Vocabularios estructurados, Web Semántica y Linked Data: oportunidades y retos para los profesionales de la documentación[1]

Carmen Caro Castro
Departamento de Biblioteconomía y Documentación
Universidad de Salamanca
c.e. ccaro@usal.es

En sus poco más de veinte años de vida la World Wide Web ha modificado nuestra forma de acceder a la información, de difundirla, de trabajar, de pasar el tiempo de ocio y de comunicarnos con los demás. La gestión de la información disponible en la Red, en constante incremento, ha supuesto un nuevo desafío para los profesionales de la Documentación que se han tenido que plantear si las técnicas y las herramientas que empleaban habitualmente en su trabajo eran aplicables al mundo digital. Simplificando un complejo panorama, las alternativas que se han propuesto para enfrentase a este reto se pueden agrupar en dos tendencias.

Una enfatiza el uso de las técnicas automatizadas como solución, basándose en argumentos como la imposibilidad de tratar por otros medios la ingente masa de documentación o la idoneidad del tratamiento automático para documentos electrónicos (nativos o de segunda generación). Este es el enfoque que, desde los trabajos iniciales de Salton, ha defendido la Information Retrieval. Desde este punto de vista, la solución radica en el perfeccionamiento de las técnicas automáticas de indización sobre texto completo y el desarrollo de sofisticados algoritmos tanto para la selección de los términos de indización, como para la organización y recuperación de la información. El objetivo sería alcanzar un nivel más alto de rendimiento de los motores de búsqueda, desarrollando algoritmos cada vez más inteligentes, porque las herramientas y métodos bibliotecarios tradicionales resultan obsoletos, al menos en la escala de la Web.

Por otro lado, muchos documentalistas y especialistas de la organización del conocimiento son de la opinión de que los motores de búsqueda no pueden tratar de manera adecuada los problemas lógico-lingüísticos de la representación y recuperación de información. Desde esta perspectiva se reivindica la necesidad de descripciones “estructuradas” de los documentos

[1] Arquivologia, Biblioteconomia e Ciência de Informação : Identidades, Contrastes e Perspectivas de Interlocução. Eduardo Ismael Murgia, Mara Eliane Fonseca Rodrigues (orgs.). Niterói: Editora da UFF, 2012, p. 139-155. ISBN 978-85-228-0882-3

que permitan saber con claridad quién es su autor, cuál su título, su materia, etc... Esta tendencia, que podríamos denominar semántica, promueve el desarrollo de modelos de metadatos para estructurar la información descriptiva sobre los documentos digitales. También promueve la implementación de vocabularios estructurados o listas de autoridad para controlar el contenido de estos metadatos y para organizar la información. De esta forma se puede conseguir una organización más eficaz de las colecciones que redundará en una recuperación de información más pertinente.

El debate estuvo bastante abierto durante la primera década de la Web, cuando el nivel de rendimiento de los motores de búsqueda era todavía excesivamente pobre. Surgieron entonces sistemas que utilizaban la indización humana para crear directorios de recursos, aunque la falta de un paradigma que sustentara las clasificaciones realizadas por personas no especializadas, contribuyó a su descrédito (Caro-Castro 1998). La utilización de clasificaciones bibliográficas y vocabularios controlados solo parecía aplicable en contextos abarcables o finitos de información, fuera su finitud debida al límite de gestión local de los documentos (una biblioteca digital propiamente dicha), a su límite físico (una intranet) o al límite temático (subject gateways). Años más tarde, y en el ámbito empresarial, se puso de manifiesto la utilidad de las taxonomías para facilitar la organización de los recursos y la navegación en las búsquedas. Estas herramientas combinan la estructura jerárquica característica de los sistemas de clasificación con la representación de las clases mediante etiquetas textuales en lugar de códigos numéricos o alfanuméricos. Además se trata de instrumentos que se adaptan a las necesidades de una comunidad de usuarios concreta (empresa, institución) y cuya simplicidad contribuye a facilitar la actualización (Zhonghon, Chaudhry y Khoo 2006)

El impresionante auge de la Web durante su segunda década y el éxito de Google parecieron poner sobre la mesa un argumento demoledor a favor de la indización automática. Sobre la base de este éxito, algunos llegaron más lejos con este argumento, hasta el extremo de declarar la obsolescencia de los vocabularios controlados en beneficio de las soluciones tecnológicas (Vatant 2010). Sin embargo, las insuficiencias de los buscadores pusieron de manifiesto las limitaciones de una Web basada en documentos HTML y enlaces hipertextuales: el exceso de información desestructurada, la opacidad del contenido informativo disponible en bases de datos o el hecho de estar basada en mecanismos de reconocimiento de cadenas de caracteres y no de conceptos. La idea de que la información de este tipo documentos no era suficiente para que los buscadores pudieran realizar un trabajo eficaz, impulsó la idea de la necesidad de información “semántica”, en la que se reconozca un sentido, un significado de los datos y de las relaciones entre ellos.

Resulta evidente que la solución al problema de la recuperación en la Web exige el desarrollo de la investigación en ambos sentidos, tanto semántico como tecnológico. Para ello, se debe combinar el desarrollo de estructuras de metainformación y la potencia del procesamiento informático de las mismas. Ambas tendencias coincidirían en el planteamiento de la Web semántica, en la que se persigue una búsqueda inteligente que debería aprovechar el conocimiento estructurado y el valor añadido de lo humano embebido en las estructuras de conocimiento.

La Web semántica y Linked Data

En su primera etapa, la Web consistía esencialmente en un gran aplicación hipertextual con multitud de documentos interconectados, a cada uno de las cuales se le asigna un Unified Resource Identifier (URI). Con la Web semántica se aspira a cambiar este escenario. Su implantación supone un cambio de paradigma, ya que significa el paso a una Red estructurada y organizada, donde el elemento principal son los objetos de información etiquetados semánticamente (Berners-Lee 01). El proyecto está basado en el modelo Resource Description Framework (RDF) e integra una variedad de aplicaciones utilizando XML para la sintaxis y URIs para las denominaciones. De esta forma se proporciona un marco común que permite que la información sea compartida y reutilizada por diferentes aplicaciones.

Además de la interoperabilidad, el gran reto reside en conseguir que los contenidos estén descritos mediante metadatos y dotados explícitamente de semántica. Gracias a la estructuración y a la normalización de la información se espera que los agentes sean capaces de buscar por conceptos, en lugar de hacerlo por simple comparación de caracteres. Si, además, se consigue que la semántica esté explícita en una ontología, los agentes de software serán capaces de deducir e inferir conocimiento. Probablemente, este último objetivo sea tan ambicioso que provoque escepticismo. Pese a todo, hay fundadas esperanzas de que el proyecto de la Web semántica vaya obteniendo resultados que, a medio o largo plazo podrían conducir a un panorama en el que la recuperación de información sea muy diferente de lo que conocemos ahora. De momento ha propiciado el desarrollo de diversos estándares (Dublin Core, XML, RDF, SKOS, OWL, etc.) que pueden contribuir a mejorar los sistemas de búsqueda, especialmente si son aplicados en entornos controlados como los repositorios, las bibliotecas digitales o las intranets.

Codina y Rovira (2006) consideran como hipótesis débil la de una Web ligada a la inteligencia artificial, cuyas páginas estén organizadas, estructuradas y codificadas de tal manera que los ordenadores sean capaces de efectuar inferencias y razonar a partir de sus

contenidos. Frente a esta visión utópica defienden la bondad de la que denominan hipótesis fuerte, ligada al procesamiento robusto para la que la Web semántica es un conjunto de iniciativas destinadas a convertir la Red en una gran base de datos capaz de soportar un procesamiento sistemático y consistente de la información. Esto significa conseguir modelos que permitan estructurar los documentos de manera que sea identificable el contenido de los diferentes campos o elementos. A partir de aquí será sencillo conseguir que la base de datos simule una cierta inteligencia de la que carecen en estos momentos los motores de búsqueda. Para Byrne y Goddard (2010), marcando la información en formatos estandarizados y altamente estructurados como RDF, se puede conseguir que los ordenadores entiendan el significado de los contenidos en lugar de simplemente identificar cadenas de caracteres. Esto permitiría que los motores de búsqueda funcionaran más como bases de datos relacionales, proporcionando resultados más precisos.

En los últimos años, la evolución ha llevado desde un espacio global de información de documentos enlazados hasta uno en el que ambos, documentos y datos, estén enlazados. Por debajo de este proceso está un conjunto de buenas prácticas para publicar y conectar datos estructurados conocido como Linked Data[2], cuyo ejemplo más visible de adopción es el proyecto Linking Open Data. Se pasa así de una Web basada en documentos HTML, donde los enlaces son relaciones entre puntos de los documentos y en la que el usuario es el destinatario de la información publicada, a una Web de Datos Enlazados que están expresados en RDF, en la que sistemas y agentes software pueden explotar estos datos de forma automática (recopilándolos, agregándolos, interpretándolos, publicándolos, etc.). Los cuatro principios de diseño en los que se basa son (Byrne, Hearth y Berners-Lee 2009):

- Utilizar URIs (*Uniform Resource Identifier*) como nombres únicos para los recursos

- Utilizar el protocolo HTTP para nombrar y resolver la ubicación de los datos identificados mediante esas URIs. Si las entidades están identificadas mediante URI's pueden buscarse desreferenciando la URI utilizando el protocolo HTTP

- Representar los datos en RDF y utilizar SPARQL como lenguaje de consulta de dichos datos. Mientras que HTML proporciona un medio para estructurar y enlazar documentos en la Web, RDF proporciona un modelo de datos genérico, con el que estructurar y enlazar datos que describen cosas en el mundo en forma de triples sujeto-predicado-objeto.

[2] http://linkeddata.org/ [Consulta 15 noviembre 2011]

- Incluir enlaces a otras URIs para permitir la localización de más datos enlazados, aprovechando que los elementos de un triple RDF pueden ser una URI o una cadena de literales.

De esta forma, si una aplicación desea obtener información sobre un dato identificado mediante una URI, cuando hace una llamada HTTP se desreferencia el recurso, obteniendo información fácilmente procesable en formato RDF. Cuando, además, se proveen puntos de consulta avanzada, con SPARQL, el resultado ante una consulta podrá ser interpretado de forma automática. Si enlazamos los datos no se quedarán aislados y se podrá compartir información con otras fuentes externas. Gracias a estos mecanismos, cualquier recurso es susceptible de ser enriquecido con cualquier tipo de información especializada, incluso la que no se espera que sea combinable.

Los datos se convierten así en los objetos básicos de información, entendidos como la unidad mínima (elemento o grupo de elementos) que constituye una entidad informativa susceptible de ser descrita mediante metadatos. Este objeto puede ser un documento, una página web, una imagen, un vídeo, una persona, una institución, un lugar o un concepto. Esto implica una deconstrucción del tradicional concepto de documento, puesto que ahora se entiende como un conjunto de átomos informativos, autónomos en sí mismos, que a su vez están interconentados o interrelacionados. El RDF Schema (RDFS[3]) y el Web Ontology Language (OWL[4]) proporcionan una base para crear los vocabularios – colecciones de clases y relaciones – que pueden utilizarse para describir estas entidades en el mundo y cómo se relacionan. Estos vocabularios pueden publicarse libremente para conectarse con otros vocabularios y con los documentos.

Vocabularios estructurados en la Web semántica

Para el mundo de las bibliotecas, archivos y centros de documentación la idea de atomizar así la información no es nueva. Los índices de materia enlazados a los tesauros y, especialmente, los catálogos de autoridades se fundamentan en un principio similar: diferentes elementos del registro bibliográfico, que describe el documento en su conjunto, enlazan con los registros del catálogo de autoridades. En este catálogo es donde se consignan las formas autorizadas y alternativas para representar un concepto, persona, lugar, institución, etc. Además, cualquiera de estas entidades se contextualiza mediante una

[3] http://www.w3.org/TR/rdf-schema/ [Consulta 15 noviembre 2011]
[4] http://www.w3.org/TR/owl2-overview/ [Consulta 15 noviembre 2011]

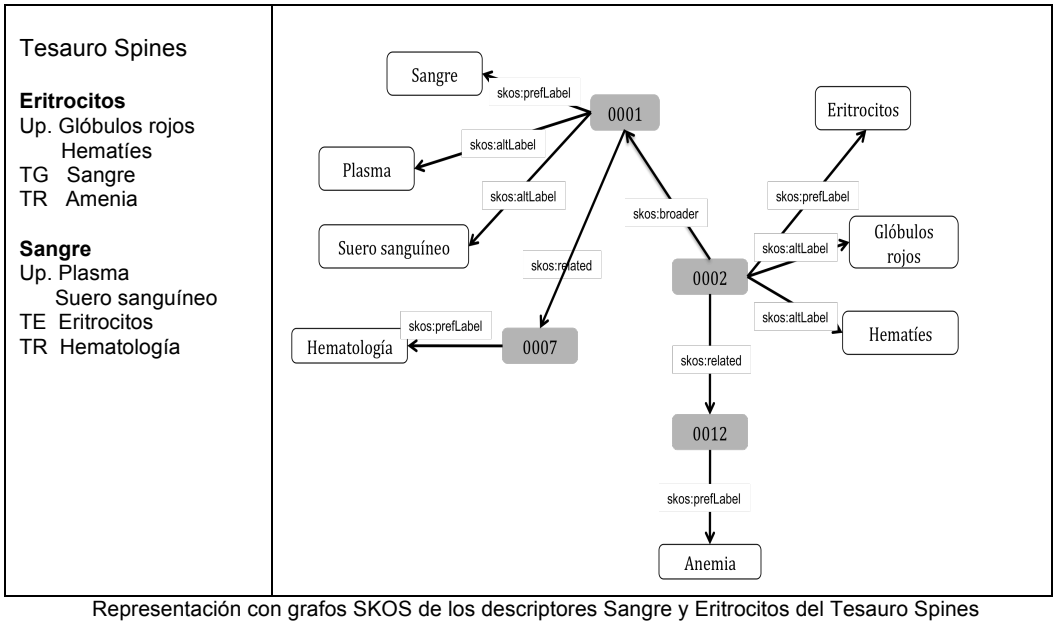
red de referencias semánticas y/o funcionales que concretan su significado, lo que permite la exploración de documentos relacionados.

Si, como hemos visto, el papel de los vocabularios controlados y estructurados en la Web ha sido objeto de debate prácticamente desde su origen, parece que en este contexto su idoneidad es poco cuestionable. El primer paso para hacer posible su utilización es publicarlos conforme a los estándares de este nuevo entorno. Con este objeto, en los últimos años han visto las luz varios formatos. Algunos ligados a la publicación de nuevas normas sobre elaboración de tesauros como el BS 8723 XML Schema[5], vinculado a la *BS 8723: Structured vocabularies for information retrieval*, o el ISO 25964 XML Schema[6], ligado a la *ISO 25964: Thesauri and interoperability with other vocabularies*. Otros, promovidos por organismos del alcance de la Library of Congress como MADS (*Metadata Authority Description Schema*[7]*)*, o el World Wide Web Consortium (W3C) como SKOS (*Simple Knowledge Organization System*[8]). Con el respaldo de la iniciativa Linked Data, SKOS se está convirtiendo en el modelo común para expresar sistemas de organización de conocimiento como tesauros, taxonomías, clasificaciones o listas de autoridad. El modelo se ha desarrollado desde el año 2003 a través de un fructífero diálogo entre las bibliotecas y las comunidades de la Web semántica (Pastor-Sánchez, Martinez-Méndez y Rodríguez-Méndez 2009). Desde 2009 es una recomendación de W3C que se ha utilizado rápidamente para llevar a cabo la migración de los vocabularios tradicionales, incluyendo RAMEAU, NUOVO SOGGETTARIO, LCSH, AGROVOC, EUROVOC y muchos otros[9].

El modelo SKOS ha adoptado un enfoque basado en conceptos que, para la Documentación, ha sido central en la metodología para la elaboración de clasificaciones, vocabularios controlados y estructuras de conocimiento (Campos 2001). Tal como se define en el Manual (2009), un concepto es "una idea o noción; una unidad de pensamiento" que se puede representar unívocamente mediante un URI. Este identificador garantiza la identidad en la Web semántica, igual que el término preferente la garantiza en un vocabulario controlado, independientemente de los términos que se empleen para nombrar los conceptos, las personas, las entidades, los lugares o las cosas. Se asume que este concepto puede estar representado por diferentes códigos, palabras o expresiones (etiquetas) en una o varias lenguas y que está relacionado, por jerarquía o asociación, con otros conceptos del mismo o de otros vocabularios controlados.

---

[5] http://schemas.bs8723.org/ [Consulta 15 noviembre 2011]
[6] http://www.niso.org/schemas/iso25964/schema-intro/ [Consulta 15 noviembre 2011]
[7] http://www.loc.gov/standards/mads/ [Consulta 15 noviembre 2011]
[8] http://www.w3.org/2004/02/skos/ [Consulta 15 noviembre 2011]
[9] Se puede consultar una lista actualizada de los vocabularios disponibles en: Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets: W3C Incubator Group Report 25 October 2011 <http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/ > [Consulta 15 noviembre 2011]



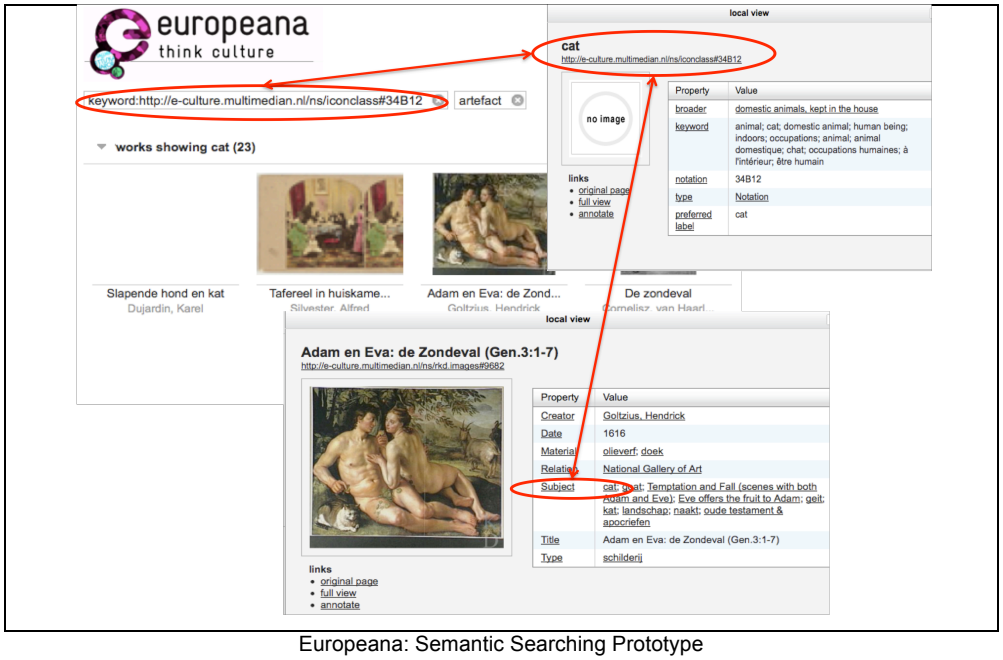
Representación con grafos SKOS de los descriptores Sangre y Eritrocitos del Tesauro Spines

No solo se pueden establecer relaciones entre conceptos dentro de un mismo vocabulario, también se pueden vincular diferentes estructuras de conocimiento mediante las etiquetas . Potencialmente, cada elemento de datos de un documento o de los metadatos con los que se describe una entidad puede ser enlazada a un valor de una estructura de conocimiento etiquetada con SKOS. Esto serviría, por ejemplo, para solucionar los problemas de ambigüedad del lenguaje natural, para lo que, en muchos casos, es necesario contar con información que no está explícita en los textos pero que las personas somos capaces de deducir a partir del contexto. Si utilizamos vocabularios controlados para asignar valores a los campos de metadatos como autor, materia, lugar, etc. se conseguirá eliminar ambigüedades derivadas de la sinonimia o de la homografía. Al mismo tiempo se contextualizará semánticamente cada concepto o cada instancia inscribiéndolo en un campo del conocimiento en el que se vinculará con otros conceptos por relaciones jerárquicas, asociativas o notas de alcance.

En el prototipo de búsqueda semántica de Europeana[10] – la Biblioteca Virtual Europea – se aplican estos principios. Los fondos de Europeana provienen de distintas bibliotecas, museos y colecciones audiovisuales europeas. Ahora bien, los registros de cada uno de los documentos, aunque comparten el mismo esquema de metadatos, utilizan vocabularios controlados diferentes para representar el contenido y, además, están escritos en idiomas distintos. El prototipo que presenta Europeana utiliza una colección de unos 150.000 registros de obras de arte de tres museos: Rijksmuseum Amsterdam, Louvre y el Instituto de Historia

[10] http://eculture.cs.vu.nl/europeana/session/search [Consulta 15 noviembre 2011]

del Arte de los Países Bajos. Para indizarlos se han empleado diversos tesauros (Joconde, IconClass, AAT, RKDArtists, WordNet) que suman más de un millón y medio de términos. Los elementos descriptivos de los metadatos se vinculan con los descriptores del tesauro lo que permitirá, en un futuro, contextualizar la búsqueda, visualizar términos relacionados y realizar búsquedas multilingües. Aunque la solución esté lejos de ser definitiva en cuanto a facilidad de uso, escala y efectividad; el desarrollo del prototipo demuestra que se ha detectado el problema y que la opción elegida para resolverlo es el uso de tesauros (Abadal y Codina 2009).



Europeana: Semantic Searching Prototype

Un paso más allá de lo que pueden aportar los vocabularios estructurados a la Web semántica estaría la contribución de las ontologías. En estas, el conocimiento implícito en las relaciones entre conceptos se expresa de manera que las máquinas puedan entenderlo y manejarlo para hacer inferencias. Si relacionamos nuestros datos con ontologías compartidas, que describen las propiedades y las relaciones entre objetos, empezaremos a permitir que los ordenadores no solo entiendan el contenido, sino que deriven nuevo conocimiento haciendo deducciones lógicas sobre el que se les aporta (Legg 2007).

Retos y oportunidades

Los bibliotecarios y documentalistas tradicionalmente han utilizado los vocabularios controlados como un elemento fundamental de sus bases de datos y catálogos, convertirlos en herramientas eficaces en la Web Semántica proporcionará un potencial ilimitado para

facilitar su utilización en múltiples nuevas formas (Harper y Tillet 2007). Sin duda, el movimiento Linked Data es una gran oportunidad para que este conocimiento se comparta y esté accesible en un ámbito tan amplio como la Web (Byrne y Goddard 2010) . En gran medida, la cuestión más importante para el mundo de la Documentación en relación con las tecnologías semánticas es si se puede transformar satisfactoriamente su experiencia en el trabajo con metadatos en experiencia en el uso de ontologías o modelos de conocimiento (Tennis y Calzada-Prado 2007).

Una parte del trabajo, la más avanzada actualmente, consiste en adaptar los vocabularios estructurados a los estándares y a las recomendaciones de la Web semántica y Linked Data. Muchos vocabularios ya están publicados conforme a las especificaciones del modelo SKOS como se desprende del informe del Linked Data Incubator Group (25 de octubre de 2011), aunque sus aplicaciones para la recuperación de información todavía son muy limitadas. En general, la acogida ha sido entusiasta, pero no faltan las voces que plantean las limitaciones y problemas de este formato. En un artículo publicado en 2009, Martha Yee se preguntaba si las especificaciones y los lenguajes de la Web semántica permiten expresar la información bibliográfica con el mismo nivel de granularidad con el que tradicionalmente se recogía en catálogos automatizados y bases de datos bibliográficas. Su respuesta es que actualmente no, y pone de manifiesto diferentes aspectos de FRBR que son difíciles de consignar en un modelo basado en RFD. Por lo que afecta a las materias señala, por ejemplo, la imposibilidad de diferenciar con SKOS si un nombre geográfico se está entendiendo simplemente como lugar o como área jurisdiccional (en MARC se identifican con etiquetas distintas), o la de identificar diferentes elementos dentro de un encabezamiento de materia (problema extensible a todos los sistemas precoordinados). También se pregunta si Internet será lo suficientemente rápida como para reconstruir toda la información referente a un registro o un documento, desreferenciándola a partir de múltiples URI's: cuanta más granularidad hay en los datos, más enlaces son necesarios para asegurar que los elementos atomizados se recuperan juntos (Yee 2009).

Sin embargo, los retos más importantes no son nuevos y están relacionados con la calidad de los vocabularios estructurados, con su uso para normalizar las descripciones realizadas con metadatos y, sobre todo, para recuperar información. En esto el panorama no es muy diferente al que se ha podido plantear en otros entornos como los OPAC's o las bases de datos bibliográficas. Si bien los vocabularios controlados y estructurados son una herramienta potencialmente útil para organizar y recuperar la información, deben ser herramientas semánticamente consistentes, adaptarse al público al que se dirigen y gestionarse de manera que sean útiles para buscar información en un entorno que es heterogéneo – en cuanto a usuarios y contenidos – y multilingüe.

Los formatos de intercambio desarrollados en el marco de la Web semántica permiten la interoperabilidad entre sistemas, pero no resuelven el problema de la compatibilidad entre vocabularios estructurados. Este problema, que es fundamental en un entorno en el que la información es multilingüe y multidisciplinar, es un "viejo conocido" para la Documentación. En 1995 Dahlberg, con ocasión del seminario *Compatibility and Integration of Order Systems*, recopilaba una amplia bibliografía sobre este tema publicada entre 1960 y 1995. Poco después Maniez recordaba que la compatibilidad semántica, el sueño de una comunicación universal entre dichos vocabularios, es el paraíso perdido de los científicos de la información. Los proyectos que se han desarrollado para lograrla han puesto de manifiesto los obstáculos para traducir los descriptores de una lengua a otra y la dificultad para establecer equivalencias (mapeos) derivada de las diferencias en la estructura, en la sintaxis, en la especificidad o en las relaciones semánticas (Maniez 1997). Aunque se ha conseguido desarrollar sistemas tan elaborados como el Unified Medical Language System (UMLS)[11], conseguir la interoperabilidad semántica sigue siendo un reto para los profesionales, especialmente en un medio tan heterogéneo como Internet.

La experiencia en el análisis de dominio que es posible aportar desde de la Documentación puede ahorrar un enorme esfuerzo de análisis, proporcionando modelos para la organización de la información en la Web (Soergel 2002). Por ejemplo, las bases teóricas y metodológicas para construir estructuras facetadas han resultado especialmente bien aceptadas por su adecuación para definir las entidades, atributos y relaciones de las ontologías (Vickery 2008, Prieto 2003). Sin embargo, todavía es necesario un gran esfuerzo para sistematizar los criterios empleados para establecer relaciones lógico lingüísticas y para que su uso sea consistente en los tesauros, encabezamientos de materia y clasificaciones que ya se están trasfiriendo a la Red. Una definición clara de los principios que sustentan las relaciones de equivalencia, jerárquicas y, sobre todo, asociativas, contribuirá a fortalecer la estructura semántica de los vocabularios estructurados y a favorecer la participación en la elaboración de ontologías, en las que estas relaciones deben hacerse explícitas para posibilitar el trabajo de los agentes de software.

La ampliación del ámbito de aplicación de los vocabularios estructurados que implica su utilización en la Web, también plantea la necesidad de que se adecuen al contexto en el que se van a emplear. Esto no solo significa recoger la terminología de un campo de conocimiento determinado, sino la adaptación a las características culturales o sociales de los potenciales usuarios: niños, trabajadores de una empresa, científicos, etc. Algunas propuestas proceden del diseño participativo, basado en la filosofía de que la implicación

[11] http://www.nlm.nih.gov/research/umls/ [Consulta 15 noviembre 2011]

directa del usuario final en la creación y en el desarrollo de los sistemas puede tener un impacto importante en términos de un uso sostenible y productivo. Desde esta perspectiva se intenta involucrar al usuario en algo más que la validación y evaluación de los sistemas, como tradicionalmente se había hecho, abogando por su participación en el diseño gracias al empleo de métodos *bottom-up* en los que los propios usuarios asignan las etiquetas léxicas y crean las categorías (Legg, 2007).

Dado el volumen de información disponible en la Web, será inevitable automatizar algunos de los procesos que supone la elaboración de grandes estructuras de conocimiento, aplicando procedimientos con los que se ha experimentado desde la *Information Retrieval*. Empresas especializadas como Sinequa[12], Endeca[13], Semantic Web Company[14] están desarrollando herramientas para la clasificación automatizada que son capaces de "ingerir" y organizar grandes cantidades de datos semiestructurados procedentes de diferentes fuentes (Regli 2009). Estas herramientas ofrecen ayudas para la construcción de tesauros a partir de la información extraída del texto de los documentos que sería deseable contrastar con los diseños obtenidos de las metodologías participativas (Srinivasan, Pepe y Rodriguez 2009).

Probablemente el aspecto al que se haya prestado menos atención hasta ahora sea a la utilización de los vocabularios estructurados y de los sistemas de clasificación para la búsqueda de información. Esta función se debería haber potenciado con el uso de la tecnología y el incremento del acceso público a bases de datos y catálogos en línea. Sin embargo, las estructuras de conocimiento han sido más un obstáculo que una ayuda para los usuarios porque no se han implementado adecuadamente ni las técnicas de recuperación ni los métodos para poner de manifiesto las relaciones entre los elementos informativos. Tanto en la bibliografía profesional (Borgman 2000, Soergel 2002) como en las nuevas normas vocabularios estructurados – ANSI/NISO Z39.19, BS 8723 o ISO 25964 – se hace hincapié en la funcionalidad que deberían soportar los sistemas informáticos para convertir estas herramientas en una ayuda para recuperar información. Tendrían que ayudar al usuario a expresar su necesidad de información guiándole para hacer un análisis conceptual de un tema mediante menús o formularios de búsqueda facetados. Deberían llevar al usuario desde sus términos de búsqueda a los descriptores del tesauro y proporcionarle la posibilidad de realizar búsquedas inclusivas, incorporando los términos específicos de un descriptor determinado. Para facilitar las búsquedas exploratorias el sistema tendría que ofrecer opciones para navegar por la estructura sistemática o por una lista alfabética, de forma que el usuario pueda conocer el espacio de información e identificar conceptos útiles al nivel de especificidad deseado. Por último, también los resultados podrían clasificarse de acuerdo a

[12] http://www.sinequa.com/ [Consulta 15 noviembre 2011]
[13] www.endeca.com [Consulta 15 noviembre 2011]
[14] http://www.semantic-web.at/ [Consulta 15 noviembre 2011]

criterios facetados, convirtiendo cualquier conjunto de documentos recuperado en subconjuntos pequeños y ordenados.

Los retos no son pocos. Muchos tampoco son nuevos. Sin embargo, es evidente que la transformación de vocabularios estructurados en ontologías jugará un papel crucial en la Web de los próximos años y será uno de los caballos de batalla de los profesionales de la Documentación. En esta tarea deberíamos recordar las palabras de Soergel (2003): no reinventemos la rueda, mejorémosla.

Bibliografía

Abadal, Ernest; Codina Lluís (2009) Búsqueda semántica en Europeana: se percibe el problema pero aún no la solución. Anuario ThinkEPI <http://www.lluiscodina.com/2009-busqueda-semantica-europeana.pdf> [Consulta 15 noviembre 2011]

Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001) «The Semantic Web», *Scientific American*, May 2001

Bizer, Christian; Hearth, Tom; Berners-Lee, Tim (2009). Linked Data - The Story So Far. *International Journal on Semantic Web & Information Systems*, 5(3), p. 1-22

Borgman, Christine L. (2000) *From Gutenberg to the global information infrastructure: access to information in the networked world*. Cambridge, MA: The MIT Press

Byrne, Gillian; Goddard, Lisa (2010) The Strongest Link: Libraries and Linked Data. *D-Lib Magazine*, 16(11/12), <http://www.dlib.org/dlib/november10/byrne/11byrne.print.html> [Consulta 15 noviembre 2011]

Campos, Mª Luiza de Almeida (2001) Linguagem Documentaria: teorias que fundamentan sua elaboração. Niterói, RJ: Editora da Universidade Federal Fluminense.

Codina, Lluís; Rovira, Cristòfol (2006) La Web semántica. En: *Tendencias en documentación digital*. Jesús Tamullas (coord.). Gijón: Trea, p. 9-54

Caro-Castro, Carmen(1998) Sistemas de clasificación y organización de la información en Internet. En: Jornadas Españolas de Documentación (6. 1998. Valencia). *Los sistemas de información al servicio de la sociedad: actas de las VI Jornadas Españolas de Documentación, Valencia del 29 al 31 de octubre de 1998*. Valencia: FESABID, p. 197-204

Dahlberg, Ingetraut (1996) Compatibility and Integration of Order Systems 1960-1995: An Annotated Bibliography. En: *Compatibility and Integration of Order Systems: Research Seminar Proceedings of the TIP/ISKO Meeting. Warsaw, 13-15 September 1995*. I. Dahlberg and K. Siwek (eds.). Warsaw: Wydawnictwo SBP

Hannemann, Jan; Kett, Jürgen (2010) Linked Data for Libraries. En: International Fedetarion of Library Associations. Congress (76 th. Gothenburg, 2010).Open access to Knowledge: promoting sustainable progress : World Library and Information Congress: 76th IFLA Conference and Assembly, 10-15 August 2010, Gothenburg, Sweden <http://conference.ifla.org/past/ifla76/2010-08-15.htm> [Consulta 15 noviembre 2011]

Harper, Corey A.; Tillet, Barbara B. (2007) Library of Congress Controlled Vocabularies and

Their Application to the Semantic Web. *Cataloging & Classification Quarterly,* 43 (3/4), p. 47-68

Legg, Catherine (2007) Ontologies on the semantic web. *Annual Review of Information Science and Technology*, 41(1), p. 407-451

Library Linked Data Incubator Group Final Report: W3C Incubator Group Report, 25 October 2011 <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/> [Consulta 15 noviembre 2011]

Maniez, Jacques (1997) Database Merging and the Compatibility of Indexig Languages. Knowledge Organization, 24(4), p. 213-224

Manual de SKOS (Simple Knowledge Organization System), 2009. Versión española de Juan Antonio Pastor Sánchez y Francisco Javier Martínez Méndez <http://skos.um.es/TR/skos-primer> [Consulta 15 noviembre 2011]

Pastor-Sánchez, J.A., Martinez-Méndez F.J. Rodríguez Muñoz J.V. (2009). "Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives" *Information Research*, 14(4) paper 422. <http://InformationR.net/ir/14-4/paper422.html> [Consulta 15 noviembre 2011]

Prieto-Diaz, R. (2003) A Faceted Approach to Building Ontologies. En: Proceedings of the *2003 IEEE International Conference on Information Reuse and Integration, October 27-29, 2003, Las Vegas, NV.*
<https://users.cs.jmu.edu/prietorx/Public/publications/BulidOntologiesRPD-ER2002.doc>
[Consulta 15 noviembre 2011]

Regli, Theresa (2009) The Death of Taxonomies, revisited. CMS Watch <http://www.cmswatch.com/Blog/1737-Death-of-Taxonomies-Revisited> [Consulta 15 noviembre 2011]

Rodriguez-Castro, B.; Glasser, H.; Carr, L.(2010) How to Reuse a Faceted Classification and Put it on the Semantic Web. En: *9th International Semantic Web Conference (ISWC2010), Shanghai, november 7-11 2010* <http://iswc2010.semanticweb.org/pdf/253.pdf> [Consulta 15 noviembre 2011]

Soergel, D. (2002). "Thesauri and ontologies in digital libraries. En: Joint Conference on Digital Libraries (JCDL 2002): Portland, OR, USA, July 14, 2002 <www.dsoergel.com/cv/B63.pdf> [Consulta 15 noviembre 2011]

Soergel, D. (2003) From Legacy Knowledge Organization Systems to Full-Fledged Ontologies (Dagobert Soergel, University of Maryland) BUILDING A MEANINGFUL WEB: From Traditional Knowledge Organization Systems to New Semantic Tool The 6th NKOS Workshop at ACM-IEEE Joint Conference on Digital Libraries (JCDL) May 31, 2003 Houston, TX

Srinivasan, R.; Pepe, A,; Rodriguez M.A. (2009) A clustering-based semi-automated technique to build cultural ontologies. Journal of the American Society for Information Science and Technology, 60(3), p. 608-620.

Tennis, Joseph T.; Calzada-Prado, Javier (2007) *Ontologies and the Semantic Web: Problems and Perspectives for LIS professionals. IBERSID: Revista de Sistemas de Información y Documentación,* 1, p. 303-311.

Vatant, Bernard (2010). Porting library vocabularies to the Semantic Web, and Back: A win-win round trip. En: International Federation of Library Associations. Congress (76 th.

Gothenburg, 2010).Open access to Knowledge: promoting sustainable progress : World Library and Information Congress : 76 th IFLA General Conference and Assembly, 10-15 August 2010, Gothenburg, Sweden, <http://conference.ifla.org/past/ifla76/2010-08-15.htm> [Consulta 15 noviembre 2011]

Vickery, Brian (2008) Faceted Classification for the Web. *Axiomathes*, 18, p. 145-160

Yee, Martha M. (2009) Can Bibliographic Data be Put Directly onto the Semantic Web? *Information Technology and Libraries*, 28(2), <http://escholarship.org/uc/item/91b1830k> [Consulta 15 noviembre 2011]

Zhonghon, Wang; Chaudhry, Abdus Sattar; Khoo, Chistopher (2006) “Potential and prospects of taxonomies for content organization. *Knowledge Organization*, 33(3), p. 160-169.