Vallez, Mari (2009). “La web semántica y el procesamiento del lenguaje natural”, en Lluis Codina; Mari Carmen Marcos; Rafael Pedraza-Jimenez (Coords) *Web Semántica y Sistemas de Información Documental*, Ed. Trea, Gijón, pp. 155-180.

Capítulo 6

La Web Semántica y las Tecnologías del Lenguaje Humano.

Mari Vállez

La implantación de la web semántica frente a la actual web supone un cambio de paradigma, ya que tiene que pasarse de una web basada y creada en lenguaje natural a una web estructurada y organizada, donde los contenidos etiquetados semánticamente serán el elemento principal. Este cambio supondrá una nueva filosofía y forma de trabajar, ya que el desarrollo y creación de contenidos para esta web requerirá una gran cantidad de esfuerzos. Es en este punto donde pueden intervenir las tecnologías del lenguaje humano para facilitar mecanismos y herramientas que ayuden a la implantación y expansión de este nuevo paradigma.

6.1 Introducción

La actual web esta formada por infinidad de documentos que utilizan el lenguaje natural para expresar sus contenidos y las etiquetas html para indicar su presentación a los navegadores. Así, el gran éxito de la WWW es haber llegado a establecer un formato universal, el lenguaje html, que es interpretado por cualquier navegador. La web semántica será una extensión de la actual, que va un paso más allá. Pretende dotar a los documentos de información y estructura semántica de una forma explícita para lograr que los sistemas informáticos, a partir de ahora agentes, puedan entender los textos (Berners-Lee, 01).

Actualmente la idea resulta más una utopía que una realidad a corto o medio plazo; sin embargo, muchos de los desarrollos realizados a raíz de la web semántica han dado lugar a nuevos servicios, implantados ya con gran éxito en la actual web. Por tanto, uno

de los logros que ya se pueden atribuir a la web semántica es haber conseguido instaurar el uso de distintos estándares para la manipulación de la información de una forma más sofisticada, que ya están siendo utilizados en diferentes áreas.

La semántica es considerada uno de los elementos clave de esta nueva fase de la web, ya que es donde reside gran parte de su potencial, pues a partir de la manipulación de elementos dotados de valor semántico pueden ofrecerse nuevos servicios que aún no existen en la actual web. De esta manera, las tecnologías semánticas, estrechamente relacionadas con las tecnologías del lenguaje humano, se han convertido en uno de los pilares de este nuevo paradigma, ya que pueden ayudar a su desarrollo e implantación.

Las tecnologías del lenguaje humano tratan de buscar mecanismos computacionales que permitan reconocer, comprender y generar lenguaje natural, para ello realizan un tratamiento automático de éste. Por tanto, intentan trasladar e integrar el conocimiento que las personas tenemos de la lengua en los agentes para que puedan emular las acciones que podemos realizar de forma innata. Para lograr este objetivo incorporan modelos teóricos, métodos y técnicas de diferentes disciplinas: lingüística, filosofía, psicología e ingeniería, ya que todas ellas están implicadas o pueden resultar útiles para tratar los diferentes procesos que envuelven el lenguaje natural. Cada una de ellas estudia la lengua desde puntos de vista y objetivos distintos, lo cual ha conllevado también el uso de terminología diferente para hacer referencia a la misma idea. La lingüística utiliza el término ‘lingüística computacional’, la ingeniería informática usa la expresión ‘ingeniería del lenguaje natural’. Sin embargo, el concepto más utilizado tradicionalmente por la comunidad científica es ‘procesamiento del lenguaje natural’, aunque actualmente está muy extendida la expresión ‘tecnologías del lenguaje humano’, especialmente en el marco de la Unión Europea. En este capítulo utilizaremos este último término al resultar más cercano y divulgativo para los profanos en esta disciplina.

Los contenidos de la actual web deben ser interpretados por las personas, ya que el uso del lenguaje natural como medio de expresión en muchos casos requiere deducir conocimiento implícito de los textos para poder ser comprendidos correctamente. Por ejemplo, para solucionar los problemas de ambigüedad propios del lenguaje natural es necesario contar con información que en algunos casos no está explícita en los textos pero que las personas somos capaces de extraer a partir del contexto.

El gran reto de la web semántica reside en conseguir que los contenidos estén dotados explícitamente de semántica para que a partir aquí los agentes sean capaces de deducir e inferir conocimiento. Como ya se ha comentado, actualmente todavía estamos en una fase muy embrionaria de la idea y los más pesimistas dudan que se llegue a buen puerto, pues la inteligencia artificial lleva décadas persiguiendo este objetivo sin gran éxito. Pero durante los últimos años, las tecnologías del lenguaje humano han madurado bastante y han logrado aplicaciones robustas y escalables que pueden ocupar un papel destacado en el desarrollo de la web semántica.

En este capítulo vamos a explorar el vínculo existente entre la web semántica y las tecnologías del lenguaje humano para ver en qué aspectos pueden complementarse ambas disciplinas. El estudio de la abundante literatura existente sobre el tema muestra principalmente la vinculación desde dos perspectivas. Por un lado las tecnologías del lenguaje humano pueden ofrecer facilidades para la creación y mantenimiento de la web semántica; y por otro lado también pueden ser la clave para acceder a los contenidos generados para la web semántica, al permitir el uso de lenguaje natural como medio de comunicación con los sistemas. También veremos como el desarrollo de la web

semántica puede elevar el nivel de las tecnologías del lenguaje humano al brindar nuevas herramientas y recursos lingüísticos que facilitarán el tratamiento del lenguaje natural en otras áreas, como pueden ser la enseñanza de idiomas, la traducción automática, etc. (Dini, 03; Buitelaar, 03b; Calzolari, 03).

6.2 La web semántica

El creador del concepto, Tim Berners-Lee, define la web semántica de la siguiente forma: «no es una web separada sino una extensión de la actual, donde la información está dotada de un significado bien definido, los ordenadores están mejor capacitados y las personas trabajan en colaboración» (Berners-Lee, 01). En estas líneas se reúnen los fundamentos de la web semántica: es una evolución de la actual web donde los ordenadores serán capaces de interpretar los documentos ya que estarán dotados de contenido semántico y el trabajo colaborativo será una realidad.

La *figura 1* muestra la arquitectura de la web semántica tal como la ha definido Berners-Lee. Se trata de una estructura de capas, donde cada nivel resulta un requisito previo para el siguiente.

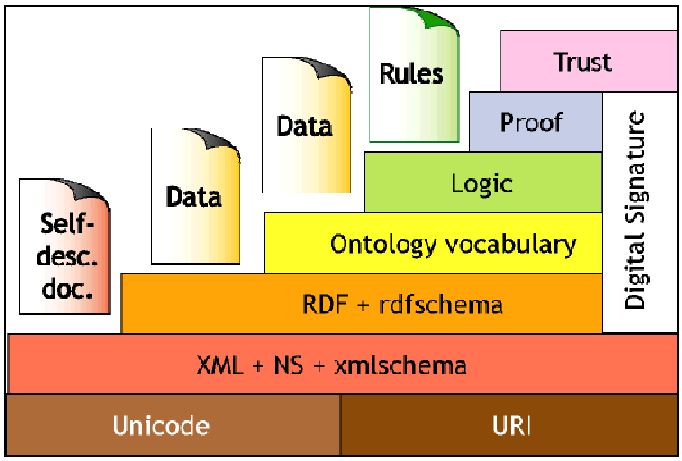

Figura 1.- Arquitectura de la web semántica (Fuente Tim Berners-Lee, http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html).

Los dos niveles iniciales hacen referencia a la base y los estándares en los que se sustenta su desarrollo: Unicode, URI, XML y RDF. Éstos permiten convertir la web en una infraestructura global donde será posible compartir y reutilizar datos y documentos entre diferentes tipos de usuarios.

La siguiente capa, referida a las ontologías, es en donde reside el contenido semántico del sistema. La cuarta capa tiene que permitir, a partir de la estructura semántica generada con las ontologías y los metadatos, realizar inferencias lógicas. Estas dos etapas son las que presentan más incógnitas actualmente y las que pueden actuar como freno para la implantación de la web semántica ya que comportan una infraestructura que actualmente no es posible realizar a gran escala. Es aquí donde las tecnologías del lenguaje humano pueden intervenir ayudando a la creación y mantenimiento de las ontologías y también vinculando éstas con los documentos, pero se trata de un camino aún en fase experimental.

Las dos últimas capas aún no se encuentran integradas en el sistema, pero si el proyecto de la web semántica avanza es previsible que sean las más rápidas en implantarse, ya

que la confianza y la seguridad son los elementos clave para dar un paso más allá en el comercio electrónico, que sería uno de los principales motores de la web semántica.

En (Codina, 06) puede obtenerse una visión global de la arquitectura, procesos e implicaciones de la web semántica.

6.3 Las tecnologías del lenguaje humano

Como ya hemos visto, las tecnologías del lenguaje humano pueden facilitar la conversión de la actual web, basada en el lenguaje natural, a la web semántica, donde la información tiene que estar estructurada.

Para observar el papel que ocupan las tecnologías del lenguaje humano en el desarrollo de la web semántica es interesante recurrir a la *figura 2* que muestra la "cadena alimentaria" de la información en el nuevo paradigma de la web semántica (Decker, 00).

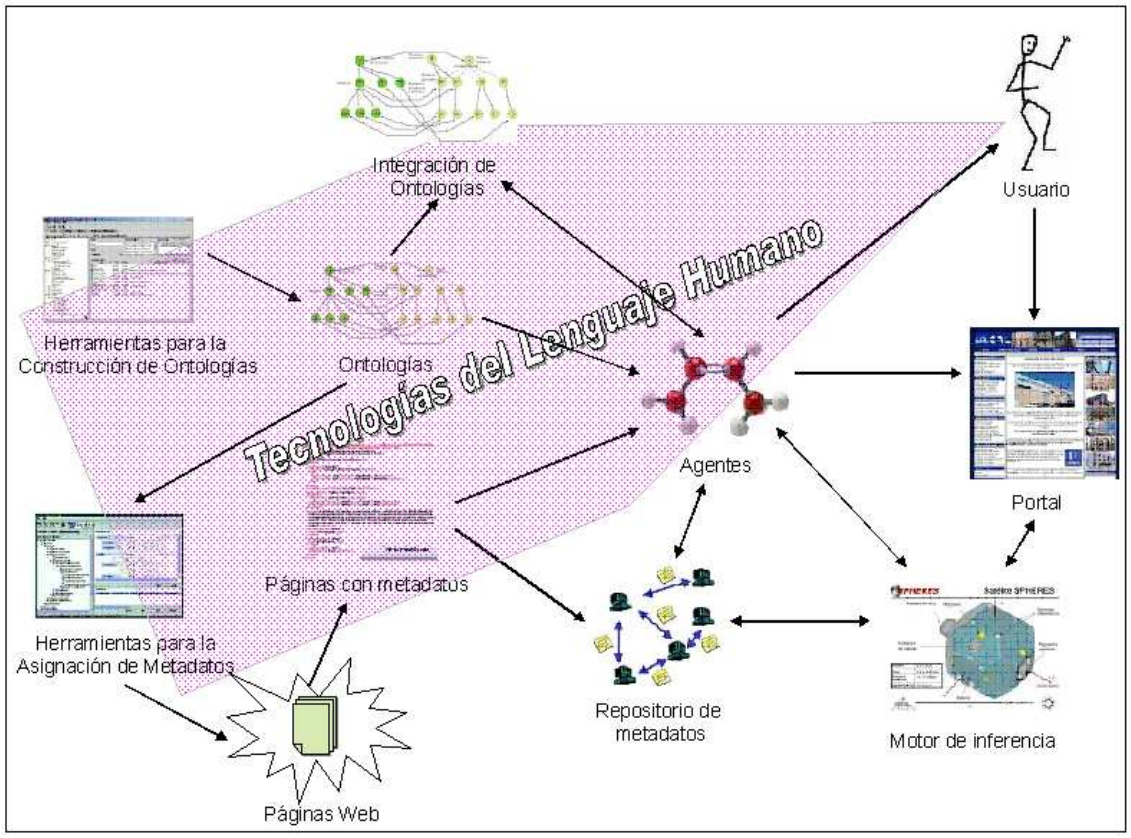

Figura 2.- Cadena alimentaria de la información.

En la *figura* se observa el proceso que deberá seguir la actual información, las páginas web, para llegar a estar dotadas de contenido semántico y de este modo tener la estructura requerida para la web semántica. Las páginas web estarán dotadas de contenido semántico gracias a herramientas que facilitarán la asignación de metadatos de una forma (semi)automática y que utilizarán las ontologías como base para ello. Por otro lado, la creación de ontologías como es una tarea también compleja contará con aplicaciones que ayudarán a su desarrollo. Y por último, estará el proceso de integración

de diferentes ontologías, ya que cada una dibuja un escenario de acuerdo con sus objetivos, y para tener una visión global de un dominio será necesario integrarlas todas.

Las páginas con metadatos pasarán a los agentes que se encargarán de procesarlas junto con más información, las ontologías, el repositorio de metadatos y el motor de inferencia. Éstos presentarán al usuario final la información pertinente, ya sea directamente porque ha formulado una necesidad de información al agente o a partir de los portales, que serán creados de forma más dinámica y facilitarán navegar por los contenidos.

En la *figura 2* además se destaca en que apartados podrán intervenir las tecnologías del lenguaje humano para facilitar la transformación de la actual web. Éstas estarán presentes en las herramientas que facilitan la creación de las ontologías, la asignación de metadatos y la integración de diferentes ontologías. Por otro lado, también formarán parte de los agentes que permiten al usuario acceder y recuperar la información pertinente de una forma más eficaz.

La tecnología del lenguaje humano más utilizada para la construcción de las herramientas y de los procesos citados es la extracción de información textual. Ésta permite obtener a partir de los distintos niveles de información extraída una representación de alto nivel de los textos, que facilita la conversión a los parámetros de la web semántica. Se trata de una técnica que ya existía antes de la conceptualización de la web semántica, y que surgió como una alternativa a los sistemas tradicionales de recuperación de información. La recuperación de información ofrece al usuario textos relevantes para una necesidad de información concreta, en cambio, los sistemas de extracción de información analizan los textos e intentan dar respuesta a la pregunta formulada, es decir, presentan únicamente la información que resuelve la necesidad de información.

6.3.1 La extracción de información

La extracción de Información es el término utilizado para la actividad de extraer automáticamente información específica de los textos en lenguaje natural con un formato previamente definido. El objetivo es extraer conocimiento estructurado de la información existente, generalmente texto no estructurado, con el fin de mejorar el uso y la reutilización de la información.

En este apartado veremos diferentes aproximaciones al proceso de extracción de información y de cómo enfocar esta técnica en la construcción de la infraestructura de la web semántica.

Los diferentes modelos utilizados en esta disciplina pueden agruparse en dos grandes categorías, los basados en el conocimiento y los que se apoyan en técnicas de aprendizaje automático. Los primeros se sustentan en la experiencia de la persona que los desarrolla para definir las reglas de extracción de información, el proceso de definición conlleva mucho tiempo, y por tanto la introducción de cambios en los sistemas es compleja porque en algunos casos puede suponer volver a redefinir el sistema. En cambio, los sistemas basados en el aprendizaje automático son creados por un proceso mecánico, con o sin supervisión de un especialista, requieren un gran número de ejemplos positivos y/o negativos para poder realizar deducciones, aprender, y así extraer y clasificar la información.

Este capítulo se va a centrar en los sistemas de aprendizaje automático que son los que se adaptan mejor a las necesidades de la web semántica, pues para que ésta llegue a

implantarse a gran escala se requieren sistemas donde la labor humana se pueda reducir al máximo, aunque actualmente hay que hablar de sistemas semiautomáticos donde la validación final del especialista es necesaria.

(Cunningham, 05 y Bontcheva, 03a) consideran la extracción de información como la técnica que podría ayudar en la fase inicial del desarrollo de la web semántica, pues facilitaría el traspaso de la actual web al nuevo paradigma como se observa en la *figura 3*. Definen la extracción de información como un proceso que toma los textos, y en ocasiones el habla, como entrada y que produce como salida formatos fijos, es decir, datos no ambiguos para ser exportados a la web semántica.

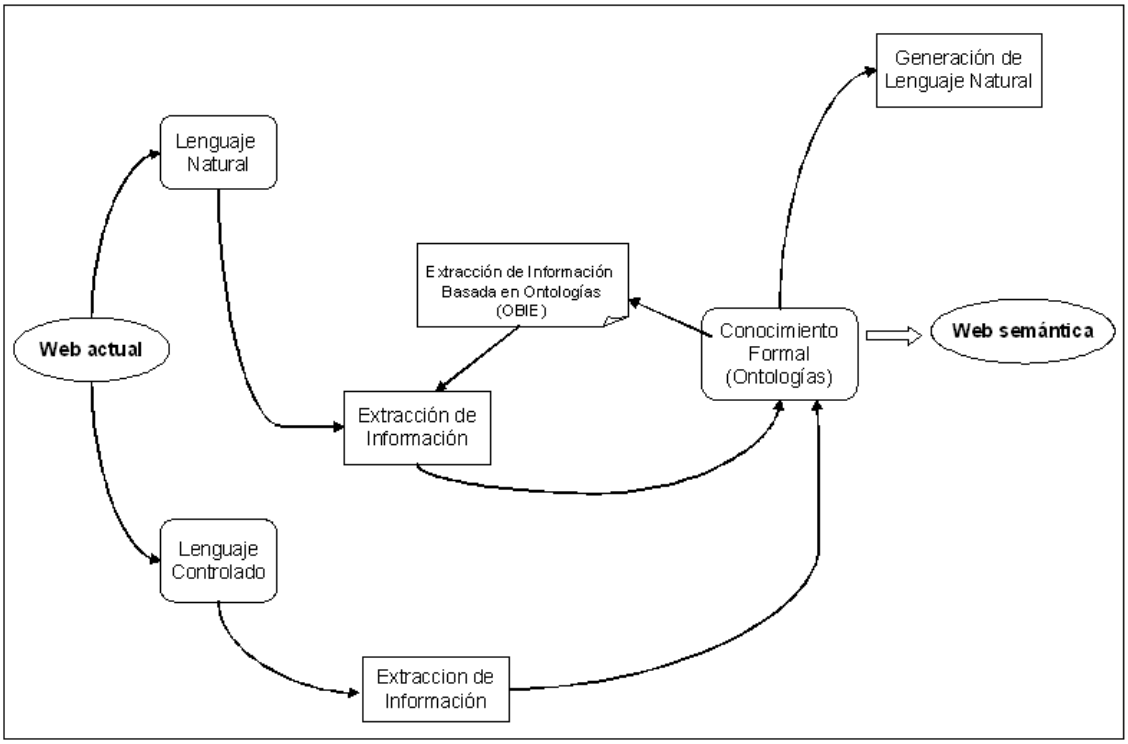

Figura 3-. Paso de la web actual a la web semántica.

Esta *figura*, adaptación de (Bontcheva, 03a), muestra como puede llegarse al conocimiento formal, articulado en las ontologías, con la aplicación de técnicas de extracción de información en los documentos expresados en lenguaje natural de la actual web. A partir de aquí pueden crearse contenidos con la arquitectura propia de la web semántica y también generar directamente contenidos en lenguaje natural en diferentes idiomas.

En la *figura 3* también se recoge la necesidad de que la extracción de información de los textos esté vinculada directamente a las ontologías, ya que la web semántica necesita que la información tenga una estructura jerárquica para poder realizar las inferencias. Se trata de una nueva versión del proceso tradicional de extracción de información, conocida con el nombre de extracción de información basada en ontologías (OBIE, Ontology-Based Information Extraction), donde las ontologías son uno de los componentes claves de los sistemas de extracción de información. Como ya veremos con más detalle, los dos retos principales de esta técnica son: identificar las instancias de una ontología en un texto y poblar automáticamente una ontología con las nuevas instancias identificadas.

Otra idea que los autores plasman en la *figura 3* es el uso de un lenguaje controlado en la actual web como alternativa para favorecer el desarrollo de la web semántica, ya que el proceso de extracción de información se simplificaría considerablemente. El concepto de lenguaje controlado es entendido como un subconjunto del lenguaje natural con algunas restricciones gramaticales y léxicas que permiten eliminar la ambigüedad propia del lenguaje natural y así facilitar la extracción de información y la interpretación de los textos. Proponen utilizar un lenguaje controlado porque es un modo bastante natural de expresarse y un mecanismo útil para controlar la ambigüedad, además de ser una opción válida para que la información sea procesada por los agentes de forma precisa.

El proceso de extracción de información, en su vertiente clásica, intenta identificar distintos tipos de información, patrones predefinidos, en los textos expresados en lenguaje natural. En (Cunningham, 05) podemos ver las diferentes fases establecidas en el proceso, explicadas con ejemplos, también indica los porcentajes promedios de éxito de cada una de las fases. La información que se extrae de los textos es la siguiente:

- las entidades, que son las unidades básicas del texto; por ejemplo, los lugares, las personas, las organizaciones, las fechas, etc.
- las menciones, que son las diferentes formas que se usan en un texto para referirse a una misma entidad.
- las descripciones de las entidades.
- las relaciones que se establecen entre las entidades.
- los eventos en los que están implicadas las entidades.

Vamos a ver con un texto de ejemplo la distinta información que se obtendría siguiendo el proceso de extracción de información expuesto:

"El banco BBVA ha hecho públicos los cambios producidos en su equipo directivo. Luis Martín ha dejado el cargo de Vicepresidente de la entidad para pasar a ocupar la presidencia. El banco ha nombrado a Rafael Bosque, que ha trabajado estrechamente durante varios años con L. Martín, nuevo Vicepresidente."

El sistema de extracción de información reconoce a *Luis Martín* y *Rafael Bosque* como entidades del concepto persona y a *BBVA* como entidad del concepto organización. Esta primera fase es el reconocimiento de las entidades nombradas y consiste en identificar en el texto los diferentes tipos de entidades que aparecen. El siguiente paso es identificar las diferentes menciones que se realizan en el texto de una entidad, en el ejemplo *banco* y *entidad* son una mención de *BBVA* y *L. Martín* hace referencia a *Luis Martín.* Estas dos primeras fases son las principales del proceso de extracción de información. La información aportada sobre cada entidad es la descripción, *Luis Martín* es descrito como anterior Vicepresidente y nuevo Presidente del Banco BBVA, y *Rafael Bosque* como nuevo Vicepresidente. Las relaciones establecidas entre las entidades son que *Luis Martín* y *Rafael Bosque* forman parte del equipo directivo del BBVA. Por último, el evento en que están implicadas las entidades es el nombramiento de los cargos directivos. La *figura 4* muestra en un esquema todo el proceso.

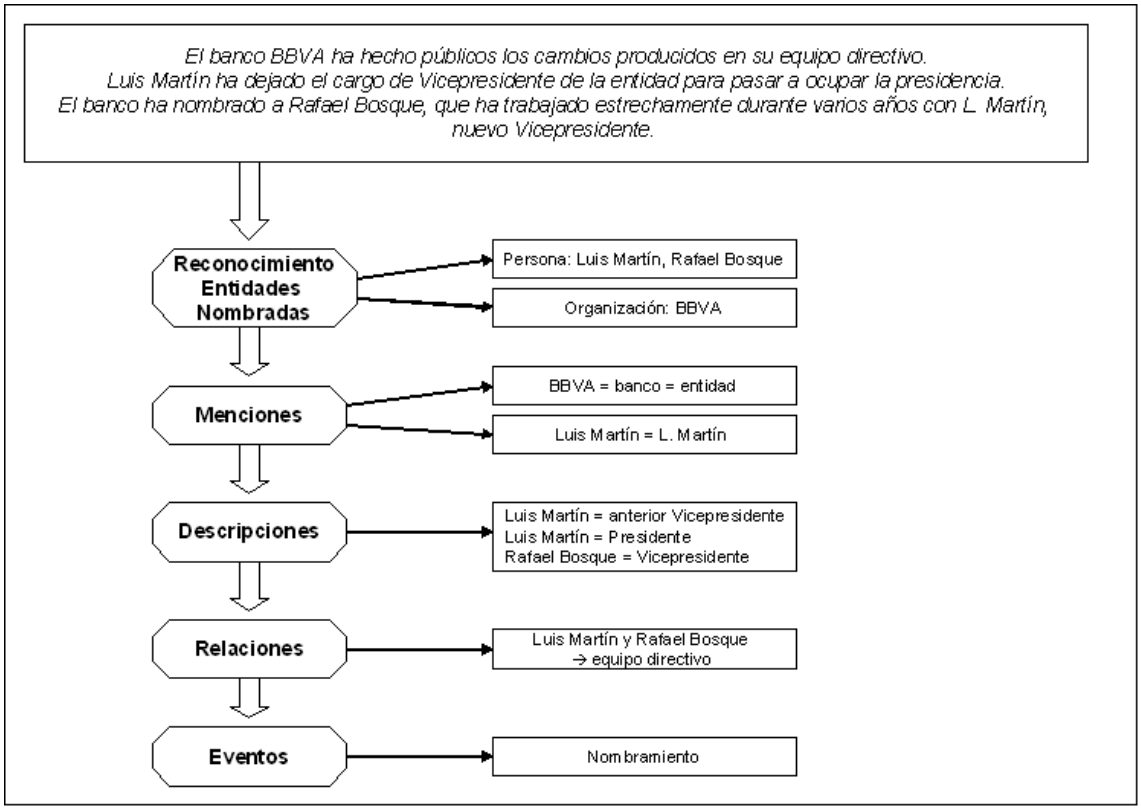

Figura 4.- Ejemplo de extracción de información.

A partir de los distintos niveles de información extraída puede obtenerse una representación de los textos, que permite explicitar formalmente el significado, y de este modo facilitar la asignación de metadatos y ayudar en el desarrollo de las ontologías.

Por otro lado, hay sistemas que se apoyan más en la extracción de información a partir de un análisis lingüístico, en (Buitelaar, 03b y Java, 07) se presenta el diseño de un modelo para la integración de información lingüística que se basa en el proceso de extracción de información a partir de ontologías. Se trata de un análisis lingüístico de los textos para obtener una representación de su contenido a partir de los diferentes tipos de información extraída.

En la *figura 5* se muestran las distintas fases de análisis lingüístico al que es sometido un texto para obtener una representación de su significado. El gráfico también recoge las fuentes de conocimiento utilizadas para realizar el análisis. Unas dependen de la lengua en que trabaja el sistema, son las reglas (morfológicas, sintácticas y semánticas) y también el léxico semántico. Las otras son independientes, se trata de las ontologías y de la mayoría de las reglas de pre-procesamiento

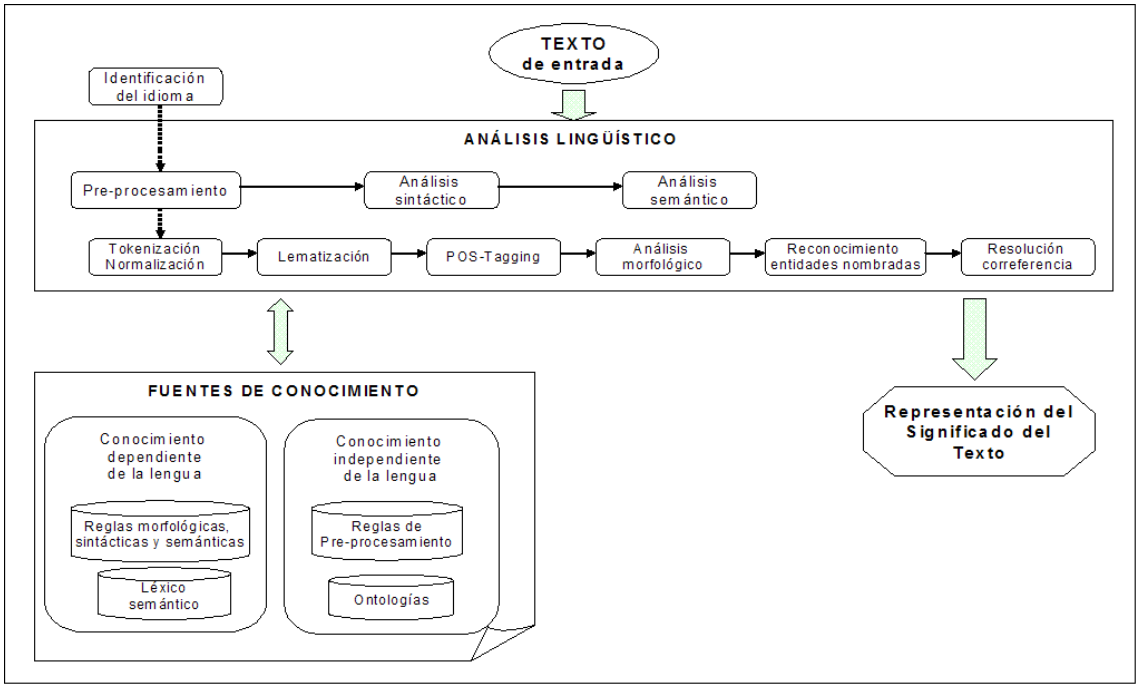

Figura 5-. Proceso de extracción de información.

A continuación, vamos a explicar las principales etapas del análisis llevado a cabo en el proceso de extracción de información que se muestra en la *figura* anterior.

Primero se *identifica el idioma* del texto para poder aplicar las reglas adecuadas. Para ello el sistema consulta los términos del texto en los diccionarios de las distintas lenguas y de este modo localiza a que lengua pertenecen.

El siguiente paso es realizar el *pre-procesamiento del texto* para después poder llevar a cabo el análisis sintáctico y semántico. Esta etapa se divide en diferentes subfases. Primero la *tokenización y normalización* de los textos para establecer los términos y límites de las frases, eliminar los elementos del documento que no aportan contenido y homogeneizar los documentos en referencia a las mayúsculas y minúsculas, acrónimos, cantidades numéricas, fechas, etc. Después se lleva a cabo la *lematización* de los términos para obtener el lema y así poder representar con una única forma diferentes variantes; por ejemplo, “informa” es el lema de “información”, “informador” e “informar”. A continuación, se realiza el *POS tagging (Part-Of-Speech)*, es decir, el análisis y etiquetaje de la categoría gramatical de las palabras de la frase para indicar si se trata de un nombre, adjetivo, verbo, etc. En algunos casos es necesario aplicar técnicas de desambiguación para saber la categoría; por ejemplo, la palabra “sobre” puede ser nombre, preposición o verbo, y depende de cada contexto el valor que se asigna. También es llevado a cabo el *análisis morfológico* para tratar la derivación y la composición de las palabras, ya que el castellano, igual que otras lenguas románicas o el alemán, es muy rico morfológicamente. Las dos últimas fases del pre-procesamiento son el *reconocimiento de las entidades nombradas* y la *resolución de la correferencia*, se trata de las fases principales de los sistemas de extracción de información, ya que se considera que las entidades son las que aportan básicamente el significado de los textos. En algunos casos hay que aplicar técnicas de correferencia para identificar las diferentes menciones utilizadas a la hora de referirse a una misma entidad. En el ejemplo anterior, de los nombramientos en una entidad bancaria, se ha explicado en qué consiste el reconocimiento de las entidades.

Después de realizar el pre-procesamiento del texto se lleva a cabo el *análisis sintáctico* para identificar las relaciones que se establecen entre las palabras al formar unidades superiores, los sintagmas y las frases. Esta etapa permite identificar las descripciones, relaciones y eventos en los que están implicadas las entidades. Se lleva a cabo un análisis sintáctico superficial, no exhaustivo, para conseguir un sistema que no consuma demasiados recursos y que además sea robusto frente a posibles errores de las estructuras sintácticas.

Por último, la fase del *análisis semántico* se encarga de dotar de contenido semántico a la frase a partir de la estructura proporcionada en el análisis sintáctico. Uno de los procesos principales de esta etapa es la subcategorización de los verbos, es decir, identificar los diferentes argumentos que acompañan al verbo y la función sintáctica/semántica que les corresponde. En este nivel también hay que resolver casos de ambigüedad para seleccionar el valor semántico correcto de un término en un determinado contexto, como se verá en el ejemplo de la *figura 6*.

El proceso de análisis lingüístico explicado con detalle permite hacerse una idea de la complejidad de estos sistemas, en especial si se implantan a gran escala. Por ello se van creando alternativas que simplifican el proceso al verlo nada más desde una perspectiva, como veremos en el siguiente ejemplo.

La *figura 6* muestra un ejemplo concreto del análisis realizado por un sistema que se apoya en la semántica ontológica (Niremburg, 01), que considera que el valor de los elementos casi exclusivamente desde el punto vista de la estructura semántica. El ejemplo se ha obtenido utilizando la tecnología del buscador Hakia[1], aún en versión beta, que permite realizar el análisis de una frase a partir de la semántica de los términos que la forman.

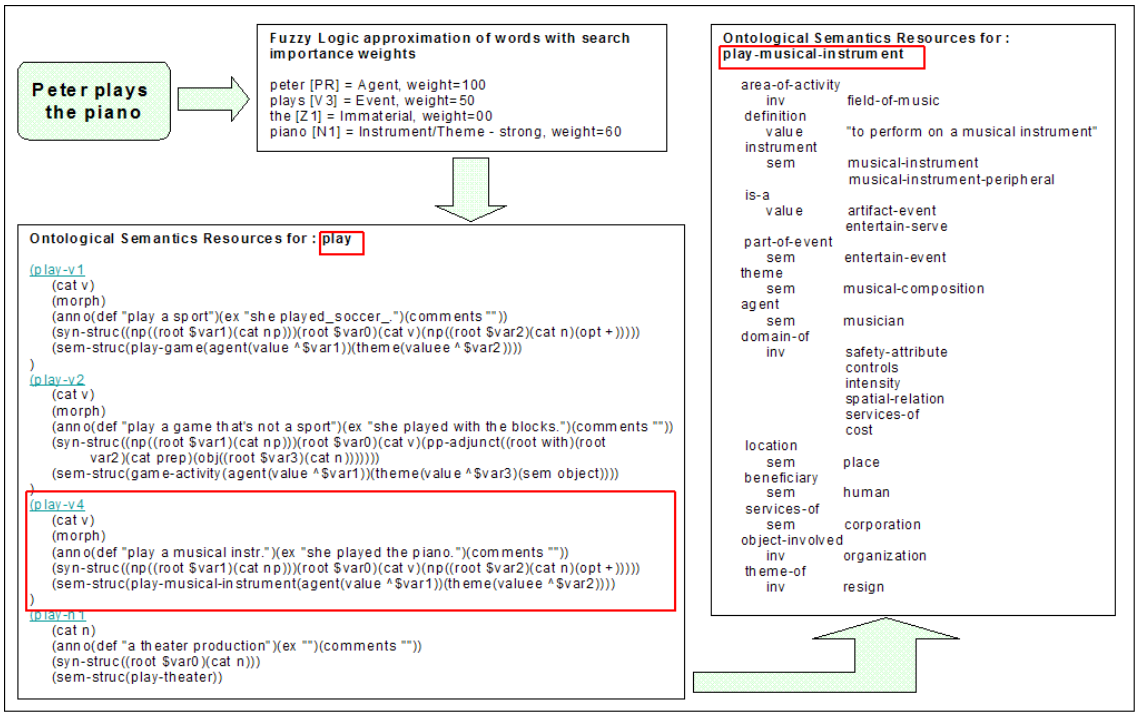

Figura 6.- Análisis basado en la semántica ontológica.

[1] http://www.hakia.com/

En el ejemplo de la *figura* primero son identificados los diferentes elementos de la frase y se otorga a cada entrada léxica una función semántica: agente, evento, tema. Después se consultan los diferentes términos en una ontología y se obtienen los diferentes sentidos de las palabras. De cada entrada se facilita: la categoría gramatical, una breve definición con algún ejemplo, la estructura sintáctica más habitual, y por último la estructura semántica, donde se explicitan los distintos elementos que requiere. En este ejemplo la palabra *play* es ambigua, pero a partir de la estructura definida para cada uno de sus posibles significados se asigna el valor correspondiente. En este caso al ir seguida de un instrumento musical se selecciona el sentido *tocar*. En la parte derecha de la *figura* se muestra la estructura de la entrada de la ontología, se trata del conjunto de propiedades que especifican el significado de un concepto relacionándolo con otros conceptos de la ontología.

En (Cimiano, 05 y Alani, 03) se detallan la implementación de diferentes sistemas de extracción de información que llevan a cabo un análisis lingüístico de los textos. El primero describe el sistema GenIE de extracción de información del ámbito de la bioquímica y se observan con detalle los diferentes procesos lingüísticos llevados a cabo para implementar un sistema de este tipo. El segundo presenta el sistema Artequakt que utiliza herramientas de procesamiento del lenguaje natural para extraer conocimiento sobre artistas a partir de diferentes textos, basándose en una ontología previamente definida sobre el dominio, y a partir de aquí poder generar la biografía de los artistas.

6.3.2 Las ontologías

Como ya se ha visto en el esquema de Berners-Lee, las ontologías son una de las piezas clave de la web semántica. Se trata de especificaciones formales y explícitas que representan los conceptos de un determinado dominio y sus relaciones (Gruber, 93). Es decir, son un modelo abstracto de un dominio, donde los conceptos utilizados están claramente definidos y pueden ser utilizados directamente por los agentes. De este modo los agentes pueden interpretar el significado de un texto y a partir de aquí buscar e integrar datos.

Las ontologías se usan principalmente para favorecer la comunicación entre organizaciones, personas y aplicaciones, permitiendo la interoperabilidad entre sistemas informáticos con el objetivo final de llegar al razonamiento automático. Gracias al conocimiento almacenado en las ontologías, los agentes podrán extraer automáticamente datos de las páginas web, procesarlos y sacar conclusiones de ellos; sin embargo, esta funcionalidad actualmente todavía está lejos de ser una realidad.

La formalización del conocimiento en las ontologías es una de las barreras iniciales para la implantación de la web semántica, ya que la construcción de ontologías es un proceso extremadamente lento, costoso y propenso a los errores, y además requiere un gran esfuerzo y especialización que muchas organizaciones no tienen a su alcance. Por tanto, es necesario crear mecanismos para promover su desarrollo de una forma (semi)automática.

Actualmente ya existen diferentes métodos y herramientas que ayudan a la creación y desarrollo de las ontologías de una forma (semi)automática; no obstante, como no existe un consenso en la comunidad científica a la hora de concretar las distintas fases implicadas en el desarrollo de una ontología, el proceso de evaluación y comparación de los diferentes sistemas se complica enormemente.

La ingeniería de ontologías es el nombre de la disciplina encargada del estudio y construcción de las diferentes herramientas que tienen por objetivo diseñar mecanismos que agilicen el proceso de construcción de ontologías de dominio. Dos de las etapas principales de este proceso son el aprendizaje y la población de ontologías. Aquí es donde las tecnologías del lenguaje humano pueden tener un papel destacado, pues facilitan la extracción de los conceptos, de las relaciones y de las instancias, como ya se ha explicado.

En la *figura 7* se muestran las dos etapas citadas anteriormente y como se relacionan entre sí. A partir de los textos de un dominio y utilizando técnicas y recursos del procesamiento del lenguaje natural se extraen los conceptos y relaciones de las ontologías. Por otro lado, desde las ontologías se recurre a los textos para extraer las instancias de los conceptos usando del mismo modo técnicas lingüísticas.

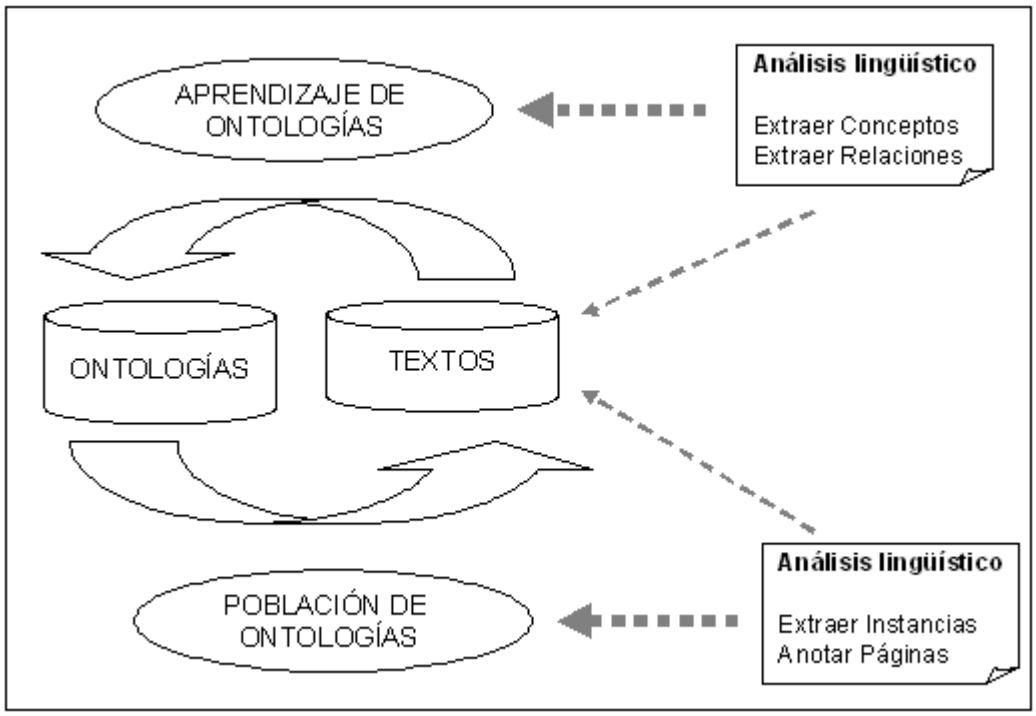

Figura 7.- Aplicación de las tecnologías del lenguaje humano en las ontologías

El análisis lingüístico aplicado a los textos puede ser de diferentes niveles según cada sistema, aunque casi todos se apoyan en el principio de que el significado de una palabra está estrechamente relacionado con el contexto en que aparece. El proceso consiste en identificar las dependencias sintácticas y semánticas de los ítems, basándose en el principio de restricción seleccional (Gamallo, 01), donde se considera que la estructura sintáctica facilita información relevante a cerca del contenido semántico. Otra de las técnicas más utilizadas es la localización de patrones léxico-sintácticos, ya que éstos implican restricciones semánticas a los términos.

6.3.2.1 El aprendizaje de ontologías

El aprendizaje de ontologías se define como la adquisición de los términos que representan un dominio a partir de datos. Tiene por objetivo principal crear una jerarquía de conceptos, es decir, una taxonomía. Existen diferentes modalidades de sistemas de aprendizaje de ontologías dependiendo de los tipos de datos de entrada

utilizados para el aprendizaje: textos, diccionarios, bases de conocimiento, datos semi-estructurados y bases de datos (Maedche, 01).

El aprendizaje de ontologías a partir de textos extrae el conocimiento de éstos, estudiando como los términos son utilizados en su contexto. Uno de los obstáculos de este proceso es la creación y diseño del conjunto de textos que deben ser utilizados para que el sistema aprenda ya que deben cubrir todos los ámbitos del dominio. Para ello algunos sistemas recurren a un glosario de términos del dominio para realizar la selección de los textos.

El proceso consiste en extraer los conceptos, los distintos términos usados para referirse a un concepto, los diferentes sinónimos de los términos, las relaciones entre los conceptos, por ejemplo las jerárquicas y las de causa-efecto, y por último las reglas y axiomas, que son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología (Cimiano, 06).

(Omelayenko, 01) realiza un análisis y comparación de las diferentes aproximaciones para el aprendizaje de ontologías. Los criterios seleccionados para la comparativa son los tipos de ontologías, las diferentes etapas del ciclo de vida de una ontología y las técnicas de aprendizaje automático utilizadas. Para acabar es interesante citar a (Gómez-Pérez, 05) ya que realiza un extenso recorrido por las diferentes aproximaciones del aprendizaje de ontologías a partir de textos y las agrupa según las tecnologías que utilizan para adquirir la información: técnicas estadísticas, algoritmos de aprendizaje automático y técnicas lingüísticas, destacando estas últimas.

6.2.3.2 La población de ontologías

La población de ontologías consiste en enriquecer una ontología con las instancias de los conceptos y de las relaciones extraídas de los textos de una forma (semi)automática. Muchas de las aproximaciones realizadas a partir de este enfoque se basan en la extracción de las entidades. Es un proceso muy semejante al descrito en la *figura 4*, aunque no tan complejo. Concretamente se centran en la identificación, extracción y clasificación de las entidades nombradas (EN); es decir, de los nombres de personas y de organizaciones, y también de los lugares, fechas y medidas (porcentajes, peso, monedas, tiempo) pues se considera que en ellas reside gran parte del contenido semántico de un texto.

El proceso consiste en reconocer las EN como instancias de alguno de los conceptos de la ontología y de este modo ir poblando una ontología. El reconocimiento de EN no es un proceso trivial ya que existen algunos conflictos que requieren la aplicación de técnicas lingüísticas para resolverlos como veremos en los siguientes ejemplos.

En un texto pueden existir distintas variantes de una misma instancia; por ejemplo, *Juan López, Sr. López, Juan, él*, son distintas formas utilizadas para referirse a la misma persona. Tienen que aplicarse técnicas de correferencia y de resolución de la anáfora pronominal para identificar que expresiones utilizadas anteriormente en el texto inciden directamente en la interpretación de los elementos que aparecen más tarde. Por otro lado, una EN puede ser ambigua, tener dos o más posibles interpretaciones, por ejemplo el término Adolfo Domínguez puede referirse tanto a una persona como a una organización, por tanto, para seleccionar el valor adecuado deben aplicarse técnicas de desambiguación.

En (Angelova, 05) pueden observarse distintas aproximaciones para el enriquecimiento de ontologías. Dos de ellas se sustentan en la información lingüística obtenida del

análisis de los textos, la primera se basa en el procesamiento a nivel de palabras, y la segunda en la estructura sintáctica de los conceptos.

6.3.3 La asignación de metadatos

Uno de los elementos fundamentales de la web semántica son los metadatos; es decir, información/datos que describen los contenidos de los documentos a los que están asociados representando de forma explícita el significado de éstos. El modelo de representación de los metadatos en la web semántica es RDF (Resource Description Framework). Se basa en describir los recursos como expresiones con la forma sujeto-predicado-objeto. El sujeto es el recurso, es decir, aquello que se está describiendo. El predicado es la propiedad o relación que se desea presentar del recurso. Por último, el objeto es el valor de la propiedad o el otro recurso con el que se establece una relación. Esta estructura presenta limitaciones, por eso se combina con RDF Schema y OWL que permite representar relaciones semánticas más complejas que las descritas. La siguiente *figura* muestra un ejemplo de metadatos en RDF y OWL.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<owl:Ontology rdf:about="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <dc:title>The RDF Vocabulary (RDF)</dc:title>
   <dc:description>This is the RDF Schema for the RDF vocabulary.</dc:description>
</owl:Ontology>

<rdf:Property rdf:about="http://www.w3.org/1999/02/22-rdf-syntax-ns#type">
  <rdfs:isDefinedBy rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
  <rdfs:label>type</rdfs:label>
  <rdfs:comment>The subject is an instance of a class.</rdfs:comment>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
</rdf:Property>
....
</rdf:RDF>
```

Figura 8.- Ejemplo de metadatos en RDF y OWL.

La asignación de metadatos es un proceso complicado que para su implantación requiere de mecanismos que agilicen esta fase, igual que para el traspaso de los contenidos de la actual web a la web semántica. Se trata de modificar la forma de crear los contenidos, ya que debe asignarse de forma explícita la carga semántica de los documentos. Actualmente los lenguajes utilizados en la web (html, dhtml, xhtml, xml, xlst,…) nada más indican como tienen que presentarse (visualizarse) los contenidos sin tener en cuenta la carga semántica de éstos, aunque hay que destacar que ha aparecido una nueva tendencia, los microformatos, que pretende facilitar el proceso de agregar contenido semántico en las páginas (x)html. Sin embargo, la web semántica va más allá, precisa que la información esté estructurada, dotada de valor semántico de acuerdo con unas estructuras previamente definidas, las ontologías, para de este modo poder tratar y manipular la información con el objetivo de ofrecer nuevos y mejores servicios a los usuarios.

Existen distintos modelos para la asignación de metadatos, aunque para resumir pueden agruparse en tres grandes categorías. El primero se basa en las ontologías y las utiliza como recurso central para relacionar y establecer conexiones entre los términos extraídos de los textos y de este modo representar el significado. El segundo modelo se basa en la anotación lingüística, concepto propio de la lingüística de corpus que tiene por objetivo etiquetar los textos a partir de los diferentes niveles de la lengua, para

identificar los términos y saber como éstos se relacionan entre sí, ya que esta información puede modificar el valor de un término. El último grupo propone el uso de un lenguaje natural controlado, como puede ser un tesauro, para agilizar y simplificar la asignación de metadatos. Éste es el modelo tradicional que ha utilizado durante años la biblioteconomía y la documentación.

El proceso que permite entrelazar los modelos semánticos con el lenguaje natural se conoce con el nombre de anotación semántica. Consiste en establecer de forma (semi)automática interrelaciones entre las ontologías y los textos; es decir, se enlazan los términos de los documentos con los conceptos de la ontología. Este modelo corresponde a las propuestas presentadas en el apartado anterior para la población de las ontologías y la técnica utilizada principalmente es el reconocimiento de las entidades nombradas.

El sistema basado exclusivamente en la anotación lingüística de los distintos niveles de la lengua (léxico, morfológico, sintáctico, semántico y pragmático), que en muchos casos se queda en el nivel sintáctico o semántico, no puede plantearse actualmente como una alternativa ya que requiere herramientas y recursos que no son aún suficientemente eficientes.

Existen los modelos híbridos (Buitelaar, 03b y Aguado, 02) donde se propone la integración de la anotación lingüística y la semántica ontológica para obtener una visión más completa del significado de los textos. Estos sistemas son los que en la actualidad están teniendo una mayor introducción. Consideran que el contenido de un texto no es exclusivamente la lista de conceptos que lo componen y las relaciones que establecen entre ellos, sino que las formas y las estructuras lingüísticas también aportan información muy valiosa que hay que tener en cuenta.

Por último, algunas propuestas que promueven el uso de un lenguaje natural controlado para la asignación de metadatos son las siguientes. (Schwitter, 05) muestra como el uso de un lenguaje natural cercano a la lógica de primer orden y el uso de una herramienta para facilitar la utilización de éste pueden facilitar la asignación de contenido semántico y además con la ventaja de hacerlo accesible e interpretable tanto para los agentes como para las personas. (Katz, 02) propone la utilización de anotaciones en lenguaje natural dentro de la sintaxis RDF, para así favorecer su implantación ya que resultaría también comprensible para las personas y no sería exclusivo de los ordenadores.

6.3.4 Herramientas y recursos para la web semántica

Las herramientas y recursos desarrollados bajo el paraguas de la web semántica han crecido muy rápidamente durante estos últimos años, aunque también incluyen los que provienen de otras áreas de conocimientos afines, la extracción de información, la clasificación y recuperación de información, el procesamiento del lenguaje natural, etc.

La mayoría de las herramientas se encuentran aún a nivel experimental y pocas son las que han pasado a la etapa comercial, lo cual implicaría su implantación a gran escala. No obstante, existen productos comerciales que se apoyan en las tecnologías semánticas y facilitan la gestión de los documentos o la recuperación de éstos a partir de los significados, pero no están demasiado introducidos en las organizaciones por su elevado coste de puesta en marcha y en algunos casos cuestionada eficiencia.

La consultoría de vigilancia tecnológica Gartner muestra en sus informes cual es la implantación de las tecnologías semánticas en las organizaciones. Las aborda con cautela ya que después de años de desarrollo considera que no han logrado implantarse

suficientemente. Sin embargo, reconoce su gran potencial y cita otras tecnologías, como las aplicaciones de código abierto, que tuvieron un largo proceso de adopción pero que actualmente están totalmente integradas. Realiza una previsión de la introducción de las tecnologías de la web semántica y especula que para el 2012 el setenta por ciento de las páginas web tendrán algún tipo de etiquetaje semántico, pero que sólo el veinte por ciento harán un uso intensivo de la web semántica basada en ontologías (Gartner, 07a/b).

Las herramientas desarrolladas se pueden agrupar en dos categorías, las que facilitan el desarrollo de las ontologías y las que ayudan a la anotación semántica de textos, aunque algunas de ellas tienen integradas ambas funcionalidades. Para conocer el estado de la cuestión, de las diferentes herramientas y aproximaciones a ambos enfoques pueden consultarse (Cimiano, 06). Asimismo, (Gómez-Pérez, 05) realiza un estudio de las diferentes herramientas y enfoques para el aprendizaje de ontologías, y (Uren, 06) ofrece un recorrido por las diferentes herramientas existentes para la anotación semántica, tanto manual como automática.

A continuación se realiza una breve selección de algunas herramientas, proyectos y recursos lingüísticos representativos en el desarrollo e implantación de la web semántica.

- GATE (General Architecture for Language Engineering)[2] es una plataforma de código abierto basada en las tecnologías del lenguaje humano que facilita la infraestructura necesaria para el desarrollo de aplicaciones que utilizan de forma intensiva el lenguaje natural. Cuenta con un módulo de extracción de información que permite la creación y gestión de las anotaciones semánticas
- KIM (Knowledge and Information Management system) [3] es una plataforma extensible que se basa en la semántica para gestionar el conocimiento. Permite la anotación semántica, lo cual facilita el acceso a la información ya que se crean enlaces que facilitan la navegación y visualización de los contenidos. También tiene por objetivo mejorar la eficiencia de la indización, recuperación y clasificación de la información sin estructurar o semi-estructurada.
- SEKT (Semantic Knowledge Technologies)[4] es un proyecto que desarrolla y explota las tecnologías semánticas para la gestión del conocimiento. La base del proyecto es la creación de sinergias para combinar tres áreas de investigación relacionadas: la gestión de ontologías, el aprendizaje automático y el procesamiento del lenguaje natural.
- Annotea[5] es un proyecto del W3C que permite al navegante realizar anotaciones (comentarios, preguntas, sugerencias) en páginas web. Permite realizar acotaciones en las páginas web sin que el documento original sufra ninguna transformación. La anotación puede hacerse a nivel de frases o párrafos concretos de la página. Los comentarios se almacenan como metadatos en otro servidor distinto y el documento original permanece inalterado. En este proyecto se ha desarrollado un navegador web propio, Amaya, para poder ver y crear las anotaciones.

[2] http://gate.ac.uk/
[3] http://www.ontotext.com/kim/index.html
[4] http://www.sekt-project.com/
[5] http://www.w3.org/2001/Annotea/

- WordNet y EuroWordNet [6] se trata de léxicos semánticos y son los recursos lingüísticos más utilizados en los que se sustentan muchas aplicaciones. El primero es una base de datos léxica del inglés, que agrupa las palabras en conjuntos de sinónimos llamados *synset*, proporciona definiciones breves y almacena las relaciones semánticas básicas que se establecen entre los elementos. EuroWordNet es una base de datos multilingüe, con WordNets en varios idiomas europeos, donde los WordNets de los diferentes idiomas están conectados gracias al Índice Inter-Lingua que permite consultar palabras similares en los otros idiomas.

Para tener una visión completa de todas las herramientas existentes puede consultarse el blog AI3 de Michael K. Bergman (http://www.mkbergman.com/)[7], donde se ofrece un listado actualizado, Sweet Tools, de más de medio millar de herramientas y recursos que tienen relación con la web semántica.

6.4 Conclusiones

Las tecnologías y la infraestructura desarrollada para la web semántica tendrán repercusión directa en muchas áreas y disciplinas que tienen como eje la comunicación y gestión de información en lenguaje natural.

Así pues, la estructura diseñada para la web semántica permitirá mejorar los diferentes mecanismos existentes para el acceso a la información puesto que la información ya estará estructurada. Existen diferentes campos de investigación relacionados con el acceso a la información en lenguaje natural que podrán evolucionar y ofrecer productos de calidad cuando la infraestructura de la web semántica esté extendida a gran escala. Aquí apuntaremos algunas ideas y haremos una breve síntesis de aquellas áreas que tienen una relación más estrecha con el lenguaje, aunque hay que tener en cuenta que no son una novedad sino que la mayoría llevan años de investigación, pero que con la implantación de la web semántica recibirían un gran impulso.

La disciplina dedicada a la arquitectura de la información podría ofrecer de forma automática distintas visualizaciones para adaptarse a diferentes contextos de acuerdo con el tipo de usuario y sus expectativas, y también del dispositivo utilizado para acceder a la información. La clasificación automática de la información según categorías predefinidas, sustentadas en las ontologías, permitiría acceder a la información de una forma más eficaz y además ayudaría a gestionar las ingentes cantidades de información existentes. La generación de resúmenes para sintetizar el contenido más relevante de un texto sería una forma de ayudar a la difusión de la información.

Además, puede hablarse de nuevos servicios que ya son una realidad y que facilitan enormemente la gestión de la información. Hay que destacar el gran éxito del formato RSS, basado en XML; es utilizado para sindicar contenidos a los suscriptores de un sitio web y se trata de una forma de distribución selectiva de la información. Este sistema ha supuesto una nueva cadena de valor en el sector de los contenidos ya que está cambiando la forma de relacionarse con la información tanto para los profesionales como para los usuarios.

Los sistemas actuales de recuperación de información también sufrirían una revolución, porque la utilización de las ontologías y de un motor de inferencia (Finin, 02) otorgarían a estos un potencial hasta ahora desconocido. (Ding, 05) describe como funcionaría la

[6] http://wordnet.princeton.edu/
[7] consultado el 20-11-2007.

búsqueda semántica, utilizando como ejemplo el buscador Swoogle[8], detalla los cambios que se producirían en la búsqueda de la web semántica respecto a la búsqueda tradicional. La recuperación de información multilingüe, donde la pregunta y/o los documentos están en diferentes idiomas, también estaría más próxima a ser una realidad gracias a las ontologías (Guyot, 06). También podrían sufrir un impulso los sistemas de búsqueda de respuestas (McGuinness, 04), lo que permitiría ir un paso más allá en la búsqueda efectiva de información al responder directamente a las preguntas formuladas por los usuarios. Es decir, en lugar de devolver el documento completo, devolver nada más la zona del texto donde se encuentra la información requerida. Estos sistemas abrirían la puerta a la comunicación en lenguaje natural, que es la forma más ágil y sencilla de expresarse para las personas.

El desarrollo de ontologías, que se ha visto activamente impulsado con la web semántica, tendría también repercusión directa en el ámbito del procesamiento del lenguaje natural, ya que las ontologías como representaciones de un dominio e independientes de la lengua pueden ser el punto de encuentro entre dos o más lenguas. Por tanto, las ontologías consideradas como un repositorio de conceptos que establece conexiones entre los símbolos de una lengua y sus referentes en el dominio que refleja, permiten llevar a cabo un tratamiento adecuado del fenómeno de la sinonimia y también son un soporte básico para la desambiguación, tanto léxica como estructural. Ello podría resultar útil para avanzar en otras disciplinas en las que el lenguaje natural es un elemento clave, como por ejemplo la traducción automática, la recuperación de información, la interacción persona-ordenador, la enseñanza de segundas lenguas, etc.

La generación de lenguaje natural es otra de las áreas que se sufriría un interesante desarrollo con la implantación de la web semántica, pues a partir de los datos estructurados y recogidos en una base de conocimiento, se podrían generar textos de acuerdo con las necesidades especificas del público a quien se dirige o de la tecnología con que va a ser consultada la información (Mellish, 06).

Por último, uno de los grandes logros que también se podría atribuir a la web semántica es el de promover la diversidad cultural en una sociedad camino a la globalización, ya que la representación de diferentes culturas se vería favorecida por las ontologías, puesto que éstas permitirían mostrar como los conceptos son usados en diferentes culturas. Ésta sería una forma de evitar la estandarización de los contenidos, además éstos podrían hacerse accesibles en diferentes lenguas (Bauteliaar, 03a). También las técnicas de la web semántica podrían servir para preservar las lenguas en peligro de extinción, ya que ayudarían a almacenar de forma permanente su documentación y facilitarían su difusión y acceso (Lu, 04).

[8] http://swoogle.umbc.edu/

6.5 Referencias

AGUADO DE CEA, Guadalupe; ÁLVAREZ DE MON, Inmaculada; PAREJA, Antonio (2002): «Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de web semántica: OntoTag», *Revista Iberoamericana de Inteligencia Artificial*, num. 17, 37-49.

ANGELOVA, Galia (2005): «Language Technologies Meet Ontology Acquisition» en *Conceptual Structures: Common Semantics for Sharing Knowledge*, *Lecture Notes in Computer Science*, vol. 3596, 367-380.

ALANI, Harith; et al. (2003): «Automatic Extraction of Knowledge from Web Documents», en *Proceedings of 2nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island, Florida, USA.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora (2001): «The semantic Web», *Scientific America*, num. 501 (May), 29-37.

BONTCHEVA, Kalina (2003a): «Semantic web enabled, open source language technology», en *Proceedings of the 2nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island, Florida, USA.

BONTCHEVA, Kalina; CUNNINGHAM, Hamish (2003b): «The semantic web: a new opportunity and challenge for human language technology», en *Proceedings of the 2nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island, Florida, USA.

BONTCHEVA, Kalina; et al. (2006): «Semantic annotation and human language technology», en Davies John, Rudi Studer, Paul Warren: *Semantic web technologies: trends and research in ontology-based systems*, England: John Wiley & Sons, Ltd., pp [29]-[50].

BUITELAAR, Paul, et al. (2003a): «Towards a language infrastructure for the semantic web», en *Proceedings of the 2nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island, Florida, USA.

BUITELAAR, Paul; DECLERCK, Thierry (2003b): «Linguistic Annotation for the Semantic Web», en Siegfried Handschuh and Steffen Staab: *Annotation for the Semantic Web*, Frontiers in Artificial Intelligence and Applications Series, vol. 96, IOS Press, 2003.

BUITELAAR, Paul; CIMIANO, Philipp; MAGNINI Bernardo (2005): *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications Series, vol. 123, IOS Press.

CALZOLARI, Nicoletta (2003): «Language resources in the semantic Web vision», en *Proceedings of International Conference of Natural Language Processing and Knowledge Engineering*, IEEE, 16- 18.

CIMIANO, Philipp; REYLE, Uwe; ŠARIĆ, Jasmin (2005): «Ontology-driven discourse analysis for information extraction», *Data & Knowledge Engineering*, vol. 55, num. 1 (October), 59-83.

CIMIANO, Philipp (2006): *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*, Germany: Springer, 2006.

CODINA, Luis; ROVIRA, Cristòfol (2006): «Web Semántica: Visión global y Análisis comparativo» en Jesús Tramullas: *Tendencias en documentación digital*, Gijón: Ediciones Trea, pp [9]-[54].

CUNNINGHAM, Hamish; BONTCHEVA, Kalina; LI, Yaoyong (2005): «Knowledge management and human language: crossing the chams», *Journal of Knowledge Management*, vol. 9, num. 5, 108-131.

DECKER, Stefan, et al. (2000): «An Information Food Chain for Advanced Applications on the WWW», en *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, vol. 1923, 490-493.

DING, Li, et al. (2005): «Search on the Semantic Web», *IEEE Computer*, vol. 38, num. 10 (October), 62-69.

DINI, Luca (2003): «NLP technologies and the Semantic Web: Risks, Opportunities and Challenges», en *8th Conference of the AI*IA*.

FININ, Tim; et al. (2002): «Information Retrieval on the Semantic Web», en *Proceedings of the eleventh international conference on Information and Knowledge Management, ACM* , 461-468.

GAMALLO, Pablo; Agustini, ALEXANDRE; PEREIRA, Gabriel (2001): «Selection restrictions acquisition from corpora», *Lecture Notes in Computer Science*, vol. 2258, 30-43.

GARTNER RESEARCH GROUP (2007a): «Taking stands on the Semantic Web», Id Number: G00148696.

GARTNER RESEARCH GROUP (2007b): «Finding and exploting value in Semantic Technologies on the web», Id Number: G00148725.

GÓMEZ-PÉREZ, Asunción; MANZANO-MACHO, David (2005): «An overview of methods and tools for ontology learning from texts», *The knowledge engineering review*, vol. 9, num. 3, 187-212.

GRUBER, Thomas (1993): «A translation approach to portable ontologies», *Knowledge Acquisition*, vol. 5, num. 2, 199-220.

GUYOT, Jacques; RADHOUANI, Saïd; FALQUET, Gilles (2006): «Conceptual Indexing for Multilingual Information Retrieval», en *Accessing Multilingual Information Repositories, Lecture Notes in Computer Science*, vol. 4022, 102-112.

JAVA, Akshay, *et al.* (2007): «Using a Natural Language Understanding System to Generate Semantic Web Content», *International Journal on Semantic Web and Information Systems*, vol. 3, num 4 (November), 50-74.

KATZ, Boris; LIN, Jimmy; QUAN, Dennis (2002): «Natural language annotations for the Semantic Web», en *On the Move to Meaningful Internet Systems, Lecture Notes In Computer Science*, vol. 2519, 1317-1331.

LU, Shiyong; et al. (2004): «Language engineering for the Semantic Web: a digital library for endangered languages», *International Journal of Information Research*, vol. 9, num. 3 (April).

MAEDCHE, A.; STAAB, S. (2001): «Ontology Learning for the Semantic Web», *IEEE Intelligent Systems, Special Issue on the Semantic Web*, vol. 16, num. 2, 72-79.

MCGUINNESS, Deborah (2004): «Question Answering on the Semantic Web», *Intelligent Systems*, vol. 19, num. 1 (Jan-Feb)**,** 82- 85.

MELLISH, Cris; SUN, Xiantang (2006): «The Semantic Web as a *Linguistic* Resource: Opportunities for Natural Language Generation», *Knowledge Based System*, vol. 19, 298-303.

NIRENBURG, Sergei; RASKIN, Victor (2001): «Ontological semantics, formal ontology, and ambiguity», en *Proceedings of the international Conference on Formal ontology in information Systems, ACM*, vol. 2001, 151-161.

OMELAYENKO, Borys (2001): «Learning of ontologies for the Web: the analysis of existent approaches», en *Proceedings of the International Workshop on Web Dynamics*.

UREN, Victoria, et al. (2006): «Semantic annotation for knowledge management: Requirements and a survey of the state of the art», *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, num. 1 (January 2006), 14-28.

SCHWITTER, Rolf (2005): «A controlled natural language layer for the semantic web», en *The 18th Australian Joint Conference on Artificial Intelligence*, *Lecture Notes in Artificial Intelligence*, vol. 3809, 425-434.