

Gestión de la Información Multimedia en Internet
Gestión del conocimiento
DAML y ontologías consensuadas

Autor: Pablo Barrera González
Profesor: Carlos Delgado Kloos

Fecha de presentación: 7 de Febrero de 2003

Índice

1. Introducción	2
2. La Web semántica	2
3. Ontologías	2
3.1. Agrupación de ontologías	3
3.2. Búsqueda semántica	5
4. DAML	6
4.1. Orígenes de DAML	6
4.2. Comparativa	6
4.3. Estado de DAML	8
4.4. Utilización de DAML	8
5. Conclusión	9

1. Introducción

Conseguir que la interacción de los hombres y de las máquinas sea más profunda en el ámbito de la web es el objetivo que persigue la web semántica. En ella, la información será accesible tanto para los hombres como para las máquinas, pudiendo ambos comprenderla. Para construir estos sistemas, se están empleando descripciones de cada uno de los campos de conocimiento, denominadas *ontologías*. Por otra parte, distintas iniciativas se encargan de desarrollar lenguajes para incluir metadatos en los recursos web, con información acerca de su contenido, empleando una determinada ontología.

El presente trabajo consta de dos partes. En la primera se revisará el artículo *Consensus Ontologies: Reconciling the Semantics of Web Pages and Agents* de Larry M. Stephens y Michael N. Huhns, en el que se describe un método para consensuar ontologías de diferentes fuentes. En la segunda parte se analizará el estado del proyecto *DAML*, un lenguaje de descripción de contenidos basado en RDF.

2. La Web semántica

El término *web semántica*, acuñado por Tim Berners-Lee, se refiere a una posible futura web, vista como una gran web de datos, similar a lo que podría ser una gran base de datos global. El objetivo de esta web es proveer de una infraestructura a la actual, para que permita realizar tareas de procesado del conocimiento contenido en la misma, hasta ahora impensables con los medios actuales. Tanto los hombres como las *máquinas* serán capaces de *consultar* y *organizar* la información e incluso *realizar deducciones* a partir de la misma.

Las páginas web actuales están diseñadas por personas, para ser interpretadas por personas. No existe un formato común de representación para mostrar la información, sino que en cada caso se emplea uno diferente. Esto presenta un problema evidente a la hora de automatizar tareas tales como la extracción de la información de los contenidos web, objetivo perseguido por la web semántica. Con ésta, se intenta obtener una representación abstracta de los datos contenidos en la tradicional *WWW (World Wide Web)*, de forma que la información almacenada en ella pueda ser usada y "*comprendida*" por una máquina sin necesidad de supervisión humana. La definición de los datos será tal que permita descubrir nueva información en otras partes de la web semántica, automatizar procesos, integrar aplicaciones con la información y reutilizar recursos de manera mucho más sencilla de la que nos proporcionan los métodos con los que contamos hoy en día. Con todo esto se busca convertir la información almacenada en la web en *conocimiento* utilizable por los ordenadores, y así mejorar nuestra utilización del mismo.

3. Ontologías

Para conseguir alcanzar la web semántica es necesario que los sitios web describan de manera más detallada cuales son sus contenidos, y que esta descripción sea completamente comprensible por los ordenadores, esté consensuada y sea reutilizable. De esta forma las máquinas alcanzarán el punto de comprensión buscado.

Todas las aproximaciones para conseguirlo pasan por la utilización de *ontologías* para describir tanto las fuentes de información como las relaciones existentes entre ellas. El término ontología proviene de la filosofía, estando relacionado con la teoría del ser. En el contexto de la web semántica, se refiere a una serie de enunciados que definen las relaciones entre conceptos y que proporcionan reglas lógicas para razonar con ellos. Con

estas herramientas los ordenadores son capaces de “comprender” el significado de los datos semánticos de un sitio web, siguiendo las relaciones con una serie de ontologías especificadas.

Una definición más exacta del significado de la ontología podría describirla como “*una especificación explícita y formal sobre una conceptualización compartida*”. Esta definición nos indica es que las ontologías definen conceptos y relaciones en algún dominio de conocimiento, de forma compartida y consensuada; y que esta conceptualización debe ser representada de una manera formal, legible y utilizable por las máquinas.

Para representar estas relaciones se han propuesto una serie de lenguajes de descripción, cuyo principal componente es *RDF (Resource Description Framework)*, un estándar del *W3C (WWW Consortium)*. El *RDF* proporciona un marco general para describir conceptos y las relaciones existentes entre los mismos. El principal problema que presenta, como veremos más adelante, es su excesiva generalidad que puede llevar a inconsistencias a la hora de definir ontologías. Por este motivo han surgido otra serie de iniciativas, definiendo lenguajes de marcado específicos para ontologías, basadas en *RDF*: *DAML (Darpa Agent Markage Language)*, *OIL (Ontology Inference Layer)*, *DAML+OIL* y *OWL (Ontology Web Language)*.

3.1. Agrupación de ontologías

Para conseguir que la web semántica funcione es necesario que todo su contenido esté convenientemente marcado, representando el conocimiento de la web mediante el uso de ontologías, como ya se ha dicho. Para que esto sea posible es necesario clarificar dos puntos:

- En primer lugar es necesario proveerse de una serie de métodos (lenguajes) que nos permitan describir estas ontologías de manera comprensible por las máquinas y, a ser posible, estándar.
- En segundo lugar, y más importante si cabe, es necesario que estos métodos puedan combinarse aún teniendo orígenes diferentes. La interoperabilidad es un pilar fundamental para que la web semántica comience a funcionar; de hecho es una de los requisitos que pone el *W3C* para cualquier lenguaje sobre ontologías. En caso que no fuésemos capaces de relacionar términos, por no usar la misma ontología para describirlos, aún perteneciendo al mismo dominio de conocimiento, ninguna de las bondades de las que hemos hablado hasta ahora serían alcanzables. Por tanto parece deseable la existencia de alguna forma de conseguir que documentos de diferentes fuentes puedan relacionarse mediante una ontología común.

El primero de estos puntos está actualmente solucionado empleando lenguajes como *RDF* y sus derivados. El *RDF* provee un marco común para descubrir los contenidos, pero dada su amplia capacidad puede presentar ambigüedades, por lo que es necesario encontrar otras vías para representar este tipo de información de forma concisa. *DAML*, y el resto de los lenguajes comentados anteriormente, solucionan estos problemas. Con ellos, el primero de los puntos se encuentra en vías de resolución.

En cuanto al segundo punto, que todavía no se encuentra resuelto, es necesario clarificar como se puede realizar el consenso entre ontologías de diferentes fuentes. A la hora de introducir información acerca del contenido de un página web los desarrolladores pueden optar entre tres posibles métodos:

- Usar la misma terminología en todas las páginas web, con una semántica prefijada y consensuada. De esta forma no existirá ambigüedad entre las ontologías de documentos pertenecientes a fuentes distintas.

- Emplear una terminología diferente en cada página, pero incluir un método de traducción hacia una ontología global prefijada.
- Cada sitio web emplea una pequeña ontología propia, que pueden relacionarse con las demás indirectamente mediante la asistencia de agentes.

Parece, a día de hoy, improbable que los dos primeros métodos lleguen a ser empleados debido, sobre todo, a la complejidad de implantación que conllevan. En primer lugar se presupone la existencia de consenso en una determinada ontología global para marcar las páginas pertenecientes a cada uno de los dominios de conocimiento, lo que parece difícil de alcanzar en todos ellos. Además se presupone que ese modelo será aceptado, y utilizado, por todos.

Ahora bien, el tercer método parece más fácil de implantar, puesto que su simplicidad es una importante baza a su favor. Lo que hace falta aclarar es si, a pesar de su sencillez, es capaz de proporcionar una herramienta útil en el camino hacia la web semántica.

El marco de trabajo que este sistema propone parte de un conjunto de páginas web, en las que se definen pequeñas ontologías para una parte específica de un dominio de conocimiento. La semántica empleada para ello es también local y no depende de ningún acuerdo específico entre diferentes sitios web. Se parte de la hipótesis que estos fragmentos de ontologías, representadas con semánticas diferentes, pueden relacionarse de forma automática sin la necesidad de definir una ontología global. La forma de hacerlo consiste en relacionar conceptos de diferentes fuentes y, de esta forma, ir construyendo un árbol de conceptos mayor agregando estas piezas. La nueva ontología que emerge de esta unión establece una relación entre todos los conceptos que la forman, independientemente de cual fuera su origen.

Incluso cuando no existe una forma directa de determinar la relación entre un par de ontologías, la existencia de un conjunto suficientemente grande de ellas puede ayudar a aclarar estas relaciones o, incluso, construirlas. Puede, de esta manera, formarse un puente semántico entre conceptos de diferentes fuentes.

Una vez agrupadas todas las fuentes, la ontología emergente proporciona una caracterización para el conjunto de todas las web originales y el dominio involucrado. Además se crea una única gran ontología que sirve como nexo para las interacciones. Esta metodología establece un medio por el que agentes y otros componentes de sistemas de información interoperen.

A la hora de establecer conexiones, existen siete tipos de relaciones básicas entre dos conceptos cualesquiera. Estas son las siguientes:

Sub-Clase En este caso un concepto presenta particularidades específicas sobre uno más general.

Super-Clase Caso contrario a la Sub-Clase.

Equivalencia Cuando dos conceptos son iguales o equivalentes.

Parte-De A diferencia de una Sub-Clase, un concepto forma parte de otro cuando es una pieza para crearlo y no sólo un tipo más específico del concepto concreto.

Tiene-Parte De nuevo, se trata del contrario de Parte-De.

Hermano En este caso, dos conceptos son hermanos cuando cuelgan del mismo nodo en un árbol de ontología.

Otra En esta última clase se unen el resto de posibles relaciones no contempladas en el resto.

En la figura 1, a modo de ejemplo, puede verse como funcionaría el sistema en una situación muy sencilla. En primer lugar hemos extraído la información de dos sitios web en los que se habla de las partes de un automóvil. En el primero se indica que una “Rueda” es parte de un “Turismo” y en el otro que una “Llanta” es parte de un “Deportivo”. En principio el sistema no es capaz de relacionar los conceptos “Turismo” y “Deportivo” entre sí. Para ello es necesario encontrar una fuente, o varias, que relacione algunos de los conceptos de las anteriores de tal forma que seamos capaces de encontrar un nexo entre estos términos. De esta forma, si encontramos la información necesaria de que una “Llanta” forma parte de un “Turismo”, podemos establecer los enlaces necesarios entre las fuentes que teníamos con anterioridad y concluir que existe una posible equivalencia entre los términos “Turismo” y “Deportivo”.

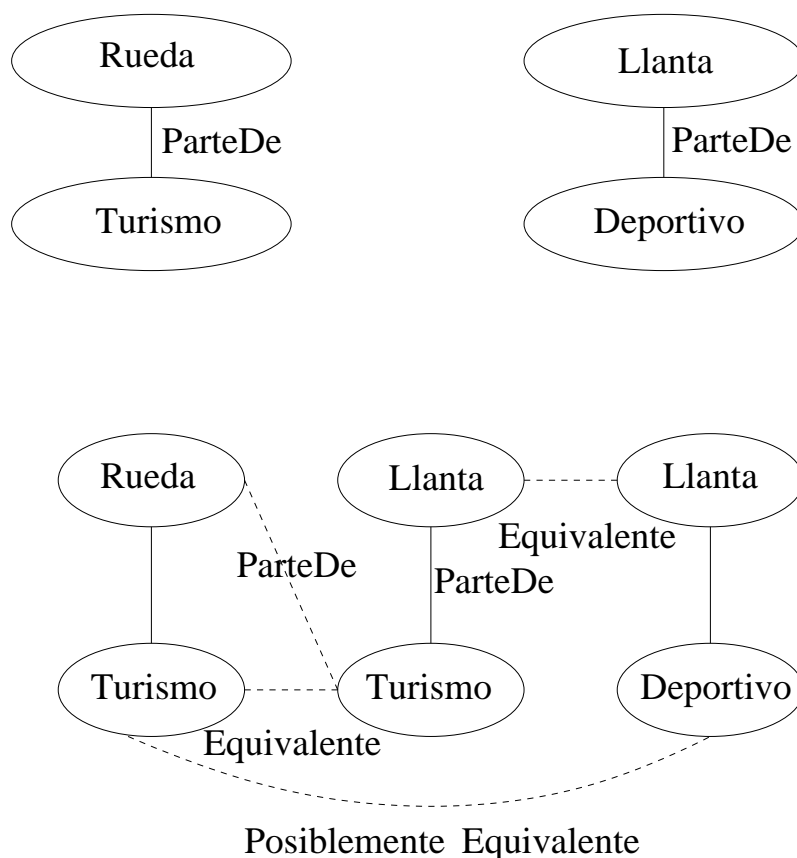


Figura 1: Relación entre ontologías

3.2. Búsqueda semántica

Una de las ventajas más evidente que se está buscando con la web semántica es mejorar la capacidad de búsqueda en la información contenida en ella. Sabiendo que significado tienen los datos, los sistemas de agentes serán capaces de proporcionarnos respuestas más adecuadas y acertadas ante nuestras preguntas.

El uso de ontologías consensuadas también puede ayudarnos en esta labor. Al juntar en una misma ontología descripciones de diferentes fuentes, pueden establecerse distancias semánticas entre los conceptos. De esta forma si para llegar desde un término a otro, un nodo en el árbol que construimos, deben atravesarse muchos otros términos podemos concluir que existe una distancia semántica superior que en un caso en el que el número de saltos sea inferior. De esta manera puede realizarse una ordenación de los resultados

de una búsqueda tradicional atendiendo al significado de los datos contenidos en un sitio web.

Otra ventaja que podemos obtener al emplear estos sistemas es que la información estará colocada atendiendo a la ontología local usada. Esto puede servir para solventar problemas como la colocación de la información. Existe una pregunta básica que debemos hacernos a la hora de colocar la información: ¿dónde poner cada cosa? ¿Es mejor colocarla en el sitio correcto, dónde le corresponde, o sería preferible que estuviera en el lugar dónde más probablemente van a buscarla? En el caso de emplear ontologías consensuadas los agentes de búsqueda tendrán en cuenta todas las formas de relacionar los conceptos que encuentren en las diferentes ontologías con las que trabajen. De esta forma, si, equivocadamente, existen sitios web que relacionan términos mal (por ejemplo, considerar a un cocodrilo como un mamífero, error extendido entre un porcentaje alto de la población) el consenso hará que se encuentre la información de forma correcta.

4. DAML

Como ya se ha comentado en el camino hacia la web semántica, hacia conseguir una red de datos que maneje de forma más inteligente la información, necesitamos incluir, de alguna forma, una representación de las relaciones entre las entidades que forman cada recurso de la web. *DAML* proporciona un lenguaje de marcado para representar estas relaciones dentro de los sitios web, empleando ontologías para indicarlas.

4.1. Orígenes de DAML

HTML plano proporciona, simplemente, una manera de representar información textual, pero no incluye ningún mecanismo por el cual se pueda describir de alguna forma que significado tiene dicha información. Este conocimiento puede ser muy beneficioso para facilitar el procesado como las búsquedas sobre estas fuentes de datos.

XML proporciona un sistema de metamarcado, con el cual se introducen datos sobre datos, pero que tiene una capacidad limitada para representar las relaciones entre elementos (esquemas u ontologías). La utilización de ontologías representa una herramienta muy poderosa para describir objetos y sus relaciones con otros.

RDF sí que es útil a la hora de describir conceptos y cuales son las relaciones que les unen, pero dada su generalidad puede ser muy ambiguo en muchos casos, a parte de no estar dotado con las suficientes herramientas para trabajar con ontologías de manera eficiente.

Para solucionar las limitaciones que presenta RDF en este contexto, surgen varios lenguajes específicos para ontologías. El proyecto *DAML*, o *DARPA Agent Markup Language*, tiene su origen en agosto de 2000, y su objetivo principal consiste en desarrollar un lenguaje para facilitar el concepto de la web semántica, así como un conjunto de herramientas útiles para tal propósito. *DAML* está escrito en RDF, en concreto es un tipo específico de marcado RDF, lo que facilita su estandarización y aplicación.

4.2. Comparativa

En el cuadro 1 se resumen cuales son las características de una serie de lenguajes con los que podemos comparar *DAML*. Estas son las siguientes:

Listas acotadas *DAML* emplea una estructura para representar listas, no ordenadas, de un tamaño máximo o determinado.

	DTD	Esquemas	RDF(S)	DAML+OIL	RDF(S) 2002	OWL
Listas acotadas				X	X	X
Cardinalidad restringida	X	X		X		X
Expresiones para clases				X		X
Tipos de datos		X		X		X
Definición sobre clases				X		X
Enumeración	X	X		X		X
Equivalencia				X		X
Extensibilidad			X	X	X	X
Semántica formal				X	X	X
Herencia			X	X	X	X
Inferencia				X		X
Restricciones locales				X		X
Restricciones de límite				X		
Reutilización			X	X	X	X

Cuadro 1: Comparativa de características

Cardinalidad restringida DAML es capaz de limitar el número de sentencias con el mismo tema y predicado. Los operadores “?”, “*” y “+” de los DTD proporcionan una funcionalidad básica de este tipo.

Expresiones para clases Siempre que se define una clase, DAML+OIL permite expresiones que combinen expresiones del tipo “uno de”, “disconjunto”, “intersección” o “complemento de”.

Tipos de datos En RDF los literales son esencialmente cadenas de caracteres. En DAML+OIL se añaden los tipos de datos de los Esquemas de XML.

Definición sobre clases DAML permite definir nuevas clases en función de restricciones de clases existentes (por ejemplo, “Niño” es una “Persona” con edad menor de 18 años).

Enumeración Los DTD permiten especificar un conjunto de valores restringidos para un atributo dado. DAML proporciona la expresión “uno de”.

Equivalencia Para permitir razonar entre ontologías y bases de conocimiento, DAML permite la relación “equivalente a” para clases, propiedades e instancias.

Extensibilidad RDF y DAML permiten que se añadan nuevas propiedades a las clases existentes. De esta forma DAML+OIL se ha definido a partir de RDF.

Semántica formal La semántica de DAML+OIL está expresada en forma de un modelo teórico.

Herencia Los grupos de esquemas de XML formalizan el uso de entidades en las definiciones de atributos, pero eso no es completamente herencia. RDF y DAML permiten el uso de “sub-clase de” y “sub-propiedad de”.

Inferencia DAML+OIL construye propiedades como la “transitiva”, la “no ambigüedad”, la “inversa de” y “disyunción con” que proporcionan información adicional para razonar. Se espera que futuras versiones de DAML proporcionen reglas, métodos de comprobación de las mismas, etc.

Restricciones locales DAML permite que las restricciones puedan estar asociadas con el par clase/propiedad, por ejemplo, que la propiedad del color pueda emplearse para los coches y para los ojos con dominios diferentes.

Restricciones de límite Las restricciones de DAML permiten expresiones del tipo “todos los hijos de X son de tipo Persona”. Las propiedades de “tiene clase”, “cardinalidad”, “mínima cardinalidad”, “máxima cardinalidad” proporcionan limitaciones del tipo “por lo menos tres de los hijos de X son de tipo Doctor”.

Reutilización RDF y DAML posibilitan que las declaraciones sean utilizadas como sujetos de otras sentencias. La ... proporciona un mecanismo estándar para almacenar fuentes de datos, marcas de tiempo, etc, sin entrometerse en el modelo de datos.

4.3. Estado de DAML

Actualmente el proyecto DAML cuenta con DAML+OIL como última versión lenguaje desarrollado para el marcado de páginas, revisada por última vez en mayo de 2001. DAML+OIL proporciona un rico conjunto de constructores con los que crear las ontologías y para marcar la información. Este lenguaje intenta consensuar los lenguajes DAML y OIL (*Ontology Inference Layer*), buscando el punto de partida para la construcción de un único lenguaje para el W3C.

A partir de esta iniciativa apareció el primer borrador de OWL (*Web Ontology Language*), publicado por el W3C y que está basado en DAML+OIL. Desde entonces el desarrollo de DAML parece parado (recordemos que la última versión de DAML+OIL es de hace casi dos años), puesto que los esfuerzos se están centrado en OWL. Aún así, otras partes del proyecto DAML continúan en desarrollo. Entre las más importantes en las que se sigue trabajando podemos citar las siguientes:

- Una biblioteca de ontologías.
- DAML-S, una ontología para servicios basados en web.
- DAML-Time, una ontología para conceptos temporales.
- Herramientas para recolección de ontologías.
- Mantenimiento de manuales, cursos y listas de correo.

4.4. Utilización de DAML

Cuando le dices algo a una persona, él puede combinar esta nueva información con otros datos que conociera anteriormente. A partir de esta combinación pueden realizarse deducciones, de gran utilidad para el manejo de toda la información que vamos adquiriendo.

Cuando le dices algo a un ordenador en XML, él puede ser capaz de darte una respuesta nueva combinando esa información con otra, pero sólo basándose en algún tipo de software que no forma parte de la especificación XML. El problema que esto plantea es que la respuesta variará de una implementación a otra de estos programas, lo que hace difícil emplear estos sistemas de manera general.

Cuando le dices algo a un ordenador en DAML, él es capaz de devolver nueva información basándose únicamente en el estándar de DAML. Es capaz de realizar conclusiones sencillas basándose únicamente en las herramientas que DAML proporciona.

Un determinado conjunto de conclusiones es necesario para cualquier sistema conforme con DAML. Los sistemas deben ser capaces de proporcionar todo tipo de servicios adicionales y respuestas más allá de los requisitos del estándar, pero un determinado

conjunto básico de conclusiones siempre serán necesarias. DAML proporciona a los ordenadores un pequeño grado extra de autonomía que pueden ayudarles a realizar unas actividades más útiles para las personas.

Tener conocimientos que puede aplicarse de manera dinámica para encontrar una respuesta, en vez de emplear procedimientos predefinidos, es extremadamente potente. DAML proporciona una infraestructura básica que permite a las máquinas hacer la misma clase de inferencias simples que a los seres humanos. Es únicamente un comienzo, pero su desarrollo es un pilar crítico para la web semántica.

5. Conclusión

La web semántica está cada vez más cerca. Herramientas como DAML, OWL y RDF sentarán las bases para esta futura red de datos. Los sistemas de búsqueda que se crearán a partir de esta tecnología serán más inteligentes que los actuales, algo necesario ya a día de hoy, para lidiar con la inmensa cantidad de datos que hoy en día inundan Internet.

En cuanto a los métodos de consenso de ontología presentado en el artículo aquí comentado, aunque su implantación parece prometedora, el texto no termina de ser lo suficientemente riguroso en cuanto a los resultados reales de la combinación de ontologías. Las únicas conclusiones que presentan están obtenidas a partir de experimentos muy acotados y específicos, lo que hace que podamos plantearnos la gran dificultad de su implementación más allá de su utilización de entornos simulados. Por esta razón, debemos mostrarnos escépticos en cuanto a la aplicación de este método en particular. Aún así es conveniente resaltar la bondad, e incluso la necesidad, de que exista un consenso entre las diferentes ontologías empleadas. No utilizar una ontología global se vislumbra como una ventaja clara, aún así puede que el consenso directo no sea tan fácil como muestran las primeras pruebas en entornos simulados.

Por último, DAML, como lenguaje de marcado, parece mostrar una serie de ventajas claras frente a RDF. Aún así, a día de hoy, parece un proyecto abandonado en favor de OWL. Como OWL hereda muchas de las características de DAML, las ventajas que hemos comentado para este último pueden aplicarse al anterior. Lo que resulta evidente es la necesidad de estandarizar un lenguaje de marcado para ontologías basado en RDF, como base imprescindible para incluir información necesaria para construir la abstracción de datos para que las máquinas interpreten los datos de la web semántica.