Web semántica y ontologías en el procesamiento de la información documental

Por **Rafael Pedraza-Jiménez, Lluís Codina** y **Cristòfol Rovira**

Resumen: *La carencia de un modelo bien definido de representación de la información en la web ha traído consigo problemas de cara a diversos aspectos relacionados con su procesamiento. Para intentar solucionarlos, el W3C, organismo encargado de guiar la evolución de la web, ha propuesto su transformación hacia una nueva web denominada web semántica. En este trabajo se presentan las posibilidades que ofrece este nuevo escenario, así como las dificultades para su consecución, prestando especial atención a las ontologías, herramientas de representación del conocimiento fundamentales para la web semántica. Por último, se analiza el papel del profesional de la biblioteconomía y documentación en este nuevo entorno.*

Palabras clave: *Web semántica, Ontologías, Rdf, Owl, Sistemas de información.*

Rafael Pedraza-Jiménez *es miembro del grupo de investigación DigiDoc y profesor del Área de Biblioteconomía y Documentación de la Universidad Pompeu Fabra. Imparte docencia en las titulaciones de comunicación audiovisual y publicidad y relaciones públicas, así como en el máster online en documentación digital. Sus principales líneas de trabajo son las taxonomías y la generación semiautomática de ontologías, uno de los temas centrales de su tesis doctoral.*

Lluís Codina *es profesor titular de universidad. Imparte docencia en los estudios de periodismo y en la Facultad de Comunicación Audiovisual de la Universidad Pompeu Fabra de Barcelona. Es el investigador principal del Grupo de Investigación DigiDoc de la misma universidad. Participa en el máster interuniversitario UB/UPF en gestión de contenidos digitales, en el programa de doctorado del Departamento de Periodismo y de Comunicación Audiovisual y es co-director del máster online de documentación digital.*

Cristòfol Rovira *es profesor de la Universidad Pompeu Fabra en el Área de Biblioteconomía y Documentación. Imparte docencia en las titulaciones de publicidad y relaciones públicas y traducción e interpretación. Es coordinador del máster interuniversitario UB/UPF en gestión de contenidos digitales y director del máster online de documentación digital. Es investigador del grupo DigiDoc de la Universidad Pompeu Fabra y director del Laboratorio DigiDoc del mismo grupo.*

Title: Semantic web and ontologies in document information processing

Abstract*: The lack of a well defined model of information representation on the web has produced several problems related to processing information. In an effort to resolve these problems, the W3C has proposed the semantic web project. This new scenario offers both possibilities and difficulties for the future. Special attention is given to ontologies, fundamental tools for the representation of knowledge on the semantic web. Finally, the role of library and information professionals is considered in this new context.*

Keywords*: Semantic web, Ontologies, Rdf, Owl, Information systems.*

Pedraza-Jiménez, **Rafael**; **Codina, Lluís**; **Rovira, Cristòfol**. "Web semántica y ontologías en el procesamiento de la información documental". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 569-578.

DOI: 10.3145/epi.2007.nov.04

1. Contexto

Hasta la primera mitad del siglo pasado la gestión de la información documental fue un dominio casi exclusivo de bibliotecarios, archiveros y documentalistas. Pero la introducción de los ordenadores en la segunda mitad del siglo XX, la continuada adaptación de los procesos de trabajo a las nuevas tecnologías y, principalmente, la creación de la web en los noventa supuso la incorporación de nuevas disciplinas (muy particularmente la teoría de la recuperación de información) a este entorno. La consecuencia inmediata ha sido la proliferación, desde entonces, de multitud de investiga-

Artículo recibido el 11-07-07
Aceptación definitiva: 18-10-07

ciones centradas en el desarrollo de tecnologías y métodos que permitan la organización y la gestión de la información documental.

No obstante, a pesar de los importantes avances aportados por las nuevas tecnologías, el usuario de la web aún carece de un sistema que permita procesar y acceder a la información documental contenida en sitios web de una manera fiable. El problema estriba en al menos tres aspectos: en primer lugar la web es un sistema descentralizado y heterogéneo completamente distinto de los escenarios para los que estaban más o menos bien preparadas las disciplinas clásicas vinculadas con la documentación y la recuperación de la información. En segundo lugar, lo que sucede en la www es una recuperación de información "con adversario" (*adversarial information retrieval*), otro aspecto nunca contemplado por la recuperación de información clásica. Por último, originalmente el método de marcado de la información, html, combina elementos de contenido con otros de presentación. Para un ser humano no hay ningún problema en interpretar el título de un documento a partir, por ejemplo, de su preeminencia, su formato y su lugar en la página, pero si el autor ha marcado el título con un elemento de formato (<*font*>) en lugar de uno semántico (<*h1*>), para un ordenador resultará imposible identificar el título.

"La web es un escenario heterogéneo completamente distinto de los que usaban las disciplinas clásicas vinculadas con la documentación"

Posteriores correcciones, como la firme recomendación del *W3C* de separar contenido de presentación mediante el uso de "nuevas" versiones de html y xhtml, así como el conjunto de especificaciones en torno al estándar xml se espera que, poco a poco, vayan alterando este panorama hacia una web donde el marcado de los documentos se realice de forma "semántica", es decir, utilizando etiquetas que expresen el significado de los elementos y no su formato (que queda a cargo de normas como xsl o css). Además, con el fin de facilitar la gestión documental de estos nuevos recursos y controlar su heterogeneidad, se ha propuesto el diseño de herramientas que faciliten reconocer, comparar y combinar recursos web con diferente estructura: las ontologías (siendo la principal recomendación del *W3C* para su construcción el lenguaje *OWL* (*web ontology language*). Se espera que este nuevo escenario, caracterizado por la existencia de contenidos autodescritos y herramientas automáticas capaces de comprenderlos, facilite el proceso de recuperación de información y, entre otras cosas, termine con las estrategias fraudulentas de posicionamiento web.

La web semántica

Berners-Lee (2001) publicó un artículo programático en el que anunciaba el proyecto de la web semántica como una extensión de la actual, dotada de una estructura que permitiera expresar el contenido de las páginas de una forma que los ordenadores pudieran "entenderlas" y que posibilitase tanto la interacción entre ordenadores como entre éstos y los usuarios. Proponía así un nuevo modelo en el que todos sus contenidos estarían descritos y estructurados de un modo que las máquinas podrían comprenderlos.

Para que ello fuera posible, **Berners-Lee** suponía que en la web de un futuro cercano los ordenadores tendrían acceso a información semánticamente marcada y estructurada, a ontologías que expresarían conceptos, y a conjuntos de reglas de inferencia útiles para llevar a cabo razonamientos automáticos sobre las páginas web que permitiesen a los ordenadores desarrollar tareas inteligentes.

Ahora bien, de acuerdo con las previsiones iniciales este panorama descrito en el 2001 debería empezar a hacerse evidente siete años después. Tal vez porque esta transformación no ha tenido lugar, el *W3C* (que no olvidemos está dirigido por **Berners-Lee**) presenta ahora una visión mucho más prudente, orientada hacia la codificación semántica de los documentos y a la aplicación de nuevas tecnologías y procedimientos de representación del conocimiento con el fin de mejorar el acceso a los recursos de la web. Muchos de ellos se muestran a continuación.

2. Tecnologías

En sólo diez años el *W3C* ha elaborado más de ochenta especificaciones técnicas para la implantación de esta nueva infraestructura. Los principales medios con los cuales se persiguen los objetivos de la web semántica son, a grandes rasgos, los siguientes: en primer lugar, mediante una codificación de páginas en la cual las etiquetas transporten una carga semántica. Este apartado corresponde al estándar denominado xml (*Xml*, 2004). En segundo lugar, aportando descripciones (metadatos) (**Rovira**, 2006) de las páginas y sitios web con un formato que sea compatible con la estructura general de la www y con diversas categorías de páginas, e interoperable entre distintos sistemas informáticos. De esto se ocupa la norma rdf (*Rdf*, 2004). Además, mediante un sistema de ontologías que permitan especificar conceptos de diversos dominios del conocimiento mediante el uso de un lenguaje fuertemente

Término lingüístico	Sujeto	Predicado	Objeto
Término lógico	Recurso	Propiedad	Valor
Ejemplo	HpDeskjet9800	Tipo de impresión	Inyección térmica de tinta

Tabla 1: Equivalencias logicolingüísticas en una declaración rdf

basado en lógica simbólica y susceptible, por tanto, de ser eventualmente "interpretado" por un ordenador. De este aspecto se ocupa el denominado *Owl* (*Owl*, 2004). Cada una de estas tecnologías ha sido definida por varias especificaciones, y constituyen la base sobre la que el *W3C* pretende construir la web semántica.

2.1. Xml

Es sin ninguna duda el elemento de la web semántica que mayor repercusión tiene ya en biblioteconomía/documentación. Es un estándar que, junto con su norma asociada *Xml schema*, permite definir tipos de documentos y los conjuntos de etiquetas necesarias para codificarlos. La idea es que, una vez están marcados o codificados con una colección de etiquetas xml, es posible procesarlos y explotarlos de forma automática con diversos propósitos, de la misma manera que un grupo de registros de una base de datos se puede emplear de formas diversas, e incluso exportarse a diferentes sistemas de gestión de bases de datos si la estructura de registros sigue algún tipo de estándar.

"Xml es el elemento de la web semántica que mayor repercusión tiene ya en biblioteconomía/documentación"

Posibilita así a sus usuarios añadir una estructura arbitraria a sus documentos, pero sin decir nada sobre el significado de la misma, por lo que se puede considerar un meta-lenguaje para la definición de estructuras textuales.

2.2. Rdf

Es el sistema que permite utilizar metadatos para describir recursos (típicamente sitios web) en la web semántica. El objetivo de esta recomendación es habilitar la extracción del significado de la estructura de un documento, descrita en xml, con el fin de garantizar la interoperabilidad entre aplicaciones sin necesidad de intervención humana (**Senso**, 2003).

Todo el sistema rdf parte de tres entidades lógicas:

– Recursos.

– Propiedades.

– Valores.

Que se corresponden con los elementos de la lingüística:

– Sujeto.

– Predicado.

– Objeto.

Con los tres elementos anteriores podemos formar declaraciones sobre los recursos del tipo: el recurso X tiene la propiedad Y con valor P. La tabla siguiente (tabla 1) expresa las equivalencias de los componentes básicos de rdf.

Los recursos pueden ser sitios o páginas web, pero también cosas que no están en la www, como personas o cualquier objeto del mundo real o conceptual. Las propiedades son las características relevantes de los recursos (por ejemplo, con relación a las páginas web: el autor y el idioma). Por último, los valores son los datos en los que se concreta un atributo determinado de un recurso determinado. La tabla 2 expresa las ideas anteriores con dos ejemplos específicos aplicados a la descripción de dos sitios web utilizando *Dublin core*.

http://www.dublincore.org

http://www.imdb.com	dc.title	Internet movie database
http://allmovie.com	dc.title	All movie guide

Tabla 2: declaración rdf sobre dos sitios web

De acuerdo con la tabla anterior, hemos descrito dos recursos (en este caso dos bases de datos cinematográficas) mediante una de sus propiedades, concretamente el título de la página web. Para que un ordenador pueda entender este tipo de estructuras, denominadas triples, será necesario representar dicha información mediante rdf/xml y *Dublin core*. En la figura 1 mostramos esta representación para uno de los triples que aparecen en la tabla 2.

Ignoramos cómo evolucionará rdf pero, afortunadamente, la amplitud de miras de esta recomendación no es un obstáculo para su aplicación al mundo de la documentación, sino todo lo contrario: una de sus más importantes y significativas utilidades consiste en la descripción de recursos digitales (**Rovira**, 2007) utili-

```
<?xml version="1.0" ?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/element/1.1">

 <rdf:Description rdf:about="http://allmovie.com">

 <dc:title>All movie guide</dc:title>

 </rdf:Description>

</rdf:RDF>
```

Figura 1: Serialización de uno de los triples de la tabla 2

zando *Dublin core*, norma que, como es sabido, consiste precisamente en aplicar la filosofía documental a la descripción de recursos.

2.3. Owl

Con el objeto de que las máquinas puedan realizar tareas de razonamiento útil sobre los recursos de la web semántica, es necesario definir un lenguaje o herramienta de descripción que vaya más allá de las semánticas básicas de rdf y permita comparar y combinar documentos (recursos) con distinta estructura (es decir, que sea capaz de reconocer, por ejemplo, el elemento <*proveedor*> y <*provider*> de dos estándares para la gestión de transacciones comerciales como iguales, permitiendo la interoperabilidad entre ambos). A estos lenguajes o herramientas se les denomina ontologías, y básicamente incluyen las definiciones de los conceptos, denominadas "clases", de un dominio y las relaciones entre ellos.

Owl es el lenguaje estándar de la web semántica para expresar y codificar ontologías. Por tanto, puede ser utilizado para representar explícitamente el significado de términos en vocabularios y las relaciones (semánticas) entre ellos.

http://www.w3.org/TR/2004/REC-owl-features-20040210/

Consigue formalizar las relaciones entre las clases aún más que rdf, indicando aspectos básicos para el razonamiento como la existencia de conceptos o clases disjuntas en un dominio. Por ejemplo, "los periféricos de salida no son periféricos de almacenamiento", esto es, la clase de los periféricos de salida es disjunta a la clase de los de almacenamiento. También es posible expresar la cardinalidad, es decir, el número de elementos que pueden componer un concepto o clase, por ejemplo, "un libro puede tener uno o varios autores" (la cardinalidad de los autores de un libro es uno o más de uno), o bien "un libro solamente puede tener un isbn" (la cardinalidad del isbn de los libros es exactamente uno). Puede expresar igualdad o equivalencia entre clases, características y restricciones de las mismas, etc.

Owl utiliza rdf para representar y codificar las ontologías. Esta recomendación sigue la tendencia tan característica del *W3C* de proceder mediante "extensiones". Por tanto, owl es una extensión de rdf que añade elementos como los mencionados anteriormente para describir características y clases.

A modo de ilustración, en la figura 2 podemos ver un gráfico que representa un ejemplo de clases y subclases de una ontología de periféricos de ordenador:

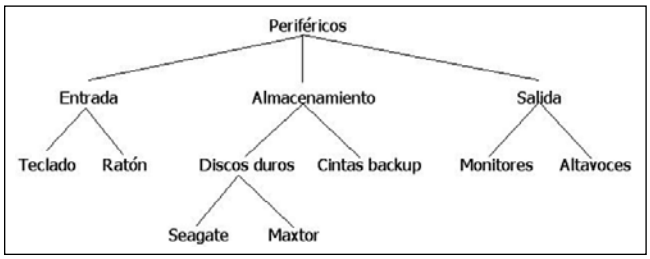

Figura 2: Clases y subclases de una ontología sobre periféricos de ordenador (Codina, 2006)

A continuación, en la figura 3, vemos parte de la ontología anterior representada mediante owl.

```
<?xml version="1.0" ?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
 ...
<owl:Class rdf:ID="Perifericos">
 <rdf:comment>
Los periféricos de ordenador están conectados a la CPU pero no forman parte de ella.
 </rdf:comment>
</owl:Class>

<owl:Class rdf:ID="Entrada">
 <rdf:comment>
Los periféricos de entrada son una subclase de periféricos de ordenador.
 </rdf:comment>

 <rdfs:subClassOf rdf:resource="#Perifericos" />
</owl:Class>

<owl:Class rdf:ID="Teclados">
 <rdf:comment>
            Los teclados son una subclase de los periféricos de entrada.
 </rdf:comment>

 <rdf:subClassOf rdf:resource="#Entrada" />

 <rdf:subClassOf rdf:resource="#Perifericos" />
</owl:Class>
...
 </rdf:RDF>
```

Figura 3: Representación en owl de la ontología de la figura 2

La idea esencial es que en algún momento del futuro (pero no parece probable que sea a corto plazo), la web no solamente estará poblada por un determinado número de ontologías que permitirán a los ordenadores realizar inferencias sobre la información publicada, sino que los agentes de usuario (esto es, los navegadores del futuro) serán capaces de realizar razonamientos fiables sobre tales ontologías. No obstante, el coste

requerido en tiempo y dinero para la generación de la infraestructura propia de la web semántica, la ausencia de alicientes que animen a los usuarios a adoptar las recomendaciones del *W3C* para el desarrollo de sus contenidos, y el desinterés de parte de los agentes técnico-comerciales más importantes de la web actual (como por ejemplo los proveedores de servicios de búsqueda y los fabricantes de navegadores), hacen augurar que incluso la fecha del 2010 que algunas veces se ha apuntado para este escenario resulte optimista.

Afortunadamente, entendemos que no es necesario esperar que se haga realidad (si es que alguna vez sucede) lo que podríamos denominar el "lado visionario de la web semántica" representado por el artículo de **Berners-Lee** del año 2001. Las ontologías son una tecnología que pueden aportar soluciones ahora a problemas reales actuales.

3. Ontologías: la nueva visión

En los últimos años, el proyecto de la web semántica ha servido para asentar lo que podríamos denominar el uso actual del término ontología (no olvidemos que ha gozado de varias vidas desde sus orígenes en la filosofía clásica). En este nuevo contexto, una de las definiciones más citadas es la de **Studer** (1998) -que completa la original de **Gruber** (1993)-, para quien una ontología es "una especificación explícita y formal de una conceptualización compartida". Una "conceptualización" es un modelo abstracto de algún fenómeno del mundo construido mediante la identificación de los conceptos relevantes a ese fenómeno (normalmente un dominio del conocimiento). "Explícito" significa que los conceptos utilizados en la ontología, y las restricciones para su uso, están claramente definidos. "Formal" se refiere al hecho de que debe ser comprensible para las máquinas, es decir, estar expresada mediante una sintaxis (como owl) que permita a un ordenador operar sobre ella. Por último, "compartida" refleja la noción de que contendrá conocimiento consensuado en algún grado (en el caso de un dominio del conocimiento, se supone que estará consensuado por los expertos en él) (**Gómez-Pérez**, 2005).

En este contexto, para que la especificación de un dominio se considere una ontología, debe presentar, al menos, dos tipos de componentes: elementos y relaciones entre los mismos. Los primeros son los siguientes (**Lacy**, 2005, pp. 32-40):

– Clases: las entidades del "mundo real" se pueden categorizar en grupos o conjuntos de objetos con similares características, forman las clases de la ontología. Las entidades pueden ser cosas físicas (p. e. automóviles) o conceptuales (p. e. teorías científicas). Algunos serían "País", "Libro" y "Automóvil". Constituyen el núcleo de una ontología y describen los conceptos de un dominio concreto. Como hemos señalado, un ejemplo de una clase podría ser "Automóvil", que idealmente representaría a todos los automóviles del mundo. De este modo, cada coche que vemos por la calle es una instancia o ejemplar de la clase "Automóvil". Una clase puede tener además subclases, que representan conceptos más específicos que el de su superclase. De este modo, podríamos dividirla por ejemplo en las subclases "Turismo", "Todoterreno" y "Deportivo".

– Propiedades: las entidades que pertenecen a una clase poseen atributos determinados, p. e. tienen un nombre, un color o un peso. Por tanto, las propiedades consisten en pares de atributo/valor y sirven para describir de forma conveniente las características relevantes de las entidades que forman las clases. Algunos ejemplos son "Población", "Isbn" y "Precio".

– Individuos, instancias o ejemplares: consisten en representaciones de objetos o elementos particulares de una clase. Se denominan indistintamente individuos o instancias de la clase. Hay que señalar que es difícil (y de hecho es discrecional) distinguir entre individuos y clases. Ejemplos de instancias son "España", "Documento 141203448-5" y "Ford Mustang".

Por su parte, las relaciones típicas de una ontología son las siguientes:

– Clase–Individuo: asocian individuos o instancias a una clase. Por ejemplo, "España" es una instancia de la clase "País", "Ford Mustang" lo es de "Automóvil" y "Documento 141203448-5" de "Libro". Se expresan mediante relaciones "es un" ("*is a*"), por ejemplo, "Ford Mustang es un automóvil".

– Individuo–Propiedad: como hemos señalado anteriormente, las instancias de una clase tienen valores asociados a propiedades. Estas asociaciones se expresan mediante relaciones "tiene el valor" ("*has value for*"). Por ejemplo, España tiene el valor "505 mil km2" para la propiedad "Extensión".

– Clase–Propiedad: la clase como un conjunto tiene propiedades. Cuando se aplican a una clase, estas propiedades se denominan restricciones porque sirven tanto para definirla como para delimitar la pertenencia de los individuos a ella. Por ejemplo, "Automóvil" posee la propiedad "tener un motor", que excluye de la misma a los vehículos de tracción animal.

– Clase–Subclase: las clases pueden tener subclases. Por ejemplo, "Todoterreno" es una subclase de "Automóvil". Esta asociación se expresa también con relaciones "es un".

Además de las anteriores, en una ontología también se dan otras clases de relaciones atendiendo a otros enfoques. En concreto, se suelen contemplar relaciones entre conceptos (clases) de sinonimia, antonimia, hipo-

nimia[1] y meronimia[2]. Algunas son similares a las que se contemplan en los tesauros. Además, cabe recordar que las relaciones clase-subclase y clase-individuo son la base de taxonomías y tesauros, de aquí la tendencia ya señalada a confundir las tres cosas.

"Las ontologías pueden considerarse lenguajes documentales con distintos niveles de estructura, pero a diferencia del tesauro tradicional están elaboradas con una sintaxis comprensible para los ordenadores"

Todas estas similitudes no deberían hacernos caer en el error de concebir una ontología como un tesauro (o como una taxonomía). Ciertamente, y al igual que un tesauro, las ontologías pueden considerarse lenguajes documentales con distintos niveles de estructura, pero a diferencia del tesauro tradicional, en primer lugar, están elaboradas con una sintaxis comprensible para los ordenadores. Además, como hemos visto, las ontologías contemplan un conjunto más amplio de relaciones que las de clase y subclase (como en una taxonomía) o las de sinonimia y meronimia (como en un tesauro) ya que en principio estas relaciones no están cerradas, sino que en parte dependen de las relaciones reales que se den entre las clases y los individuos del dominio modelado por la ontología. Por tanto, una ontología permite mayor riqueza en la definición de sus conceptos y sus relaciones que un tesauro.

Sin embargo, la diferencia más importante es el hecho de que están formalizadas, es decir expresadas mediante una rigurosa lógica formal y, por tanto, no solamente pueden ser procesadas por aplicaciones informáticas sino que, en principio (aunque con severas limitaciones) soportan procesos de inferencia automáticos.

3.1. Generación de ontologías

Acudiendo ahora a términos mucho más prácticos, y siguiendo la metodología especialmente relevante que propone **Noy** (2001), a grandes rasgos, el desarrollo de una ontología implica, al menos las siguientes fases:

– Definir las clases (conceptos).

– Ordenar las clases en una taxonomía.

– Definir las propiedades de las clases y los valores asociados a esas propiedades.

– Completar los valores de las propiedades para cada una de las instancias reales.

Debe tenerse en cuenta que no existe un modo correcto de modelar un dominio: siempre encontraremos distintas alternativas para hacerlo que nos proporcionarán diferentes resultados. Obsérvese que esta afirmación conlleva la concepción de las ontologías como instrumentos adaptados a la resolución de tareas, y por ende, la concepción de las mismas como conceptualizaciones "no" universales de los dominios que representan, lo cual hoy por hoy, choca frontalmente con la concepción universalista de las ontologías del *W3C*. Por tanto, el diseño de una ontología estará condicionado por su uso y nivel de detalle.

En cuanto a la complejidad asociada a su elaboración, el primer problema es determinar qué términos debemos enunciar y qué propiedades vamos a enumerar de éstos. Para solucionarlo es muy importante obtener a priori una lista de los términos que consideremos relevantes al dominio, sin preocuparnos de si existe solapamiento entre sus significados, o de las relaciones entre ellos. Las dos siguientes etapas serán desarrollar la jerarquía de clases (la taxonomía) y definir las propiedades de los conceptos.

Existen distintas aproximaciones para extraer la taxonomía de clases. Podemos recurrir a una aproximación arriba-abajo (*top-down*), que comienza con la definición de los conceptos más generales en el dominio para a continuación extraer conceptos más específicos. O por el contrario utilizar una metodología de abajo-arriba (*bottom-up*), mediante la identificación de las clases más específicas, y a continuación la agrupación de éstas en otras clases más generales. Aunque también se puede adoptar una aproximación mixta que combine los dos enfoques anteriores. En este caso, en primer lugar se identifican los conceptos más relevantes para el dominio de la ontología, y a continuación se generalizan o especializan según sea conveniente. No puede afirmarse que uno u otro método sea más apropiado para la extracción de la taxonomía, así que la selección del mismo dependerá de nuestra percepción del dominio.

3.2. Técnicas para la generación semiautomática de ontologías

Se espera que la aplicación de las especificaciones del *W3C*, junto con el desarrollo y generalización de las ontologías suponga el final de los problemas derivados de la ausencia de un modelo de datos bien definido en la web. No obstante, para que esto sea posible hay que solventar un nuevo problema, a saber: la estructuración y descripción de los recursos web mediante xml y rdf, así como que la elaboración manual de ontologías supone un coste tan elevado, en tiempo y dinero, que ya son muchas las voces que cuestionan que la transformación de la web actual en la web semántica pueda llegar a ser una realidad algún día.

"Muchos dudan de que la transformación de la web actual en la web semántica pueda llegar a ser una realidad algún día"

Con el fin de intentar paliar este problema ha aparecido una nueva disciplina, la ingeniería de ontologías, dedicada al estudio y diseño de aplicaciones que ayuden a su elaboración, mantenimiento y uso. Su principal objetivo es, por tanto, la creación de entornos que, mediante la automatización de ciertas tareas y el diseño del software para su gestión, agilicen el proceso. Ejemplos:

- *KAON*. *http://kaon.semanticweb.org/*
- *Hozo*. *http://www.hozo.jp/*
- *WebODE*. *http://webode.dia.fi.upm.es/WebODEWeb/index.html*
- *Protégé*. *http://protege.stanford.edu/*

Una comparación de estos sistemas puede encontrarse en **Mizoguchi**, 2004.

Mención especial merece la disciplina conocida como "Aprendizaje de ontologías" u "*Ontology learning*" (**Maedche**, 2004), una parte de la ingeniería de ontologías que investiga el desarrollo de métodos para la creación de una ontología de forma semiautomática. Concretamente, se centra en la generación de herramientas que permitan importar, extraer, podar, refinar y evaluar la taxonomía de una ontología bajo la supervisión de un experto humano, el "ingeniero ontológico", denominación acuñada en el ámbito germano, y cuyo perfil se corresponde en gran medida con el de un documentalista.

A continuación, para ilustrar el funcionamiento de estos sistemas se describe brevemente la arquitectura de una de las primeras propuestas formuladas en este ámbito (figura 4), la de **Maedche** (2001), que ha determinado las líneas básicas a seguir en este campo. La arquitectura propuesta consta de cuatro elementos:

– Interfaz gráfica: permite al ingeniero ontológico intervenir manualmente en todo el proceso de creación.

– Componente de gestión: con ella seleccionamos los datos a partir de los cuales construir la ontología (documentos html y xml, dtds, bases de datos, otras ontologías, etc.).

– Centro de procesamiento de recursos: facilita al ingeniero ontológico diferentes herramientas para procesar los documentos de entrada y extraer la terminología necesaria (conceptos).

– Por último, el sistema (véase *KAON*) dispone de una biblioteca de algoritmos cuyo funcionamiento se basa normalmente en reglas de asociación, técnicas de análisis formal de conceptos, o técnicas de agrupamiento (jerárquicas o no). Mediante la aplicación de uno o varios de estos algoritmos sobre los documentos ya procesados podrán extraerse las clases de la taxonomía y sus relaciones.

En teoría, mediante el uso de un sistema como el descrito, podría construirse una ontología siguiendo las siguientes etapas:

– Si es posible, importamos y reutilizamos las ontologías existentes en el dominio de nuestro interés, y un experto las fusiona (manual o automáticamente) en una única a partir de la cual aplicar el resto de fases.

– Extracción de ontologías: la herramienta propone diferentes entradas léxicas (términos) para la ontología, que se obtienen en función del procesamiento de los textos de los recursos del dominio seleccionados (documentos html, etc.). Independientemente de la recomendación hecha por el sistema, el ingeniero ontológico puede incluir o eliminar entradas léxicas si así lo desea. Obtenido el léxico, el siguiente paso es su clasificación taxonómica mediante técnicas automáticas de clasificación. El resultado final de esta fase es la propuesta de una taxonomía del dominio al ingeniero ontológico que éste puede modificar o rehacer como crea conveniente.

– Poda de ontologías: la arquitectura viene dotada de herramientas que permiten al ingeniero ontológico ajustar la ontología a su propósito original.

– Refinamiento: el sistema pone a disposición del profesional herramientas que permiten completar y afinar el resultado final.

– Evaluación de la ontología resultante: a través del seguimiento y observación de su uso.

– Actualización: para incluir nuevos dominios o actualizar los ya existentes.

Este modelo asume que cualquier ontología puede ser descrita por un conjunto de conceptos, relaciones y entradas léxicas. Consecuentemente, su construcción se puede agilizar utilizando tecnologías que analicen distintos tipos de recursos web (documentos html, xml, dtds, bases de datos, etc.) y extraigan los términos más significativos para nuestro dominio de interés así como sus relaciones. Todo ello bajo la supervisión de un experto humano en generación de ontologías.

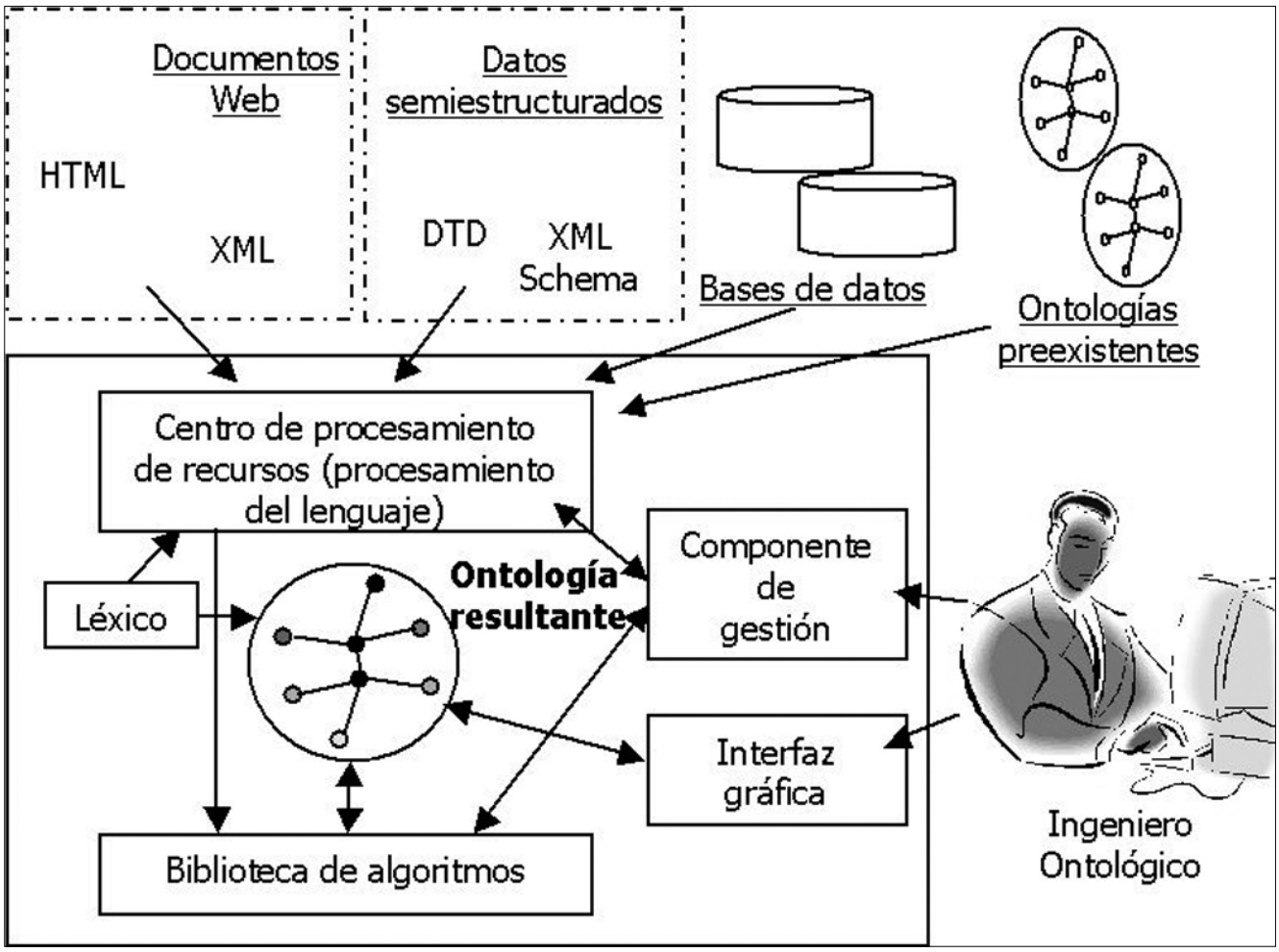

Figura 4: Arquitectura del sistema para el aprendizaje de ontologías propuesto por Maedche (2001)

4. Conclusiones

A la luz de todo lo expuesto es inevitable plantearse ciertas cuestiones que determinarán en gran parte el futuro. La primera de ellas quizás sea: cuando nos pronunciamos acerca de la web semántica ¿estamos hablando de una realidad palpable? La respuesta a esta pregunta es muy compleja. Si por realidad entendemos la existencia y funcionamiento de este entorno en la actualidad, puede afirmarse que no lo es.

En cambio, sí son realidades las iniciativas de investigación y desarrollo (públicas y privadas) puestas en marcha a raíz de la formulación de esta nueva Web y las recomendaciones del *W3C* como el lenguaje xml y rdf que están influyendo de forma activa y real en buena parte de la Web actual.

¿Significa esto que la web semántica será, entonces, una realidad en el futuro? Cada vez más analistas creen que es poco probable que se haga realidad el lado más visionario del proyecto, ni a corto ni a medio plazo.

De lo que no cabe duda es que aportará muchas cosas por el camino y provocará cambios duraderos y decisivos que ayudarán a tener una Web mucho mejor en el futuro. Un ejemplo fácil lo tenemos en la potente idea de la separación entre presentación y contenido de los documentos web.

¿Qué metodología se impondrá para la generación de esta futura Web: el procesado manual o el semiautomático? Por un lado, las descripciones de los contenidos y las ontologías elaboradas por expertos humanos son de gran calidad, aunque su coste en tiempo y dinero es inabarcable (además cabe la posibilidad de fraude); por otro, la utilización de herramientas automáticas para agilizar el desarrollo de las descripciones y las ontologías supone una disminución considerable de los costes, a cambio de descripciones y ontologías más someras (a menudo meras taxonomías o clasificaciones) que, además, pueden ser erróneas, y que difícilmente satisfacen las exigencias mínimas de las especificaciones del *W3C*.

Finalmente, quizás lo más probable (y apropiado) consista en una aproximación mixta, es decir, la coexistencia de descripciones y ontologías manuales junto con las automáticas. Las primeras, utilizadas cuando el dominio o tarea requiera descripciones y ontologías de gran calidad, y se disponga de los recursos necesarios. Las segundas, cuando el dominio o la tarea en cuestión suponga una tarea inabarcable para un experto

humano, bien por su magnitud, bien por su naturaleza cambiante.

En todo caso es de esperar que los logros aportados por este nuevo entorno web sean adecuadamente incorporados a cualesquiera otros escenarios dedicados a la gestión de la información documental. En particular, el profesional y el estudioso de la biblioteconomía-documentación no debería quedar al margen de estos avances.

"Es de esperar que los logros de este nuevo entorno web sean adecuadamente incorporados a la gestión de la información documental"

De hecho, la formación y experiencia de esta clase de profesionales hacen de ellos firmes candidatos a jugar un papel preferente en el desarrollo de la web semántica. Especialmente útil sería su participación activa en todas aquellas tareas conducentes a la descripción de los recursos de esta nueva web, poniendo especial énfasis en la importancia de las ontologías.

Probablemente el campo emergente de la ingeniería de ontologías favorecerá las aproximaciones semiautomáticas, asignando un papel preeminente al experto humano, el ingeniero ontológico, en el desarrollo de estas herramientas. Por ahora este nuevo perfil, muy similar al de un documentalista, parece vinculado exclusivamente a la informática; pero nada impide que, dada la similitud indicada, los profesionales e investigadores de nuestro campo tengamos también un papel más o menos protagonista. Existen, de hecho, otros campos profesionales y científicos que podríamos denominar "compartidos". En la arquitectura de la información, por ejemplo, podemos encontrar tanto a profesionales de la biblioteconomía-documentación como de la informática (y casos que comparten ambos perfiles, claro).

Corresponde a los documentalistas y profesionales de la información hacer visible su idoneidad para el desempeño de estas nuevas labores. Como expertos en la elaboración de lenguajes documentales y herramientas para el control terminológico deberían ser un agente más en la creación de esta nueva Web, evaluando tanto las recomendaciones como las herramientas para la descripción y recuperación de los nuevos recursos web, y asesorando a aquellos involucrados en su diseño. Sin duda alguna, los documentalistas y profesionales de la información están hoy en situación de adquirir las habilidades técnicas que les permitan desempeñar estas labores eficazmente. Si así lo hacen, probablemente asistiremos al nacimiento de una nueva dimensión que puede revalorizar considerablemente el perfil de este profesional de la información.

Notas de la Redacción

1. En semántica lingüística, se denomina hipónimo (del griego: υπον μιον, que literalmente significa 'pocos nombres') a aquella palabra que posee todos los rasgos semánticos, o *semas*, de otra más general, su hiperónimo, pero que añade en su definición otros rasgos semánticos que la diferencian de la segunda.
Ejemplo: *lunes*, *martes*, *miércoles*, etc. son hipónimos de *día*.
Fuente: *Wikipedia* en español.

2. La meronimia es una relación semántica no-simétrica entre los significados de dos palabras dentro del mismo campo semántico. Se denomina merónimo a la palabra cuyo significado constituye una parte del significado total de otra palabra, denominada ésta holónimo. Por ejemplo, *dedo* es merónimo de *mano* y *mano* es merónimo de *brazo*; a su vez, *brazo* es holónimo de *mano* y *mano* es holónimo de *dedo*.
Ejemplos: *azul* es merónimo de *color*; *doctor* es merónimo de *oficio*.
Fuente: *Wikipedia* en español.

Agradecimientos

Este trabajo ha sido financiado por el *Ministerio de Educación y Ciencia*, como parte del proyecto *HUM2004-03162/FILO*.

Bibliografía

Berners-Lee, T.; **Hendler, J**.; **Lassila, O.** "The semantic web". En: *Scientific American*, 2001, May, v. 284, n. 5, pp. 34-43.

Codina, L.; **Rovira, C.** "La web semántica". En: **Tramullas, J.** (ed.). *Tendencias en documentación digital*. Gijón: Ediciones Trea, 2006, pp. 9-54.

Extensible markup language (xml) 1.1 (W3C Recommendation 04 Feb 2004, edited in place 15 Apr 2004).
http://www.w3.org/TR/2004/REC-xml11-20040204/

Gómez-Pérez, A.; **Manzano-Macho, D.** "An overview of methods and tools for ontology learning from text". En: *The knowledge engineering review*, 2005, v. 19, n. 3, pp. 187-212.

Gruber, T. R. "A translation approach to portable ontologies". En: *Knowledge acquisition*, 1993, v. 5, n. 2, pp. 199-220.

Maedche, A.; **Staab, S.** "Ontology learning for the semantic web". En: *IEEE intelligent systems*, 2001, v. 16, n. 2, pp. 72-79.

Maedche, A.; **Staab, S.** "Ontology learning". En: **Staab, S.**; **Studer, R.** (eds.). *Handbook on ontologies*. Berlin: Springer, 2004, pp. 173-189.

Mizoguchi, R. "Ontology engineering environments". En: **Staab, S.**; **Studer, R.** (eds.). *Handbook on ontologies*. Berlin: Springer, 2004, pp. 275-296.

Noy, N. F.; **McGuinness, D. L.** "Ontology development 101: a guide to creating your first ontology". En: *Stanford Knowledge Systems Laboratory Technical report KSL-01-05*.

OWL Web Ontology Language: Overview (W3C Recommendation 10 Feb 2004).
http://www.w3.org/TR/owl-features/

RDF Vocabulary description language 1.0: RDF Schema (W3C Recommendation 10 Feb 2004).
http://www.w3.org/TR/rdf-schema/

Rovira, C.; **Marcos, M. C.** "Metadatos en revistas–e de documentación de libre acceso". En: *El profesional de la información*, 2006, marzo-abril, v. 15, n. 2, pp. 136-144.

Rovira, C.; **Marcos, M. C.**; **Codina, L.** "Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código". En: *El profesional de la información*, 2007, enero-febrero 2007, v. 16, n. 1, pp. 24-38.

Senso, J. A. "Herramientas para trabajar con rdf". En: *El profesional de la información*, 2003, marzo-abril, v. 12, n. 2, pp. 132-139.

Studer, S.; **Benjamins, R.**; **Fensel, D.** "Knowledge engineering: principles and methods". En: *Data and knowledge engineering*, 1998, n. 25, pp. 161-197.

***Rafael Pedraza-Jiménez**, Área de Biblioteconomía y Documentación, Universidad Pompeu Fabra, Barcelona.*
rafael.pedraza@upf.edu

***Lluís Codina**, Facultad de Comunicación Audiovisual, Universidad Pompeu Fabra, Barcelona.*
lluis.codina@upf.edu

***Cristòfol Rovira**, Área de Biblioteconomía y Documentación, Universidad Pompeu Fabra, Barcelona.*
cristofol.rovira@upf.edu