Bloque 1. Introducción:Información Geográfica, Sistemas de Información Geográfica y Web Semántica

Capítulo C. Las Infraestructuras de Datos Espaciales

Unidad 1.C.7: Introducción a la Web Semántica

Dr. Miguel Angel Bernabé Poveda *(Universidad Politécnica de Madrid)*

7.1. Introducción

La web tradcional (si se puede hablar de tradición en algo que para la mayoría de los usuarios tiene menos de 15 años de antigüedad) está pensada para que los humanos se comuniquen. Por ejemplo, generalmente, los buscadores en Internet realizan búsquedas de cadenas de caracteres coincidentes con los términos solicitados por el usuario.

Si en Internet, solicitamos información sobre el término "*casa*", un buscador normal recorrerá la web en busca de esa secuencia de letras colocadas en ese mismo orden. El usuario recibirá una miríada de direcciones URL donde la palabra *casa* está presente. Encontraremos que alguna de las referencias se refieren a la palabra "*casa*" como sinónimo del concepto "*edificio*" ("*Está cerca de aquella casa*") , en otros casos como sinónimo de "*hogar*" ("*Nos fuimos a casa y nos quedamos dormidos*"), en otros se referirá al tiempo verbal correspondiente del verbo "*casar*" (*"Si ella se casa con otro, yo me suicido*")

Es más, si introducimos la secuencia *"El martes por la mañana iré a buscartea"* (sin las coomillas) los buscadores me darán información no sólo sobre la palabra "*mañana*" sino todas las palabras restantes de la oración. El buscador, (que es muy eficiente), no entiende el significado de la frase al no entender el significado de las palabras. Sin embargo, si a un humano se le formula la frase anterior, no cabe duda de que no confundirá la palabra "*mañana*" con "*el día después de hoy*" pues comprenderá su significado por el contexto de la frase en la que se encuentra.

Los buscadores, generalmente, buscan solo secuencias de letras (o números) pero no entienden de significados. Esto se debe a que la web tradicional está diseñada para ser leída por humanos y no por máquinas. El Proyecto de la Web Semántica tiene como fin crear páginas web comprensibles para ordenadores, de manera que puedan buscar sitios web y realizar tareas encadenadas de una manera estandarizada. Serán tareas en las que hoy es imprescindible la implicación de un humano en su resolución.

.

7.2. La Web Semántica

La Web Semántica es un proyecto que pretende crear un medio universal para el intercambio de información mediante la colocación de documentos en la red de forma que puedan ser procesados por máquinas, al margen del idioma en el que estén redactados. El usuario podrá encontrar respuesta a sus preguntas gracias a que la información estará bién definida.

Para esa posibilidad, la web debe disponer de los significados de los conceptos de manera que las palabras no sean solo una colección de letras dispuestas en un cierto órden, sino que sean conceptos bien definidos. Una vez que el concepto de *casa* o los diferentes conceptos de *mañana* estén bien definidos, el ordenador sabrá a qué nos estamos refiriendo en cada caso.

Por ejemplo: Se podrá encargar al ordenador tareas del tipo: *"Quiero reservar a mi nombre un ticket de avión para ir a Londres a mediados de mayo, lo más barato posible"*. El ordenador debe entender los conceptos: *"mi nombre"* (que estará en algún archivo de mi computadora); "*ticket de avión para ir a*" (o tambien "*billete de avión*" o "*pasaje aéreo*"); *"mediados de mayo"* (y saber que puede ser cualquier día entre el 10 y el 20 de mayo) y "*lo más barato*" (teniendo cuidado de no comprar el más barato que encuentre ese mismo día, sino conocer el rango en el que se mueven los precios y comparalo o no comprarlo hoy, esperando a ver lo que ocurre mañana, tal y como haría el humano). Es más, debe entender esos conceptos, unirlos y debe saber el significado completo de la frase, o lo que es lo mismo,

como resultado a esa orden el ordenador debe reservar a mi nombre un billete barato de avión a mediados de mayo para Londres.

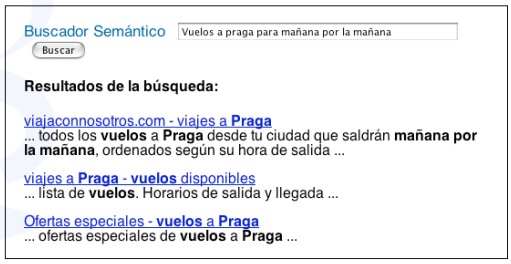

Figura 1
Resultado de búsqueda con un buscador no semántico a la cuestion: "Vuelos a Praga para mañana por la mañana"

Figura 2
Resultado de búsqueda con un buscador semántico

http://www.w3c.es/Divulgacion/Guiasbreves/WebSemantica (activa el 11/06/2006)

Adicionalmente, si tras cada palabra escrita debe estar su concepto, y existe la forma de identificar el significado de cada palabra aunque existan homografías (*haya* = árbol y *haya* = forma del verbo haber), no debe haber problema para que las búsquedas se realicen sobre cualquier idioma, al margen del que se le realice la pregunta. Eso proporcionaría una riqueza enorme a los resultados de las búsquedas.
Por ejemplo, en la pregunta *"Quiero un billete de vión para mañana por la mañana"* el sistema semántico debe saber que *"para mañana"* significa *"para el día después de hoy"* y que *"por la mañana"* significa *"antes del mediodía"*. Debe tambien saber que los conceptos asociados a *"mañana por la mañana"* equivalen a los asociados a "*tomorrow morning*", "*demain le matin*", "*morgen früh*", "*domani mattina*", "*amanhã manhã*", etc. Y debe saber que preferimos que los resultados nos los muestre en español.

En resumen, para disponer de una auténtica Web Semántica se debe disponer de datos bien documentados para que las máquinas resuelvan problemas bien definidos que solucionen preguntas bien formuladas. Los datos deben estar rigurosamente documentados y dotados no sólo de significados sino de relaciones y de reglas de identificación para casos de homografías.

7.2.1 Datos bien documentados

Como se verá más adelante en este Curso, los metadatos son informaciones que describen a los datos. Se dice que "los metadatos son datos sobre los datos". Ejemplo: Mi pasaporte contiene los metadatos que la policía de inmigración requiere acerca de mi persona. Mi historial médico contiene los metadatos que mi médico necesita también acerca de mi persona. Algunos metadatos de los que quiere el policía y de los que quiere el médico coinciden: Nombre, apellidos, fecha de nacimiento, pero otros como si pasé o no pasé la tos ferina o el sarampión cuando era niño, son irrelevantes para el policía.
Otro ejemplo: Los descriptores que vienen en los envases de comidas en los que se especifica las cantidades de elementos que la componen (agua, edulcorantes, colorantes, etc) son los metadatos de esa comida.

Los metadatos de una página web son aquellos datos que informan a los buscadores sobre los contenidos que van a encontrar en cada página web. Por ejemplo, los metadatos de esta página convertida a página web pueden ser <web semántica>, <Información geográfica> <Infraestructura de datos espaciales> <IDE> pero nunca serán <pasajes de avión> <Londres> aunque en ella aparezca estos términos que hemos utilizado como ejemplo.

7.3. La Web Semántica y la Información Geográfica (IG)

Las IDEs pretenden, como objetivo final, ayudar al asuario a localizar IG, y ponerla a su disposición, informándole sobre las características de esa información (actualidad, precio, precisiones, propietario,

limitaciones, etc). Responder a esa información de manera precisa, exige que la información esté catalogada y que existan herramientas capaces de dirigir al usuario para el uso que quiera hacer de esa IG. Si además se pretende que exista un mayor refinamineto en las respuestas, la IDE debe disponer de herramientas de Ingeniería del Conocimiento especializadas en la recuperación de la información. Nos referimos a Nomenclátores, Listas Controladas, Tesauros y Ontologías. Veamos una breve descripción de ellas.

7.3.1 Metadatos

Para lograr ese nivel de información, los datos geográficos deben estar documentados al máximo nivel posible con metadatos estandarizados. Más adelante se verán los estándares más relevantes aplicables a la IG (FGDC, Dublin Core e ISO 19115).
El concepto de metadatos es familiar a la mayoría de aquéllos que manejan temas espaciales. La leyenda de un mapa es una representación de metadatos, que contiene información sobre el editor del mapa, la fecha de publicación, el tipo de mapa, su descripción, referencias espaciales, su escala y su exactitud, entre otras cosas. También son metadatos la información descriptiva que suelen llevar adjuntos los archivos geoespaciales digitales. Son términos y definiciones usados al documentar y utilizar esos datos. Son del tipo "qué", quién", "dónde", "por qué", "cuándo" y “cómo" de los datos..

Los metadatos de la IG deben ser documentados por medio de herramientas especiales y son el ingrediente clave para lograr la localización de los datos, su evaluación y acceso.

7.3.1.1 Catálogos de Datos Espaciales

Para que los datos geoespaciales puedan localizarse y ser utilizados por medio de aplicaciones externas, deben estar documentados (metadatados, utilizando un término que puede ser un neologismo no aceptado). Además, debe existir un servicio que permita localizar y acceder a la .información geoespacial. A ese servicio se le conoce de diferentes maneras. Por ejemplo:

- El Open Geospatial Consortium (OGC) lo llama "Servicios de catálogo"
- La Australian Spatial Data Infrastructure se refiere como "Directorio de Datos Espaciales"
- El FGDC de los EEUU los llama "Clearinghouses"

Esta última se traduce al castellano como "Agencia o Almacén de Distribución" y aunque tengan nombre diferentes, el objetivos es el mismo: localizar datos geoespaciales a través de las propiedades descritas por sus metadatos.

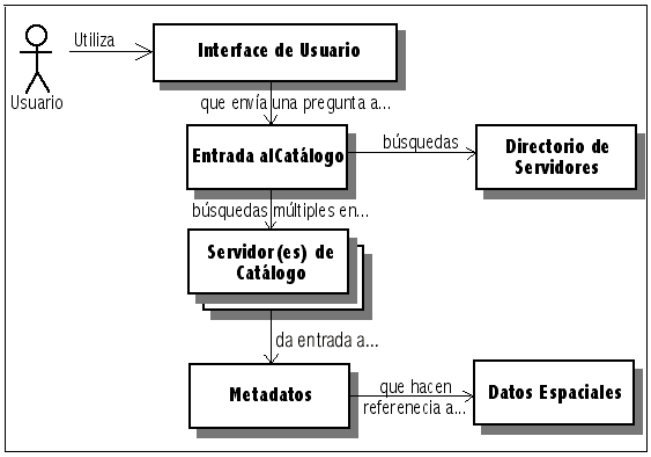

Figura 3
Funcionamiento del Servicio de Catálogo

Como se verá al profundizar en los conceptos de Servidores de Mapas y Servidores de Fenómenos, gracias a la integración de estos servicios de catálogos con otros, se podrá llegar a localizar capas de información procedentes de distintas fuentes de datos, fundirlas y solucionar problemas gracias a los datos resultantes.

7.3.2 Nomenclátores

Uno de los principales problemas de los metadatos es que no siempre están descritos con la suficiente precisión como para identificar de manera precisa la información que disponen. A veces la información de los metadatos puede ser ambigua.

En lo que se refiere a la IG, no cabe duda de que las coordenadas geográficas de un punto, describen el lugar sin ambigüedades. Las coordenadas son los metadatos más importantes de un lugar. El problema es que los humanos no tenemos mucha capacidad para recordar las coordenadas de cada uno de los sitios de los que solicitamos información. Recordamos el nombre del sitio, por ejemplo "Villaverde" pero puede ocurrir que asociado a ese nombre haya una legión de sitios. Hay también otros sistemas de referencia espacial no basados en coordenadas, como son, los Códigos INE, los Registros de Entidades Locales, los Códigos Postales, los Datos Catastrales, la Cuadrícula del Mapa Topográfico Nacional. (http://www.idee.es/show.do?to=pideep_sistemas_referencia.ES)

Los nomenclátores son sistemas para referenciar la IG basados en Identificadores Geográficos unívocos de localizaciones geográficas, (de acuerdo con la Norma ISO19112 que los define). Esos sistemas pueden estar basados en jerarquías administrativas (Países, Provincias, Municipios) y cada una de las instancias que comprenden el conjunto debe disponer al menos de:

a.- Un identificador Geográfico. Ejemplo: Urbanización Pinar del Plantío

b.- Una dscripción de su extensión geográfica. Ej.: Conjunto de viviendas del Término Municipal de Majadahonda, provincia de Madrid (España), limitadas por las calles Ronda de las Sirenas y Paseo de Paseidón.

c.- El nombre de la organización responsable de su definición. Ej. Ayuntamiento de Majadahonda

d.- Las coordenadas de un punto representativo en el caso de falta de unicidad de los descriptores anteriores. Ej.: 40º 27´20 N , 3º51´20

Estos sistemas sustituyen a los antiguos nomenclátores de los Atlas en papel en los que aparecían los nombres de los elementos georreferenciables (poblaciones, parajes, accidentes, etc), junto a sus coordenadas y la hoja del Atlas donde se encontraban.

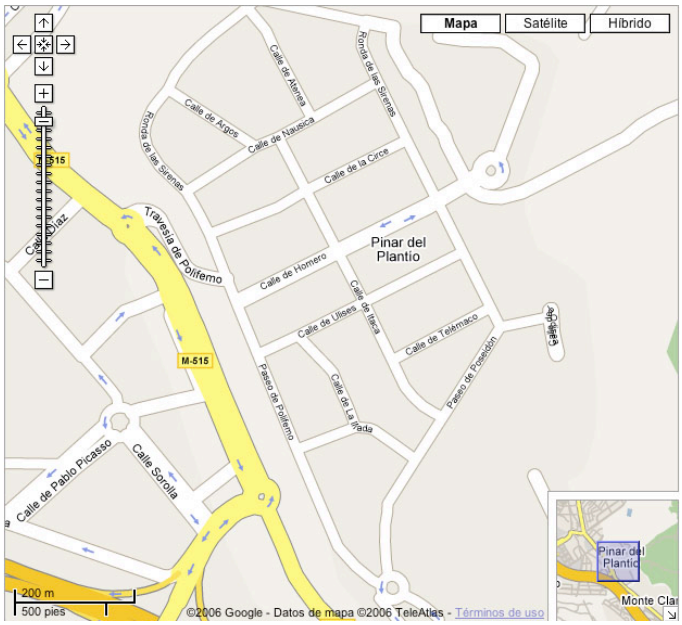

Figura 4
Urbanización Pinar del Plantío. Majadahonda (Madrid)

7.3.3 Listas Controladas

Una forma de catalogar la información sin que aparezcan términos no esperados es proporcionar al usuario herramientas que dispongan de Listas Controladas. Este concepto corresponde a listados de palabras que son los únicos que el usuario puede utilizar en un cierto entorno. Por ejemplo: si vivo en los Estados Unidos de América y en una página web me preguntan "¿País donde vive?" , si no hubiera listas controladas, se podría contestar: <USA>, <US>, <U.S.>, <Estados Unidos>, <United States>, <EEUU>,

<States> … Como vemos, no es fácil acertar con lo que quiere recibir la página web. Lo mejor es que el sistema me ofrezca una Lista de países donde el usuario pueda elegir, sabiendo que eso será lo que espera el sistema (aunque a mi me fastidie vivir en Spain en vez de en España).

En particular, la Norma ISO 19115 define 23 tipos de Listas controladas para la IG, y por ejemplo, para describir el término "Tipo de datos" existe una Lista Controlada que sólo dispone de tres elementos: "Fecha de creación (001)", "Fecha de publicación (002)" y "Fecha de revisión (003)".

Las Listas Controladas dan una enorme estabilidad a los sistemas de información, facilitando las búsquedas y el acceso a la información.

7.3.4 Tesauros

Se denomina Tesauro a una lista estructurada de descriptores o términos propios de un ámbito científico determinado, entre los cuales se establecen una serie de relaciones jerárquicas (gérero/especie, todo/parte), de equivalencia (formas de uso, formas preferidas, sinónimos) y asociativas (términos relacionados). Las palabras que componen un tesauro son el resultado del acuerdo de expertos en el tema, por lo que pueden considerarse como vocabularios controlados. Durante s construcción, se seleccionan y se recomiendan las palabras clave de la temática que aborda. Las búsquedas utilizando palabras clave se simplifican así como la recuperación de la información, pues será posible ubicar textos o datos dentro de una gran serie de volúmenes dispersos. Además, los tesauros permiten ampliar el conocimiento acerca de un tema conociendo nuevos temas generales y específicos relacionados con el tópico de interés. Esto conduce a ampliar o a precisar los criterios de búsqueda.

Aplicando la definición de tesauro a las categorías de nombres geográficos, se puede decir que un Tesauro de nombres geográficos va a mostrar el conjunto de términos que representan los conceptos geográficos clasificados en áreas temáticas determinadas y sus relaciones, estableciendo relaciones de equivalencia, de jerarquía y de asociación con el resto de términos que forman el tesauro.

Interesante: visitar este enlace: http://www.visualthesaurus.com/

7.3.5 Ontologías

La descripción de un cierto dominio de interés, esto es, de sus conceptos y de las relaciones entre ellos, se llama "modelo conceptual del dominio o del área de conocimiento"

Se define una ontología como *una especificación de una conceptualización**, esto es, un marco común o una estructura conceptual sistematizada y de consenso no sólo para almacenar la información, sino también para poder buscarla y recuperarla. (*Conceptualización: Visión abstracta y simplificada del dominio que se quiera representar. M.A. Abián (2005)

Una ontología define los términos y las relaciones básicas para la compresión de un área del conocimiento, así como las reglas para poder combinar los términos para definir las extensiones de este tipo de vocabulario controlado.. En el campo de la informática, los modelos conceptuales deben transformarse para que puedan ser almacenados en un ordenador y sobre los que puedan aplicarse algoritmos.

Se trata de convertir la información en conocimiento mediante unas estructuras de conocimiento formalizadas (las ontologías) que referencien los datos, por medio de metadatos, bajo un esquema común normalizado sobre algún área del conocimiento. Los metadatos no sólo especificarán el esquema de datos que debe aparecer en cada instancia, sino que también podrán contener información adicional de cómo hacer deducciones sobre ellos, es decir, cómo establecer axiomas que podrán, a su vez, aplicarse en los diferentes dominios que trate el conocimiento almacenado. De esta forma, los buscadores podrán obtener información al compartir los mismos esquemas de anotaciones web y los agentes de software no sólo encontrarán la información precisa, sino que podrán realizar inferencias de forma automática buscando información relacionada con la que se encuentra situada en las páginas web y con los requerimientos de las consultas realizadas por los usuarios. Además, los productores de páginas y servicios web podrán intercambiar sus datos siguiendo estos esquemas comunes consensuados e, incluso, podrán reutilizarlos. (M.J. Lamarca. 2006)

Los beneficios de utilizar antologías se pueden resumir en:

- proporcionan una forma de representar y compartir el conocimiento utilizando un vocabulario común
- permiten usar un formato de intercambio de conocimiento
- proporcionan un protocolo específico de comunicación
- permiten una reutilización del conocimiento

7.4. Componentes de la Web Semántica

Y como no voy a escribir nada mejor que lo que ha escrito María Jesús Lamarca en su página web, no puedo por menos que decir que todo lo que se quiera encontrar, con un lenguaje accesible, se encuentra en:
http://www.hipertexto.info/documentos/web_semantica.htm

5.7. Bibliografía y enlaces

Enlaces:

Un acercamiento sencillo a la web Semántica y al concepto siempre escurridizo de Ontologías http://www.wshoy.sidar.org/index.php?2005/12/09/30-ontologias-que-son-y-para-que-sirven

M.A. Abián (2005). *El futuro de la Web*. Un acercamiento sencillo y bien documentado (en español) sobre la web Semántica y sus contenidos http://www.javahispano.org/tutorials.item.action?id=55

M.J. Lamarca Lapuente (2006) Tesis doctoral: Hipertexto: el nuevo concepto de documento en la cultura de la imagen. http://www.hipertexto.info/documentos/ontologias.htm

Metadatos Geográficos y sus aplicaciones. Un sencillo artículo sobre las posibilidades de los metadatos y Google Earth (todo el siguiente URL va unido)
http://www.wshoy.sidar.org/index.php?2005/08/01/24-metadatos-geograficos-1-obteniendo-informacion-geografica-para-la-web-semantica