

*Cátedra Taller en Tecnologías en Comunicación Social
La Plata, Argentina*

Documento de la Cátedra. 2006

Forma y contenido en la web: De los lenguajes de marcado a la web semántica

Mela Bosch

Documento producido bajo licencia



Palabras Clave: TECNOLOGÍAS EN COMUNICACIÓN SOCIAL / WEB SEMANTICA / LENGUAJES DE MARCADO /

<p>Indice de temas</p>	<p>HTML: La forma de la web Para dar forma y realizar enlaces en todo lo que se publica se utiliza el Hipertext Markup Language. Que es un lenguaje de marcado. Los lenguajes de marcado están constituidos por etiquetas, marcas, que son indicaciones descriptivas sobre el aspecto físico, fechas, y otros datos de un documento, imagen, sonido, etc y son legibles por los programas informáticos. El lenguaje base es el SGML (Standard Generalized Markup Language) y de este surgieron subconjuntos como HTML (Hypertext Markup Language) y extensiones, es decir nuevas generaciones más amplias como XML que se ha transformado en la base de todos los desarrollos en Web.</p> <p>El acceso al contenido en la web Con el uso de los lenguajes de marcado las computadoras pueden realizar las tareas de rutina dentro de las páginas web pero no tienen ninguna forma confiable para procesar la semántica, es decir para describir el contenido de ellas. Los motores de búsqueda (Google, Lycos, etc) se valen de medio estadísticos a partir de las palabras que aparecen en las páginas web. Pero no siempre las palabras expresan el contenido, la semántica es algo mucho más complejo, ya que la referencia puede hacerse con otros términos, en otros idiomas, o puede ocurrir que una persona busque algo utilizando otros términos que no son los más usuales. Para facilitarlos se agregan a las páginas web etiquetas en que los generadores que describen el contenido, estas etiquetas que describen los datos se llaman metadatos. Aparte de los motores gratuitos de internet existe una gran cantidad</p>
<p><u>HTML: la forma de la web</u></p>	
<p><u>El acceso al contenido en la web</u></p>	
<p><u>Qué es la web semántica?</u></p>	
<p><u>Las tecnologías de la web semántica</u></p>	
<p><u>Para profundizar</u></p>	
<p><u>Fuentes de información utilizadas</u></p>	

en este documento

de sistemas que realiza la indización en forma automática de los contenidos. Pero estos programas son costosos y trabajan a partir de lo que ya está.

La web semántica

Hay también otra corriente: es que el contenido sea descripto por los generadores de las páginas web, o por servicios que se ocuparían de hacerlo, indicado con formas de lenguaje de marcado especial que pueda ser manejado tanto por máquinas como por seres humanos: es lo que se llama Web Semántica.

En la web semántica las marcas en las páginas web no describen únicamente como es el documento sino qué contiene, adónde está, quienes son los autores y las organizaciones etc.

La Web Semántica fue lanzada por el mismo fundador del web. Berners-Lee en el año 2000. (Berners-Lee, T.; Hendler, J.; Lassila, 2001). Según estos autores la Web Semántica brindará contenido significativo a las páginas de la red. Creará un ambiente donde agentes de software, moviéndose de página a página, puedan fácilmente efectuar tareas sofisticadas para usuarios. Según ellos la Web Semántica no será una red separada sino una extensión de la actual, en la que la información tenga un significado definido, permitiendo que las computadoras y la gente puedan trabajar en cooperación.

Hasta ahora, así como existen programas que indizan en forma automática el contenido de los documentos, existen también sistemas para manejar conocimiento, es la revolución de lo que se llama Knowledge Management y se trata de software sofisticados y costosos. Requieren que todos compartan exactamente la misma definición de conceptos. Por otra parte estos sistemas limitan el tipo de preguntas que pueden pedirse para que la computadora conteste en forma confiable. Para evitar tales problemas, los sistemas de representación del conocimiento han tenido un propio y limitado conjunto de reglas para hacer inferencias sobre sus datos.

En cambio, el desafío de Web Semántica está en proveer un lenguaje que exprese tanto los datos como una lógica para darles sentido de diferente origen y forma.

Las tecnologías de la Web Semántica

Las tres tecnologías Web Semántica son el XML y el Resource Description Framework (RDF) y las ontologías.

XML, una extensión de lenguaje de marcado tiene un doble carácter ya que permite crear etiquetas propias y ocultas, pero además es puede dar la especificación de puntos de ejecución de porciones de software tales como scripts y hasta programas completos que pueden hacer uso de etiquetas en maneras sofisticadas.

En cuanto al marco de descripción de recursos o Resource Description

Framework (RDF) es el complemento de XML, ya que éste si bien permite agregar marcas arbitrarias a los documentos no dice nada sobre qué significan. El significado es expresado por RDF, se trata de conjuntos de metadatos organizados en tripletes, cada triplete está constituido como si fuera el sujeto, verbo y predicado de una frase muy elemental. Los tripletes de RDF forman redes de información entre cosas conexas

RDF está siendo desarrollado y promovido por el Consorcio 3WC y se están desarrollando recomendaciones que se esperan serán normas ISO en breve según el RDF Interest Group.

La forma en que funciona es más o menos así: cada sujeto y cada predicado son identificados por el Universal Resource Identifier (URI), tal como se hace con un enlace en cualquier página web (URLs, Uniform Resource Locators, son el tipo más común de URI.) Los verbos son identificados también por URIs, que permite a cualquiera definir un concepto nuevo, un verbo nuevo, simplemente definiendo un URI para ellos en algún lugar de la Web.

Ya que los RDF usan URIs para codificar información en un documento, los URIs aseguran que los conceptos no son simplemente palabras en un documento sino que las vinculan a una definición única que todos pueden encontrar en Web.

Pero esto no evita la superposición ya que dos recursos, por ejemplo dos bases de datos en línea, pueden usar diferentes identificadores para lo que de hecho lo mismo. Un programa que quiere comparar o combinar información a través de las dos bases de datos tiene que saber que uno o varios términos están siendo usados para significar la misma cosa. El programa debe tener una manera para descubrir tales significados comunes.

Una solución a este problema es provisto por el tercer componente básico de la la Web Semántica, las ontologías.

Este término tiene su origen en la filosofía y para los sistemas informáticos de última generación es la especificación de una conceptualización. Cada uno de los conceptos son definidos en una red terminológica que explicita sus atributos y comportamiento, pero además tienen una forma de establecimiento los comportamientos por medio de reglas, que permiten que la ontología deduzca, o por lo menos proponga, a qué clase o categoría puede pertenecer cada nuevo concepto que ingresa.

En síntesis la ontología está formada por una taxonomía, es decir de un organización jerárquica de los objetos de un sistema y de un conjunto de reglas de que dicen qué cosas pueden hacer esos objetos. Las ontologías pueden facilitar el funcionamiento de la Web para mejorar la exactitud de la recuperación, el programa realiza la búsqueda sólo en las páginas que se refieren al concepto preciso en vez de todas las palabras claves ambiguas. Las aplicaciones más avanzadas usarán las ontologías para relacionar la información de

	<p>una página con el conocimiento asociado a ellas en otras páginas aunque no esté indicado en forma explícita.</p> <p>Además, este marcado hace mucho más fácil de desarrollar programas que puedan abordar preguntas complicadas cuya respuesta no radica en una única página.</p>
--	--