemas de gestión de contenidos (CMS): del mainframe al open source. // El profesional de la información. 16:3 (2007) 213-215.

Smith, M.K., Welty, C. y McGuinness, D.L. (2004). OWL Web Ontology Language Guide. W3C Recommendation 10 Feb 2004. // World Wide Web Consortium. http://www.w3.org/TR/owl-guide/ (2009-03-27).