>>> revista tradumàtica
Traducció i Technologies de la Informació i la Comunicació
Nómero 2 : La Documentació : Novembre 2003 : http//www.fti.uab.es/tradumatica/revista

Internet invisible y web semántica: ¿el futuro de los sistemas de información en línea?

Lluís Codina, Universitat Pompeu Fabra

Resumen
Los cambios en la Web Invisible, cuyas fronteras no dejan de retroceder, y el desarrollo de la web semántica marcarán buena parte del futuro de los sistemas de búsqueda en línea de los próximos años. La Web Invisible es cada vez menos invisible porque, poco a poco, y de una forma u otra, sus contenidos se van incorporando a los motores de búsqueda. La web semántica es la más ambiciosa e importante apuesta tecnológica y científica del W3 Consortium y afectará en gran medida a los desarrollos futuros en sistemas de representación y acceso a la información

Palabras Clave
Web Invisible, Formatos de documentos, Motores de búsqueda, Web semántica

---

Resum
Els canvis a la Web Invisible, les fronteres de la qual no deixen de retrocedir, i el desenvolupament de la web semàntica, marcaran bona part del futur dels sistemes de cerca en línia en els propers anys. La Web Invisible és cada vegada menys invisible, perquè lentament i d'una o altra manera, els seus continguts es van incorporant als motors de cerca. La web semàntica és la més ambiciosa i important aposta tecnològica i científica del W3 Consortium, i afectarà en gran mesura els desenvolupaments futurs de sistemes de representació i accés a la informació

Paraules clau:
Web Invisible; Formats de documents; Motors de cerca; Web semàntica

0. Metodología

Para este trabajo nos hemos basado ampliamente en los resultados obtenidos de un proceso de análisis sistemático de funciones de búsqueda y representación de la información en sistemas de información documental en línea, entre los que destacan los análisis realizados a motores de búsqueda, multibuscadores y bases de datos en línea. Este trabajo se ha beneficiado también de las discusiones de un grupo de expertos[1] que se llevan a cabo en un seminario de sistemas de información documentales desarrollado a lo largo del año 2003 y coordinado por el autor en el seno de dos proyectos de investigación financiados que se llevan a cabo en el Instituto de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF).

1. Internet invisible

Internet invisible es un nombre claramente inadecuado para referirse al sector de sitios y de páginas web que no pueden indizar los motores de búsqueda de uso público como Google o AltaVista. Pese al nombre, afortunadamente, la web invisible es perfectamente visible ya que los contenidos de tales páginas y sitios web pueden ser vistos o bien mediante un navegador

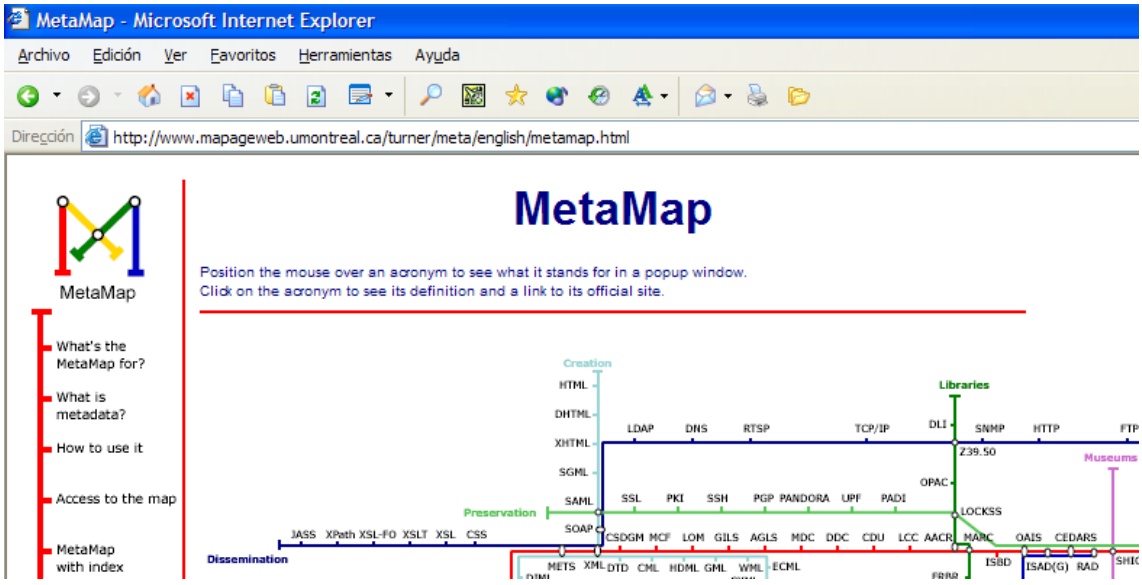
[1] Participaron habitualmente en el *Seminario de Documentación-IULA* en el período cubierto por este trabajo (enero-junio 2003): Miquel Centelles, Mercè Lorente, Mari Carmen Marcos, Gemma Martínez, Maria del Valle Palma y Cristòfol Rovira.

convencional o bien mediante un navegador complementado con algún programa adicional (*plugin*).

Por tal motivo, debería denominarse, en realidad, la web "no indizable", lo cual es un término mucho más adecuado, pero claramente alejado de la capacidad sugeridora del término "invisible". Dado que, sin embargo, es el término más habitual incluso en la bibliografía técnica, usaremos en este trabajo el término Web o Internet invisible para referirnos a la información publicada en servidores Web que por diversos motivos no puede ser indizada y, por tanto, no puede ser encontrada por los motores de búsqueda convencionales.

Veamos ahora por qué hay contenidos no indizables en la Web. Hay al menos tres motivos. En un orden no significativo, podemos decir que el primer motivo son los formatos de los documentos. Los motores de búsqueda fueron creados originalmente para descargar, leer e indizar páginas HTML. Cualquier otro formato era ilegible, es decir, *invisible* para tales motores. Todos sabemos de la proliferación de formatos no HTML en la Web (que sin embargo se integran con toda facilidad en el navegador). Es el caso, por ejemplo, de los cada vez más abundantes documentos en formato .pdf (documentos Acrobat) e incluso en formato .doc (documentos Word). En la medida en que una parte de los contenidos de la Web está formada por documentos no HTML, esa parte es candidata a ser Internet invisible.

Figura 1. Parte de un documento en formato no HTML (svg) visto en un navegador
(http://www.mapageweb.umontreal.ca/turner/meta/english/)



El segundo motivo son las páginas que se generan de forma dinámica; típicamente, a través de la consulta a una base de datos. Por ejemplo, si usamos *All Movie* (www.allmovie.com) para buscar información sobre un film obtendremos una URL como la que indica la figura siguiente:

Figura 2. URL de un documento de la web invisible

http://www.allmovie.com/cg/avg.dll?p=avg&sql=A169

Los motores de búsqueda no pueden indizar contenidos que se generan de ese modo. Antes de lanzar la búsqueda, el contenido existe en el formato binario (y propietario) de alguna base de datos. Solamente después de la consulta, y como resultado de ejecutar una instrucción

como la que muestra la figura anterior, se creará una página en formato HTML. El lector puede hacer la prueba, si copia la URL de la figura anterior (que contiene una consulta a una base de datos), y la introduce como dirección en un navegador obtendrá una página HTML que le informará sobre un film determinado. Antes, sin embargo, esa página no existía. En la imagen siguiente puede ver el resultado.

Figura 3. Resultado de la página generada dinámicamente con la URL anterior (http://www.allmovie.com/cg/avg.dll?p=avg&sql=A169)

En el caso de bases de datos como la anterior, los motores de búsqueda pueden proporcionar acceso a la página de inicio (*home page*) de la misma. Si hacemos una consulta por el término **movies** obtendremos entre los resultados (aunque en este caso hemos necesitado llegar hasta la tercera página) una entrada que se refiere a AllMovie, como podemos ver en la ilustración siguiente:

Figura 4: Uno de los resultados de buscar en Google por el término movies

Es decir, podemos acceder a las páginas principales de los sitios web que proporcionan acceso a bases de datos, porque tales principales son páginas HTML convencionales, pero no podemos acceder al resto del sitio a través del motor de búsqueda; y el resto del sitio puede ser (en ocasiones) una enorme base de datos.

Por ejemplo, si lanzamos la consulta **2001** en Google, en ninguno de los resultados obtenemos la ficha del film correspondiente de *All Movie*. De hecho, obtendremos una diversidad de resultados que refleja que el término 2001, fuera de contexto, tiene muchos significados y no necesariamente el de título principal de un film de Kubrick.

Figura 5: Resultado de una búsqueda en un motor por el término "2001". (Obsérvese, por encima del primer resultado, la remisión a una categoría del directorio)

Por último, forma parte de la web invisible el conjunto de sitios o de páginas web que, de forma expresa, se excluyen de la actividad indicadora de los motores de búsqueda. Algunos servidores excluyen a los motores de búsqueda de todos o de parte de sus carpetas y directorios mediante el uso de un protocolo de exclusión que, en general, respetan los programas rastreadores (*spiders* o *crawlers*) de tales motores de búsqueda. Tal protocolo consiste en un pequeño número de valores que puede adquirir el atributo *content* como parte de una etiqueta **meta** cuyo otro atributo, *name*, obtiene el valor "robots". Estas indicaciones se guardan en un simple archivo de texto de nombre robots.txt que se sitúa en el servidor de página web y que se supone que leen y respetan los rastreadores (robots). La figura siguiente muestra el uso de tal protocolo para indicar a los robots de los motores que no indicen la página en cuestión ni sigan ninguno de los enlaces que pueda contener tal página.

Figura 6: Ejemplo de exclusión de motores de búsqueda de un sitio web

```
<meta name="ROBOTS" content="noindex,nofollow">
```

Además del protocolo que acabamos de ver, hay otras razones por las cuales los motores no pueden entrar en un sitio. En general, cualquier sitio web que requiera el uso de contraseñas o *passwords* quedará fuera de la capacidad indizadora de los motores. Estos sitios pueden ser extranets o servicios que requieren no solamente una suscripción previa, sino que exigen el pago de una cantidad en concepto de abono, etc.

Los motores también tienen dificultades para interpretar los sitios que usan marcos (*frames*), aunque son de otro tipo y no las consideraremos aquí.

La cuestión es que, en total, algunos analistas señalan que la Web Invisible puede ser hasta 500 veces más grande que la Web visible (Bergman, 2001). Desde el punto de vista del acceso al conocimiento y de la clase de búsqueda y obtención de la información que nos interesa aquí, no hay ningún problema con que una parte de la Web Invisible siga siendo invisible.

Por ejemplo, no es ninguna tragedia para el desarrollo de la ciencia o del conocimiento humanos que la extranet o la intranet de una corporación sea invisible a los motores de búsqueda. No solo no es un problema, sino que es deseable que siga siendo así. Nadie quiere que los motores de búsqueda puedan indizar documentos administrativos particulares o informaciones confidenciales.

Por tanto, de las tres razones por las cuales tenemos una Internet Invisible, una de ellas no es ningún problema, pero las otras dos sí. Recordemos: documentos con formato no HTML y páginas generadas dinámicamente (típicamente a través de bases de datos).

Con la imposibilidad de indizar documentos no HTML tenemos, efectivamente, un auténtico problema. Muchos informes y estudios que contienen información valiosa están publicados y disponibles en la web de forma pública y abierta; sin embargo, si no son indizados de forma adecuada, son inaccesibles a casi todo el mundo a casi todos los efectos prácticos.

Por otro lado, no deja de ser un problema que, pese a disponer de un cliente universal de acceso a la información: el navegador web, no exista, en cambio, nada similar a una interfase universal de acceso a la información desde el momento en que, para cada una de las varias decenas de miles de bases de datos existentes en Internet sea necesario: primero, un acceso diferenciado y segundo un sistema de consulta (en parte) diferente.

En este último caso, obsérvese que las barreras al conocimiento son dos: el conocimiento de las fuentes y el dominio de la interfase de usuario de cada fuente. En efecto, en primer lugar, para que un usuario pueda beneficiarse de los contenidos de una base de datos es necesario, al menos, que sepa de su existencia. Pero, suponiendo que sepa de su existencia, entonces deberá tener habilidades de uso de tal base de datos y cada base de datos no solamente presenta una interfase de usuario diferente, sino un conjunto de funciones distintas.

2. Acceder a los contenidos de Internet Invisible

2.1. Formatos no html

Pese a todo, se puede acceder a cada vez mayores "porciones" de la Web Invisible. Examinemos primero el caso de los formatos de documentos. Afortunadamente, en este aspecto, las fronteras de la Web Invisible no hacen más que retroceder. La tabla siguiente ilustra los formatos que, en estos momentos, son capaces de indizar (o al menos de buscar) dos de los motores más potentes de la Web:

Figura 7. Tabla de formatos "buscables" a través de Google y AllTheWeb (además de html**)**

Motor	***Formatos***
Google www.google.com	Acrobat (pdf)
	Postscript (ps)
	Word (doc)
	Excel (xls)
	PowerPoint (ppt)
	Texto Enriquecido (rtf)
AllTheWeb www.alltheweb.com	Acrobat (pdf)
	Flash (swf)
	Word (doc)

Vemos que, en el momento de realizar este trabajo, *Google* busca (y probablemente indiza) 6 formatos distintos de documentos (además, claro, del formato HTML) y *AllTheWeb* (uno de los

alumnos no solamente aventajados, sino respondones de Google) busca y/o indiza 3 formatos distintos.

En este sentido, parece que la tendencia es clara: poco a poco, la mayor parte de los formatos de documentos significativos en el mundo científico y cultural serán indizados por los motores de búsqueda y, por tanto, esa zona de la Web Invisible dejará de serlo pronto. Además, hay dos factores más que confluyen en este aspecto: por un lado, los navegadores cada vez incorporan con mayor facilidad documentos no HTML. Es ejemplar, en este sentido, la integración de las últimas versiones de los navegadores y el formato pdf. Por otro lado, el progresivo ancho de banda disponible en manos de los usuarios (ADSL, por ejemplo) hace que esa integración sea transparente.

De este modo, si los motores tienden a lo que podríamos llamar una "indización universal" y los navegadores (o agentes de usuario) tienden a poder mostrar cualquier tipo de documento, podemos concluir que este aspecto de la Web Invisible está llamado a ser marginal.

Ahora bien, a veces las soluciones a los problemas aportan también problemas nuevos. A medida que formatos como pdf y word se integran en la Web con mayor naturalidad, para beneficio de los usuarios, desciende el grado de conectividad general de la Web.

Es decir, una de las virtudes de la Web es la facilidad con la cual se pueden publicar páginas web (o sitios enteros) ricamente interconectados de forma interna, así como la facilidad para conectar páginas y sitios web remotos. Sin embargo, parte de esas facilidades desaparecen con formatos como pdf y word. Es cierto que un documento pdf, por ejemplo, puede contener enlaces internos o externos, pero en la práctica, se publican documentos pdf como una forma fácil de obtener una publicación de calidad tipográfica con mínimo esfuerzo. En la práctica, por tanto, la inmensa mayoría de documentos pdf están muy pobremente interconectados.

2.2. Bases de Datos

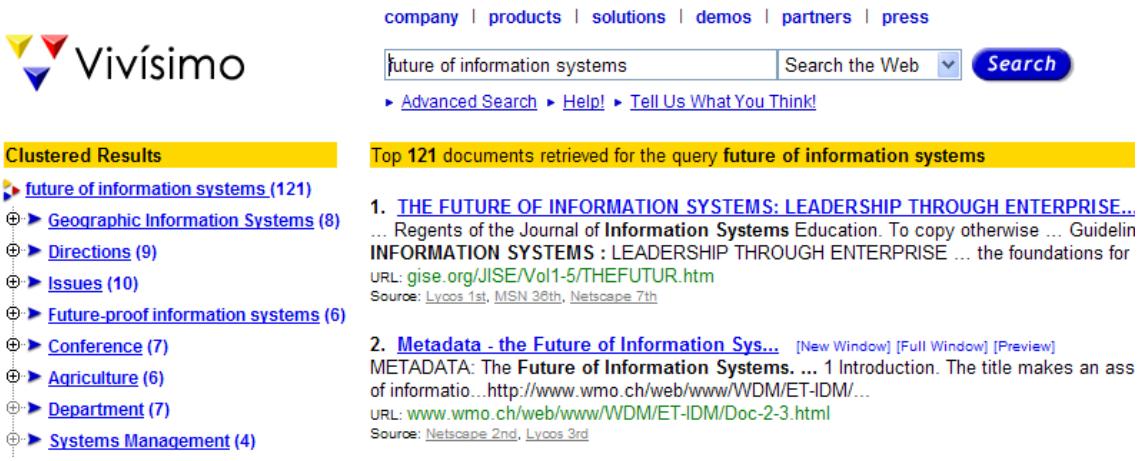
También tenemos indicios de solución al segundo gran "problema" de la Web Invisible: el acceso al contenido de las bases de datos, pero desde motores convencionales.

La solución aquí proviene de este enfoque: si bien es difícil o imposible indizar por parte de los motores de búsqueda el contenido de bases de datos ajenas, no debería haber mucha dificultad en generar interfases de consulta unificadas que enviaran una misma consulta a diferentes bases de datos desde, por ejemplo, una misma página web. El modelo en este caso son los multibuscadores, también (mal) llamados metabuscadores.

Un multibuscador es un sistema que acepta como entrada la pregunta de un usuario y devuelve en una respuesta unificada las respuestas de diversos motores de búsqueda.

Un buen ejemplo de multibuscador es Vivísimo (www.vivisimo.com). Una búsqueda en *Vivísimo* por los términos **future of information systems** muestra como resultado una compilación de la información ofrecida por diversos buscadores.

Figura 8: El resultado de una búsqueda en Vivísimo
(www.vivisimo.com)



Compilar información, en el caso de Vivísimo significa que no se limita a volcar los resultados que envía cada buscador, sino que: (a) unifica resultados (o sea, elimina duplicados); (b) distribuye los resultados por grupos o pseudo categorías que el sistema de agrupación (*clustering)* de Vivísimo es capaz de generar de manera automática.

Pero lo que nos interesa aquí examinar es la siguiente idea: Vivísimo no intenta explotar directamente los índices de los distintos motores de búsqueda. En su lugar, hace algo más viable: envía la pregunta a diversos motores y procesa los resultados antes de ofrecerlos al usuario. Esta operación le permite ofrecer un resultado unificado cuyas fuentes, sin embargo, tienen procedencias muy diversas.

Figura 9: Opciones de búsqueda en Vivísimo
(www.vivisimo.com)

Ahora bien, si observamos con atención la figura n. 9 podremos ver que entre las fuentes que utiliza *Vivísimo* (tomamos este sistema solamente a título de ejemplo) vemos que hay, al menos tres clases de fuentes: (1) motores de búsqueda como *AltaVista* (hasta aquí ninguna novedad), (2) sitios web de noticias como *Reuters* y (3) sitios web de bases de datos como

PubMed. ¿Qué significa esto? Simplemente, que *Vivísimo* es solamente una muestra de cómo se están derribando parte de las fronteras de la Internet Invisible.

2.3. Sindicación de contenidos

Otro ejemplo sumamente interesante y buena muestra de lo que, probablemente, nos espera en los próximos años es el motor de búsqueda *Scirus* (www.scirus.com). Es aún pronto para saber si *Scirus* será un experimento efímero, como tantos otros proyectos esperanzadores en la web (esperemos que esta vez no) o solamente un avance de una nueva generación de sistemas de búsqueda en línea que rompa de una vez por todas las barreras de la Web Invisible.

Scirus es un proyecto de una importante editorial científica, *Elsevier*, que ha producido un motor que es capaz de enviar las preguntas de los usuarios a las bases de datos que indica la tabla de la Figura 8.

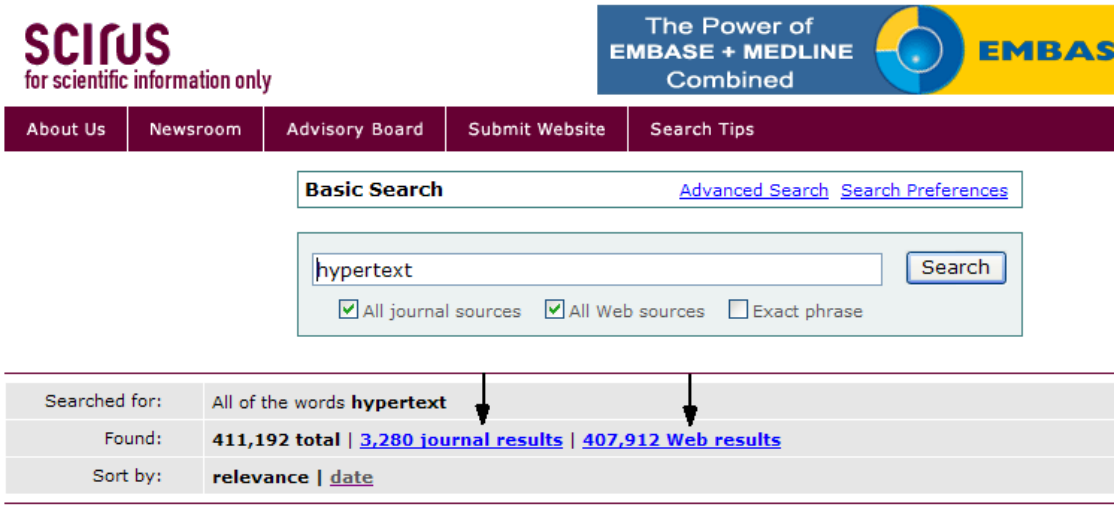
Figura 10. Bases de datos que puede interrogar *Scirus* de forma simultánea

- Medline
- Sciencedirect
- Uspto
- Beilstein Abstracts
- E-Print Arxiv
- Nasa Technical Reports
- Cogprints
- Biomed Central
- Mathematics Preprint Server
- Chemistry Preprint Server
- Computer Science Preprint Server

Además, *Scirus* indiza casi 90 millones de páginas web, es decir, documentos en formato HTML publicados en servidores de páginas web convencionales, pero siempre vinculados con instituciones académicas o científicas. De este modo, el usuario de *Scirus*, típicamente un investigador o un profesional, cuando realiza una búsqueda en este motor obtiene dos tipos de resultados: (1) páginas o sitios web relacionados con la ciencia, la universidad, etc.; (2) artículos de revista o registros referenciales procedentes de bases de datos de ciencia y tecnología (o sea, una parte de la Web Invisible).

Scirus, por tanto, es uno de los mejores ejemplos que tenemos ahora a nuestro alcance de lo que pueden ser los futuros sistemas de información en línea: una interfase unificada de información a fuentes diversas.

Figura 11: Un típico resultado en *Scirus* puede incluir artículos en texto completo procedentes de diversas bases de datos



Podemos concluir, en relación a este apartado, que las barreras de la Internet Invisible probablemente van a ir cediendo, una a una, hasta que los contenidos no indizables de Internet sean exactamente los que deben ser: porciones de la web que sus administradores o propietarios, en uso legítimo de sus prerrogativas, no desean que sean indizados.

En cambio, los contenidos de la Internet Invisible correspondientes a formatos no HTML y parte del contenido que se encuentra en el formato binario de distintas bases de datos, serán accesibles desde motores de búsqueda públicos, del tipo *Google* o *Scirus*.

Lo que esto último significa es que los productores de bases de datos deberán comenzar a plantearse si desean, por así decirlo, sindicar sus contenidos a los motores de búsqueda. Un modelo puede ser el que representa *Scirus.* Los productores de bases de datos pueden decidir que entra en sus intereses permitir la recepción de consultas y el envío consiguiente de resultados a uno o más motores de búsqueda, conscientes que los usuarios finales siempre persiguen, de una forma u otra, la idea (en parte utópica) de la interfase de consulta universal.

Naturalmente, sindicación de contenidos implica también un modelo de negocio. Implica que los motores de búsqueda como Google o bien estén dispuestos a retribuir a los productores de las bases de datos, o bien que, a partir de un momento dado, una parte de los resultados ofrecidos por el sistema sea de acceso libre y otra sea de acceso condicionado al pago de una cierta cantidad o la condición de ser abonado o suscriptor.

Esto último es lo que hace *Scirus*. Cuando un usuario lanza una búsqueda en *Scirus* puede encontrar tres tipos de resultados: (1) documentos de acceso totalmente libre, por ejemplo, un estudio publicado como una página web en un servidor web convencional y de acceso libre; (2) documentos a los que tiene acceso debido a que su institución posee una suscripción a la publicación correspondiente, por ejemplo un artículo de una revista suscrita por la biblioteca de su institución; (3) documentos a los que tiene acceso mediante pago con tarjeta de crédito.

3. La web semántica

3.1. Definiciones

Ante todo, veamos la definición oficial de web semántica (*semantic web*). Según el *W3 Consortium* (el organismo promotor de la idea):

Definition: The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.

Dos cosas sobre la definición anterior. En primer lugar, como se puede observar no dice absolutamente nada: ¿qué significa que alguna cosa sea "la representación de datos en la World Wide Web"? Nada. El resto de la supuesta definición es peor. Abandona claramente el intento de decir lo que es la web semántica (dado el antecedente, tal vez sea lo mejor) y se limita a señalar, entre otras cosas sumamente informativas "que integra una variedad de aplicaciones"(!).

La segunda cosa que corresponde señalar es que la web semántica no existe. No sabemos si la web semántica será realidad algún día, pero hoy por hoy, ni existe "ni se la espera" (al menos de manera inminente). Pese a ello, se debe reconocer en ella a una auténtica idea-fuerza, en el sentido de que es una idea que ya ha sido capaz de movilizar muchas energías (y muchas ilusiones) y que, sin duda no dejará de arrojar resultados durante los próximos años porque sin duda seguirá movilizando energías.

Es una idea tal, por decirlo de alguna forma, semejante a los viajes que tienen sentido por sí mismos, independientemente del destino previsto. Dicen los expertos en narrativa que toda auténtica aventura es en realidad un viaje en el cual, al final del mismo el protagonista ha sufrido alguna transformación (se supone que para bien). La web semántica puede verse, así, como un viaje que inicia ahora la World Wide Web y tal vez no alcance nunca (del todo) su destino, pero que, entre tanto, la transformará profundamente.

Si tuviésemos que proponer una definición de la web semántica, nosotros empezaríamos con esta:

Definición: La Web Semántica es un conjunto de iniciativas, tecnológicas en su mayor parte, destinadas a crear una futura World Wide Web en la cual los ordenadores puedan procesar la información, esto es, representarla, encontrarla, gestionarla, como si los ordenadores poseyeran inteligencia

En lo que sigue, intentaremos presentar una aproximación a la idea de la web semántica; para ellos nos hemos basado en un trabajo previo (Codina, 2003) pero, sobre todo, en la información que sobre la web semántica puede encontrarse en el ya mencionado organismo promotor de la idea, el *W3 Consortium* (www.w3.org/2001/sw/), y en un famoso y citadísimo artículo publicado en *Scientific American* (Berners-Lee, 2001). Hemos consultado también otros autores que se indican en la bibliografía.

3.2. Estado actual

Si la web semántica no existe, ¿qué es en estos momentos? De momento, es el nombre de una aspiración; el nombre de un objetivo muy ambicioso que, de cumplirse, cambiaría de forma radical la Web tal como la conocemos hoy. ¿En qué consiste esta aspiración? Ni más ni menos que en conseguir que las páginas que forman la Web dejen de ser simples cadenas de caracteres para los ordenadores y se conviertan en textos con sentido, es decir, texto provisto de semántica, tal como, de hecho, lo es para los seres humanos.

¿Porqué un objetivo semejante? Tal como se codifican las páginas web actuales, principalmente mediante el lenguaje HMTL, tienen muy poco sentido para las máquinas. En efecto, si vemos el código fuente de una página web actual, encontramos, por ejemplo, un trozo de código como el siguiente:

```
…
<b><i>Superar la brecha digital</i></b>
…
```

cuando el ordenador lo interprete, a través del programa navegador, aparecerá como un texto en negrita y cursiva, como éste:

```
…
Superar la brecha digital
…
```

Con esto casi se acaba casi todo lo que es capaz de hacer un ordenador con las páginas HTML. Como saben bien informáticos y documentalistas, otra cosa que pueden hacer los ordenadores es construir índices con las palabras que aparecen en las páginas web. Después cuando alguien envía una pregunta a un motor de búsqueda, lo que hace este último es comparar las palabras de la pregunta con las palabras de su índice. Por ejemplo, supongamos que el responsable de un programa de gobierno sobre el problema de la brecha digital decide indagar en Internet para ver si encuentra estudios o informes sobre la brecha digital.

Supongamos que accede a Google y entra la siguiente pregunta: "brecha digital". Lo que hará Google es comparar las palabras de su pregunta, con las palabras de su índice. Si encuentra un documento que tenga la "brecha digital", lo devolverá como respuesta. Esto es casi todo lo que pueden hacer los ordenadores que tenga que ver con procesamiento de información en páginas web.

Con estas limitaciones, la búsqueda en Internet, como todo el mundo sabe, está repleta de frustraciones. Si alguien busca por "caballos", no encontrará nada que trate sobre "yeguas". Si alguien busca sobre cómo evitar la guerra, no encontrará un documento sobre cómo conseguir la paz, etc. La web semántica quiere solucionar esto. Esto suena a inteligencia artificial. Por tanto, aunque no quieran llamarlo así, con la web semántica se está buscando el mismo objetivo, a saber, que los ordenadores entiendan que un documento sobre "yeguas" puede ser muy relevante para una necesidad de información sobre "caballos", y que la semántica de la pregunta "¿es posible evitar la guerra?" es la misma que la de la pregunta "¿es posible conseguir la paz?".

Además, se espera que los ordenadores puedan desarrollar tareas de gestión que requieran interpretar información y tomar decisiones adaptándolas al contexto. Se trata ni más ni menos que de un objetivo al que la informática ha denominado hasta ahora inteligencia artificial.

3.3. Infraestructura

Los medios con los cuales se supone que se conseguirá la web semántica son los siguientes: primero, un nuevo lenguaje de codificación de páginas, un nuevo lenguaje de marcado. Este lenguaje, como es sabido, se denomina XML. Con XML se pueden diseñar lenguajes de marcado muy estructurados y muy explícitos en los cuales, en lugar de etiquetas como <b> e <i>, habrá etiquetas como <título>, <subtítulo>, <capítulo>, <subcapítulo>, <autor>, <institución>, <ciudad>, etc.

Como harán falta etiquetas específicas para cada tipo de información -por ejemplo, las páginas web de las compañías aéreas necesitarán etiquetas como <vuelo>, <hora de salida>, <destino>, etc.-, se ha creado, como es sabido, una especificación, una especie de metalenguaje, XML, que permite definir lenguajes específicos, es decir conjuntos de etiquetas específicos para cada necesidad de información. Por ejemplo, los editores de diarios disponen ya de su propio conjunto de etiquetas, así como los matemáticos para expresar ecuaciones, etc.

El segundo elemento con el que se cuenta son los metadatos. Como saben muy bien los documentalistas, los metadatos son información sobre la información y son, en realidad, una antigua fórmula. Los catálogos de las bibliotecas son metadatos. La venerable norma ISBD es una norma sobre metadatos, los descriptores asignados a un documento son metadatos, los tesauros y clasificaciones son lo que ahora en el argot de los metadatos se denominan también *schemes*, etc.

La cuestión es que las páginas web ya tienen metadatos. Al menos, suelen tener el metadato título, en forma de etiqueta <title> en una zona de las páginas web invisible para las personas, pero visible para los ordenadores. Además, algunas páginas, muy pocas, suelen tener otros metadatos, como <keyword>, <description>, etc.

Como es sabido, existe una ambiciosa norma de alcance internacional, *Dublin Core*, que proporciona una lista unificada y normalizada de hasta quince metadatos del tenor de los ya comentados para que los editores y autores que lo deseen los incluyan en sus páginas web. La idea es simple: si las páginas web tuvieran metadatos del tipo <título>, <autor>, <tema>, <lugar de publicación>, etc., los usuarios podríamos hacer preguntas mucho más precisas a los motores de búsqueda. Podríamos, por ejemplo, hacer peticiones de información de este tenor: "búscame documentos publicados en tal o cual lugar y que traten de este y este tema, bajo este punto de vista".

Pero los metadatos actuales no tienen ni semántica ni sintaxis ni están unificados bajo una norma común que agrupe la diversidad de plataformas de metadatos existentes.

Para dotarlos de esas tres cosas, se han desarrollado otras normas. La más importante se denominada RDF (*Resource Description Framework*). Esta norma especifica una gramática lógica para que los autores de páginas web puedan describir las propiedades semánticas de los documentos en una notación estándar y común para cualquier tipo de metadatos. Se trata de una notación basada en nociones fundamentales. Básicamente: hay objetos, tales como páginas web, y los objetos tienen propiedades, tales como un responsable intelectual, una fecha de publicación o un contenido expresado en palabras clave, etc. Así mismo, hay relaciones entre los objetos, como una página web que forma parte de una serie o es una versión en otra lengua de otra página web, etc.

Para describir el contenido semántico y otras propiedades de una página web, se puede utilizar la norma RDF mediante el procedimiento de etiquetado XML para expresar los temas de un documento, entre otras cosas.

En síntesis, la gran esperanza de la web semántica se basa, al menos, en tres cosas: XML para hacer los documentos más explícitos; metadatos (expresados también en XML) para hacer los documentos más fáciles de representar, indizar y buscar y, finalmente -se desprende de lo anterior, aunque suele obviarse- una nueva generación de software -programas y métodos de representación del conocimiento- que sepa explotar las dos cosas precedentes.

La representación del conocimiento necesitará, a su vez, procedimientos normalizados, ya sea para representar conocimiento complejo o de sentido común. Estas representaciones suelen denominarse ontologías. Un campo interdisciplinario donde suelen confluir diversas disciplinas cognitivas, desde la inteligencia artificial hasta la lingüística.

Ahora bien, en el esquema de la web semántica se supone que los metadatos los ponen principalmente los propios autores de los documentos. ¿Cuál es el problema? Varios: en primer lugar, los autores no suelen estar entrenados para poner metadatos, y se necesita mucha formación para saber elegir buenas palabras clave.

En segundo lugar, los autores -no todos, ni mucho menos- mienten. Así de simple. Quieren que sus páginas web queden muy alto en los buscadores, de manera que colocan treinta veces la misma palabra, con pequeñas variantes, para que queden muy alto en los *rankings* de los motores de búsqueda para los temas que a ellos les interesa, aunque su página no tenga en realidad mucho (o nada) que ver con ese tema.

En tercer lugar, las personas nos equivocamos, y los autores de las páginas web se equivocan: se olvidan de poner metadatos, los ponen mal, los ponen en unas páginas sí y en otras no, se equivocan en la ortografía, etc. Conclusión: casi ningún motor de búsqueda se fía de los metadatos para generar los resultados de sus *rankings*.

3.4. Posibilidades reales a corto y a medio plazo

El lector ya habrá deducido que, al menos según la opinión de quien esto escribe, las posibilidades a corto y medio plazo de la web semántica son reducidas.

Efectivamente. Una cosa es que se trate de un objetivo que vale la pena perseguir y otra que se trate de un objetivo factible. Permítanme un ejemplo muy significativo. Sin duda es un buen objetivo (al menos, muchos lo creemos así) acabar con la pobreza en el mundo. Es un ejemplo de un fin loable, con el que todos deberíamos comprometernos. Pero que sea un objetivo magnífico y muy deseable en sí mismo, no lo convierte automáticamente en alcanzable; al menos no en su totalidad y no a medio o a corto plazo. ¿Debe por ello abandonarse? Ni mucho menos. Todo lo contrario. Debe perseguirse con ahínco, porque es la única forma de conseguir progresos en tales terrenos, aunque sean parciales.

El problema con la web semántica, tal como la presentan algunos de sus defensores (notablemente, el *W3 Consortium,* que parece haberse especializado en arrojar confusión sobre todos sus proyectos recientes) es la inmensa cantidad de ingenuidad o de ignorancia que exhibe. En comparación, los programas contra la pobreza y a favor de los derechos humanos son obras maestras de pragmatismo (y sabiduría).

Sigamos, por ejemplo, con los metadatos: si casi nadie usa metadatos ahora, ¿por qué razón, de pronto, todo el mundo va a poner metadatos en sus páginas? Para peor, si los autores de páginas web han demostrado su incapacidad para usar una norma relativamente simple como era la primera versión de *Dublin Core*, ¿por qué van a hacerlo ahora que ha llevado su complejidad al límite de lo impracticable?

Por último, respecto a las ontologías y su explotación mediante motores de inferencia o sistemas expertos: si la inteligencia artificial suma ya varias décadas de fracasos, por lo menos en relación a la hipótesis fuerte, o sea en relación a su objetivo declarado a bombo y platillo de lograr que los ordenadores piensen, ¿por qué va a tener éxito ahora?

Por tanto, las posibilidades de que la web semántica sea una realidad tal como la presenta el *W3 Consortium*, sin que se produzca antes, al menos un cambio de paradigma de gran calado en las ciencias de la computación, son ridículas. Además, necesitaremos en paralelo cambios no menos importantes en otras áreas, incluyendo, por supuesto, en las ciencias de la documentación.

Sin embargo, no nos engañemos: el objetivo de la web semántica es magnífico, producirá importantes avances en algunos o en todos los terrenos relacionados con la representación y el acceso al conocimiento y en mi opinión, desde las ciencias de la documentación, debería obtener todo nuestro apoyo.

3.5. ¿Labor de ONG?

¿Cuál es el problema general, casi diríamos filosófico, de la web semántica? Si no se produce algún cambio pronto, el problema de la web semántica es que no proporciona ningún beneficio individual, aunque promete grandes beneficios sociales.

Lo anterior es una definición del fracaso. La historia nos dice que casi siempre que para alcanzar algún objetivo socialmente deseable se requiere un sacrificio individual, el fracaso estará servido. Según los economistas (no es que la economía tenga un historial muy brillante de predicciones, pero vamos a escucharlos por si acaso), es casi imposible conseguir una sociedad viable a base de esperar que los ciudadanos, espontáneamente, vayan contra sus intereses individuales.

Si acaso, podemos esperar resultados si la clase de sociedad que queremos es posible construirla mediante el hecho de que cada ciudadano persiga la consecución de su interés egoísta. A esta visión, que en caso de tener algo de cierto, ayudaría a explicar porqué ha triunfado el capitalismo y se ha hundido el comunismo, se opone la realidad de las ONG.

Suponiendo que sea cierto que, espontáneamente (es decir, excluyendo procedimientos *manu militari*) los ciudadanos dan preferencia a sus intereses individuales, incluso si van contra los sociales, tenemos el ejemplo de las ONG. Las ONG, como es sabido, son a las ciencias sociales lo que la vida es a la física. Para los físicos, todos los sistemas en el universo tienden a la entropía, pero la vida es un fenómeno que niega la entropía. Con el permiso de los físicos, se podría decir que las ONG van contra el principio entrópico de la economía y se nutren de ciudadanos que se sacrifican individualmente (o sea, se "autoperjudican") a favor del bien social.

Por el momento, y hasta que no aparezcan incentivos claros, es difícil que los editores, autores, productores, etc. de sitios web: (1) utilicen de manera responsable metadatos; (2) utilicen XML, o al menos XHTML, en lugar de HTML; (3) desarrollen o apliquen a sus sedes web o bien ontologías, o bien taxonomías o bien tesauros, según corresponda y (4) los representen, según convenga, en formato *RDF*, *Topic Map*, etc.; a menos que confiemos en el "efecto ONG".

4. Conclusiones

En el futuro de los sistemas de información hay una larga lista de innovaciones a las que merece la pena prestar atención. Señalaremos las que son más importantes en nuestra opinión por tener mayor impacto en las Ciencias de la Documentación:

1. Internet Invisible. Se ha producido un gran avance en la variedad de formatos que pueden indizar los motores de búsqueda. Por otro lado, es previsible que motores de búsqueda como *Scirus* sean solamente un ejemplo de la clase de sistemas de acceso a la información que podemos esperar en el futuro. Sin embargo, hay varios frentes en los cuales deberíamos empezar a colocar nuestras energías y esfuerzos. Por un lado, los documentos no HTML son potenciales enemigos de la hipertextualidad. Deberíamos considerar si los avances por un lado, no son retrocesos por otro. En ese caso, deberíamos considerar qué hacer, o al menos, considerar qué hacer en el terreno de la investigación y las políticas de información. Seguro que tenemos un amplio y bonito programa de investigación por ese lado. Por otro lado, las interfases de consulta de los motores de búsqueda están a años luz de las posibilidades reales y del know-how sobre el tema. Otro terreno sobre el cual, al menos, pensar y, mejor aún, actuar.

2. Web semántica. Aunque sea con mentalidad ONG, ¿qué podemos hacer a favor de la web semántica si creemos en sus beneficios a escala social aunque, por ahora, aporte escasos beneficios individuales? Al menos, los organismos vinculados al mundo de la promoción del conocimiento y la ciencia y el patrimonio cultural (universidades, archivos, bibliotecas, centros

de investigación, museos, etc.) deberían sentirse obligados por la visión de la web semántica. Por tanto, al menos a corto y medio plazo, las organizaciones vinculadas con el mundo de la ciencia, la cultura, el patrimonio, la educación, etc., debería sentirse obligadas a: (1) interesarse al menos por cosas tan aparentemente inocentes como el lenguaje XHTML en unión con las hojas de estilo (CSS) y (2) estudiar políticas de metadatos en relación a todas sus publicaciones digitales.

3. ¿Qué nos enseña la web semántica? En mi opinión, nos enseña algo que, en realidad, ya sabíamos, esto es: si tomas un conjunto de datos y los etiquetas sistemática y exhaustivamente, tienes lo más parecido a la inteligencia. Si las bases de datos exhiben un notable grado de inteligencia en comparación con la Web es porque en una base de datos, todos los datos están "etiquetados", o sea, forman parte de los valores de un campo. Cada campo, a su vez, tiene unos atributos bien definidos: es un campo de texto, o es un campo numérico, o lógico, etc. Por último, todos los datos en una base de datos están sistematizados: cada registro responde a la misma estructura, así que la mera posición (la sintaxis) genera sentido (semántica). Así que, lo que es (genialmente) nuevo en la web semántica es la idea de convertir toda la Web es la más gigantesca base de datos que la humanidad pudiera haber soñado jamás.

5. Bibliografía

ABADAL, Ernest. *Sistemas y servicios de información digital*. Gijón: Trea, 2001, 147 p.

AGUILLO, Isidro (2001). "Información científica en la web: retos y tareas para los documentalistas del siglo XXI". En: Fuentes, M.E. (dir.). *Anuario de biblioteconomía, documentación e información*. Barcelona: COBDC, 2001, p. 33-50

BERNERS-LEE, T.; HENDLER, J. ; LASSILA, O. "The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new posibilities". *Scientific American*, May 2001 (se puede consultar a través de la página web de la revista: http://www.sciam.com)

CODINA, Lluís. "Web semántica: una mirada crítica". *El profesional de la Información*, 2003

CODINA, Lluís; PALMA, María del Valle. "Web y cine: análisis comparativo de dos bases de datos para la investigación en línea". *Formats* n. 3, mayo 2001 <http://www.iua.upf.es/formats/formats3/cod_e.htm>

FUENTES, Maria Eulàlia; GONZÁLEZ QUESADA, A.; JIMÉNEZ LÓPEZ, A. (2000). "Documentación e información electrónica". En: J.A. Moreiro (coord.). *Manual de documentación informativa*. Madrid: Cátedra, 2000, p. 345-422

HÍPOLA, P.; EÍTO, R. (2000). "Edición digital: formatos y alternativas". *El profesional de la información*, v. 9, n. 10, octubre 2000, p. 4-15

GEROIMENKO, V.; CHEN, C. *Visualizing the semantic web: xml-based Internet and information visualization.* London: Springer, 2002

LÓPEZ YEPES, José (ed.) (2000). *I Congreso universitario de Ciencias de la Documentación. Teoría, historia y metodología de la Documentación en España (1975-2000), Madrid, 14-17 de noviembre de 2000*. Madrid: Universidad Complutense de Madrid. Facultad de Ciencias de la Información, 2000, 822 p.

MOREIRO, José Antonio (coord.) (2000). *Manual de documentación informativa.* Madrid: Cátedra, 2000, 458 p.

PALMA, María del Valle (1999). "Integración de la gestión documental en la administración pública: un estudio de caso". En: Fuentes, M.E. (dir.). *Anuari de biblioteconomia, documentació i informació*. Barcelona: COBDC, 1999, p. 179-212

PALMA, María del Valle (2002). "Bases de datos y servicios de información disponibles en Internet". En: *Curso de Documentación Digital* (CD-ROM). Barcelona: UPF, 2002

NUNBERG, G. (comp.) (1998). *El futuro del libro: ¿esto matará eso?*. Barcelona: Paidós, 1998, 314 p.

ROVIRA, Cristòfol (2001). "Herramientas de ayuda a la navegación". *Temes de Disseny*, n. 18, abril 2001, , p. 66-73

TRAMULLAS, Jesús; OLVERA, M. Dolores (2001). *Recuperación de la información en Internet*. Madrid: Ra-Ma, 232 p.

SHERMAN, Chris (1999). "The future of web search". *Online,* v. 23, n. 3, May/June 1999, p. 54-61,

SHERMAN, Chris (2000). "The future revisited: what's new with web search". *Online*, May 2000, <http://www.onlineinc.com/onlinemag/OL2000/sherman5.html>