Xarxa Punt TIC

MÓDULO 1 NIVEL AVANZADO
Búsqueda Y Recuperación de la Información en Internet (Avanzado)

Apuntes Completos

plan avanza2 | Gobierno de España | Plan E | Ministerio de Industria, Turismo y Comercio | Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información | Generalitat de Catalunya

➔ ÍNDICE

➔ A. Problemática documental de la información en la Web

Con la aparición de los BBS (sigla de *Bulletin Board System*), que permitía acceder rudimentariamente a datos remotos usando un módem, el acceso a documentos desde una computadora hogareña se simplificó bastante. Las bocas de producción de documentos electrónicos se multiplicaron, así como la cantidad de usuarios con computadora que intercambiaban información. Sin embargo, cada BBS actuaba de manera autónoma (década del '80 y los primeros años '90.)

Y entonces llegó la Web. Para bien y para mal. Para bien, porque Internet conectó a todas las máquinas que producían (y recababan) información, porque abarató y democratizó la producción y recolección de información y porque la cantidad de información creció (y sigue creciendo) en proporciones antes nunca imaginadas. Pero también para mal, porque entonces la información no estaba toda en el mismo formato, porque no necesariamente era verdadera, porque podía estar desactualizada o con errores.

En ese contexto aparecieron los buscadores de Internet, para tratar de ponerle un poco de sentido a todo ese inabarcable océano de información *online*. Los buscadores evolucionaron rápidamente, tratando de ayudar, cada vez mejor, a organizar los miles de millones de documentos que se producen.

Tipología. Estructura

A grandes rasgos, hay dos tipos de buscadores en Internet: *los directorios y los motores de búsqueda*.

Los directorios

Son buscadores organizados a partir de una jerarquía temática (taxonomía). El más conocido de los directorios es Yahoo en su página http://es.dir.yahoo.com y Google en las páginas http://www.google.com/dirhp?hl=ca o http://www.google.com/dirhp?hl=es. Se puede navegar por un directorio adentrándose en sus categorías y subcategorías o ingresando una palabra clave que mostrará las distintas categorías y sitios en los que esa palabra aparece. Hay directorios generalistas como Yahoo y especializados, como Ariadna, un buscador de recursos periodísticos en www.periodismo.com/buscador.

Aunque la mayoría de los buscadores imitan la taxonomía de Yahoo, no hay un estándar en este sentido. Tampoco hay ningún tipo de homogeneidad o criterio común entre los buscadores especializados. El directorio muestra los resultados de su búsqueda basándose únicamente en el título y la descripción

del sitio. Además, los directorios incluyen sitios completos, no páginas o secciones dentro de un sitio. Otra característica de los directorios es que las páginas son revisadas por seres humanos e incorporadas sólo si cumplen con los criterios de pertinencia de ese buscador. Esto hace que la cantidad de sitios de los directorios sea pequeña en comparación con la totalidad de sitios existentes.

El otro tipo de buscadores son los motores de búsqueda.

El más conocido actualmente es Google. Los motores de búsqueda no tienen una taxonomía, se puede acceder a los resultados sólo a partir de una palabra clave. Todo el proceso de indizado de páginas es automático, no hay seres humanos revisando sitio por sitio. A diferencia de los directorios, los motores de búsqueda buscan en toda la página de un sitio web, no se limitan al título y la descripción. Si bien no indizan todas las páginas de un sitio, tampoco se limitan a indexar una sóla página. La cantidad de páginas indizadas es enorme: al momento de escribir esto, Google tenía indizadas más de 8.000.000.000 páginas web. Los motores de búsqueda tienen robots buscadores que exploran los sitios web y los incorporan a sus bases de datos. A esta acción se la llama indizar o indexar (la Real Academia acepta los dos términos).

En la actualidad muchos buscadores combinan la potencia de los motores de búsqueda con la lógica de los directorios. Si Yahoo! no encuentra resultados en su directorio, muestra los resultados de su motor de búsqueda. Pero también los motores de búsqueda recurren a los directorios: para quién necesite resultados más ordenados, Google usa los datos del directorio Dmoz, también conocido como *Open Directory*. (http://dmoz.org/).

Cómo funcionan

Robots

En una escala microscópica los robots del siglo XXI son invisibles e inmateriales. Estos robots se dedican a hacer algo clave: indexar las páginas que se visitan. (Piscitelli, 2005).

Hoy la situación es muy diferente de hace unos cuantos años atrás. En ese momento, la fe en los robots buscadores, hacía suponer que si buscadores de aquella época como Altavista o Hotbot no encontraban aquello que se buscaba, era sencillamente porque tal información no existía en la red.

Sin embargo, se empezó a dar importancia a la calidad por encima de la cantidad. Esto determinó que era preferible indexar sitios de calidad, antes que apilar meramente la mayor cantidad de sitios posibles, acudiendo a los buscadores. El universo Web estaba lleno de páginas que no valía la pena visitar nunca.

En ese momento interesaba el área del aprendizaje robótico y se construyó un robot llamado Inquirus, capaz de interrogar a otros robots acerca de la existencia de documentos que cumplieran con cierta estructura de búsqueda; este robot podía arrojar un beneficio secundario más valioso que el buscado

originalmente estimando el tamaño real de la red, un número que en ese momento nadie conocía a ciencia cierta. Entre los resultados que cosechó el Inquirus aplicado al buscador Hotbot, fue descubrir, en 1997, que la Web contaba con alrededor de 320 millones de documentos (por lo menos el doble de lo que se creía entonces) .Y no solo eso, Hotbot se preciaba de ser el más exitoso y exigente de los robots en aquella época pero, de pronto, se vio devaluado al descubrirse que sólo indexaba el 34% de toda la Web. Como premio de consuelo, pudo jactarse de que a los otros robots les iba aún peor: Altavista solo cubría un 28%, y otros buscadores —como Lycos, que pronto caería en manos de Terra y Telefónica— apenas cubrían un 2% de la red.

En febrero de 1999, cuando se repitió el mismo ejercicio, los investigadores encontraron que la red había crecido (tenía 800 millones de documentos) pero la capacidad de indexación de sitios a manos de los robots había empeorado.

Un excelente buscador de la época, Northern Light, ocupó entonces la *pole position* cubriendo el 16% de la Web, pero Altavista había bajado al 15% y Hotbot reseñaba apenas el 11% de las páginas existentes. Mientras tanto Google, que era un benjamín entre los pesos pesados, apenas veía en ese entonces un 7,8% de las páginas estimadas. En junio de 2001 Google cubrió por primera vez 1.000 millones de documentos, seguido de cerca por Alltheweb. Hoy, Google está cerca de alcanzar los 9.000 millones de documentos.

Por muy impresionante que sea la capacidad de indexación de los motores, la Web crece infinitamente más rápido que la posibilidad de éstos de analizarla. Además, existe la Web profunda que es por lo menos 550 veces más grande que la que los robots alcanzan, por lo que la asimetría entre lo visible y lo existente se ahonda muchísimo más.

En el año 2000, 6 de cada 10 páginas no habían sido visitadas nunca. Hoy los resultados deben alcanzar cifras de entre 8 y 9 de cada 10.

Los buscadores en su rol de gatekeeper

La noción de *gatekeeping* (cuidado de la puerta o del acceso) investiga la manera irregular en que las informaciones circulan y se encuentran sometidas a instancias que las demoran o “traban” en algún punto de la cadena comunicacional, y la fluidez con que circulan luego aquellas que consiguen pasar la barrera. Estos lugares de demora o nudos que actúan como barrera y filtro en la circulación de la información serían los *gatekeepers* o porteros (“arquero” en el fútbol).

El concepto de *gatekeeper* fue introducido por el psicólogo Kurt Lewin en 1947 trabajando en dinámica de grupos, y observó que la información circulaba de una manera muy irregular, ya que por momentos podía interrumpirse debido a los nudos o fluir de manera muy amplia después de superarlos.

Hay que imaginar al buscador como a un *gatekeeper*: dentro del universo de todas las páginas de la Web, el buscador tiene el poder de orientar en el camino hacia la búsqueda de la información.

Ya se sabe que los buscadores no indexan todas las páginas de la Web. Ahí ya hay una primera selección. El *gatekeeper* buscador deja afuera de sus resultados una enorme cantidad de contenido. La segunda selección está en la relevancia: el buscador define que determinadas páginas son más importantes que otras. Y que este criterio, aunque automatizado, es subjetivo (valga la paradoja) se comprueba fácilmente comparando los resultados de los distintos buscadores.

¿Cómo desafiar estos criterios? El segundo criterio es más fácil de burlar: utilizando diversos buscadores y directorios se puede llegar a una "intersubjetividad de resultados". Hay metabuscadores como kartoo, turbo10, webcrawler, dogpile, clusty entre otros, por ejemplo clusty muestra al usuario las mejores posiciones en que figura cada página en los distintos buscadores.

Desafiar la lógica del buscador en cuanto a los contenidos que deja afuera de sus resultados, lleva a internarse en la llamada Internet Invisible.

Cómo buscar

El primer paso en una búsqueda es saber qué es lo que se busca. No necesariamente hay que saberlo con precisión. Puede interesar, puntualmente, encontrar la bandera de Rusia o, más vagamente, hallar legislación sobre jubilación privada en América Latina.

Después de conceptualizar lo que se va a buscar, se sabrá hacia dónde ir. Si la búsqueda es más específica, se empezará con un motor de búsqueda que conduzca hacia el sitio que se necesita. Si es más general, sería conveniente empezar por un directorio que agrupe a todos los sitios comunes al tema que se investiga.

Si lo que se necesita es local o regional, habrá que restringir las búsquedas a esos países o regiones o, mejor, consultar buscadores de la zona en cuestión.

En el mismo sentido, si la temática es específica, habría que partir de un buscador generalista, buscar allí un directorio temático y realizar en el directorio temático una segunda búsqueda, más acotada.

Cada buscador tiene sus reglas (sintaxis) y por eso es recomendable leer la documentación y páginas de ayuda para entender bien sus opciones de búsqueda.

La mayoría de los buscadores aceptan los operadores booleanos. Estos son AND, OR y AND NOT (esto último en algunos buscadores funciona al poner solamente NOT o colocando el signo -).

Por defecto, la mayoría de los buscadores funciona con el operador AND o +. Esto quiere decir que poner en un buscador las palabras mapa argentina o mapa AND argentina o mapa + argentina es equivalente. Esta búsqueda traerá sólo los documentos que contengan ambas palabras.

Colocando la palabra OR, mostrará los documentos que contengan al menos una de esas palabras. Por ejemplo, si se pone argentina OR uruguay mostrará las páginas que contengan la palabra argentina, las páginas que contengan la palabra uruguay (y en consecuencia también las que contengan ambas

palabras). El operador OR es también útil si no se sabe cómo se escribe una palabra (volswagen OR volkswagen): traerá documentos que contengan al menos una de las grafías.

El operador NOT o el signo - excluye palabras de la página de resultados. “Cindy Crawford” -sex -porn -adult -xxx -nude traerá documentos que mencionen a la supermodelo, pero no contenido pornográfico. A esto se llama filtrar o refinar una búsqueda. También puede usarse si, por ejemplo, se quiere información solamente sobre Windows XP pero no Windows Vista, habrá que poner windows +XP –Vista.

Las comillas (como en el ejemplo anterior) sirven para denotar una frase exacta: términos que uno sabe que deben ir juntos como el título de un libro o de una película o de una canción o de un juego. Hay que tener la certeza que se escribe de ese modo, porque si no, ignorará el pedido.

Los operadores AND y OR deben escribrse con mayúsculas.

Todos estos operadores pueden ser muy útiles usándose en forma combinada. Así, por ejemplo, si se quiere encontrar todas las páginas en las que aparezca mencionado Estados Unidos en su forma catalana, excluyendo la grafía en inglés, se podría poner: **“Estats Units” OR EE.UU. OR EEUU -“United States” -USA**.

Buscadores y directorios

- Gigablast Inc. Gigablast <http://www.gigablast.com/>
- Periodismo.com. Ariadna <http://www.periodismo.com/buscador/>
- Google. Google <http://www.google.cat>
- Grub Buscador <http://www.grub.org>
- IAC Search & Media. Ask.com <http://www.ask.com/>
- IAC Search & Media. Excite <http://www.excite.com/>
- LookSmart, Ltd. Wisenut <http://www.wisenut.com/>
- Lycos, Inc. Hotbot <http://www.hotbot.com/>
- Lycos, Inc. Lycos search <http://www.lycos.com/>
- Microsoft. MSN <http://www.msn.com>
- Overture Services, Inc. Alltheweb, find it all <http://www.alltheweb.com/>
- Overture Services, Inc. Altavista <http://www.altavista.com/>
- The New York Times Company. About <http://www.about.com/>
- Walt Disney Internet Group (WDIG). Go.com <http://go.com/>
- WebFile.com. Webfile <http://www.webfile.com/>
- Yahoo! Inc. Yahoo! <http://www.yahoo.com./>

Metabuscadores

- Copernic Technologies, Inc. Copernic <http://www.copernic.com/>

- Digital Tsunami, Inc. Quickfindit <http://www.quickfindit.com/>
- Ez2find.com. ez2finf <http://ez2find.com/>
- InfoSpace, Inc. Dogpile <http://www.dogpile.com/>
- InfoSpace, Inc. Metacrawler <http://www.metacrawler.com/>
- Intelliseek, Inc. ProFusion <http://www.profusion.com/index.htm>
- Mamma, Inc. Mamma <http://www.mamma.com/>
- Surfboard BV. Ixquick <http://www.ixquick.com/>
- Netscape Communications Corporation. DMOZ Open Directory Project <http://dmoz.org>

Buscadores de buscadores

- Multibuscador.com <http://dir.multibuscador.com/>
- Buscopio <http://www.buscopio.net/esp/>

Herramientas de segunda generación: clasificación documental

Nos encontramos con un conjunto totalmente nuevo de herramientas, diferenciadas de las anteriores porque son *client-side*. Se trata, por tanto, de programas totalmente independientes que se instalan en el ordenador cliente, lo que redunda en un mayor control y personalización de sus funciones. El hecho de que, a veces, algunas de estas herramientas pueden funcionar de forma autónoma respecto al cliente en el que estén instaladas ha llevado a que incorrectamente se generalice el nombre de agente o bot, que podría identificar sólo a alguna de ellas, no a todas.

En general, el conjunto resulta relativamente heterogéneo lo cual permite construir una clasificación muy descriptiva.

Además, puesto que algunos de los mecanismos son paralelos a los que existen como servidores, dicha segregación resulta especialmente útil y admite análisis comparativos de prestaciones. Sin embargo, el valor añadido de alguno de ellos no se restringe únicamente a un incremento de la capacidad de automatización de tareas y personalización, sino que ofrecen posibilidades totalmente novedosas. Algunas de opciones inéditas resultan imposibles de implementar desde un servidor.

Entre las novedades más singulares destacaremos:

- La posibilidad de extraer información de la Internet invisible (infranet), el conjunto de registros de bases de datos o catálogos de biblioteca accesibles mediante formularios web, pero que no son indexadas por los motores.

- El uso de los verdaderos agentes, que de forma autónoma, mediante mecanismos inteligentes pueden recorrer la red, extraer información e incluso “aprender” con ayuda del operador humano. La mayoría de los programas revisados son productos comerciales disponibles bajo el sistema *Shareware* (evaluar antes de adquirir), lo que significa que puede obtenerse una copia de los mismos, más o menos operativa, de la red Internet. El precio no es excesivamente caro y son, precisamente, los programas más sofisticados los de mayor coste. Lamentablemente, para este tipo de programas apenas se ofrece soporte técnico y no es infrecuente que algunos títulos desaparezcan con gran rapidez.

A continuación presentamos una clasificación comentada de las citadas herramientas utilizando como criterio sistematizador las potencialidades y aplicaciones documentales de las mismas. Este criterio excluye otros programas, relativamente numerosos en la actualidad, a veces reunidos bajo la categoría de “utilidades de Internet” que son potencialmente interesantes. El interés de los mismos, fundamentalmente informático, puede ser más evidente en un futuro no muy lejano. Atendiendo a los usos documentales distinguimos cinco grandes grupos, por orden de complejidad:

- Clientes Z39.50
- Volcadores
- Metabuscadores
- Indizadores
- Mapeadores

Caso aparte lo constituyen las herramientas canalizadoras, que tienen un carácter mixto. Basadas en la tecnología *push* podríamos calificarlas de híbridas, al requerir tanto de una instalación cliente como de un servidor.

La incorporación de este tipo de servicios a los clientes universales (Netscape y Explorer) nos ha llevado finalmente a excluir estas interesantes herramientas de nuestra clasificación, donde previamente las considerábamos volcadores sofisticados. Se puede estar al corriente de las principales novedades de este tipo de programas visitando periódicamente alguno de los principales depósitos de *software* en la Internet.

➔ B. La web opaca

Páginas que pueden ser indexadas, pero no son incluidas en los buscadores. Los motivos para que los buscadores “decidan” no incluirlas pueden ser:

Profundidad de exploración: los sitios tienen “profundidad”. La *home page* o página principal es el primer nivel; ahí llegan todos los buscadores. De ahí se lincan a páginas internas del sitio, ése sería el segundo nivel, al que no llegan los directorios y algunos motores de búsqueda. Esas páginas, a su vez, enlazan con páginas “más” internas, que no estaban en la *home page*. A este nivel llegan muy pocos buscadores. Cuanto más profundo sea el nivel, menos buscadores lo indexarán.

Frecuencia de exploración: un sitio puede cambiar todos los días, pero muchos de los robots de los buscadores que exploran los sitios los visitan una vez por mes, o menos, por una cuestión de costos. Todos los cambios entre una visita y otra no figuran en los buscadores.

Supera el número máximo de resultados: cada buscador define qué cantidad de páginas de un sitio mostrará. Si un sitio tiene más páginas que las que el buscador incluye, las restantes quedarán sin indexar.

Errores de exploración: puede haber un problema en el sitio o en el robot del buscador (o en la compatibilidad entre ambos) que impida que una página (o hasta un sitio completo) sea incluida en la base de datos del buscador.

Hoy en día la mayoría de los accesos a sitios web se realiza a través de motores de búsqueda, por lo tanto, es fundamental para sus responsables asegurarse de aparecer bien posicionados en los resultados de búsqueda, tanto desde el punto de vista del márqueting como para dar mejor servicio a sus usuarios.

En consecuencia, el lugar que un sitio web ocupa en el listado de resultados de un motor de búsqueda, cuando un usuario realiza una consulta en el buscador, es un aspecto de gran relevancia para los responsables de sitios web, y las acciones para su mejora vienen explicadas por el “posicionamiento web” (Codina, 2004; Codina, Marcos 2005; Arbildi, 2005). El eje central está en preguntarse cuáles serán las palabras que previsiblemente utilizarán los potenciales usuarios en sus búsquedas; una vez determinado esto, podrá hacerse uso "lícito" de las técnicas de optimización para que un determinado sitio web aparezca en una buena posición cuando los usuarios busquen información relacionada con los contenidos de dicho sitio web.

Los sitios web que albergan bases de datos terminológicas también pueden –y deben– hacer uso de las técnicas de mejora del posicionamiento para facilitar a sus posibles usuarios encontrar estos recursos a través de los buscadores. Nuestra investigación ha tomado como muestra diez bases de datos terminológicas cuya consulta es libre en la Web, presentan multilingüismo y pertenecen a diferentes temáticas (tabla 1).

Bases de datos	*URL*
CercaTerm(català)	http://www.termcat.net

Eurodicautom	http://europa.eu.int/eurodicautom
EuskalTerm	http://www1.euskadi.net/euskalterm
OncoTerm	http://www.ugr.es/~oncoterm/
TerminoBanque	http://www.cfwb.be/franca/bd/bd.htm
TIS: Terminological Information System	http://tis.consilium.eu.int
UBTerm	http://www.ub.edu/slc/ubterm
UNTerm	http://unterm.un.org
WTOTerm	http://wtoterm.wto.org

Tabla 1. Bases de datos terminológicas estudiadas

Metodología de la investigación

Para facilitar ser encontrado en buscadores, los sitios web deben mejorar todos los aspectos que estos tienen en cuenta a la hora de establecer el *ranking* de resultados. Sin entrar a explicar en detalle las técnicas para la optimización de sitios web, nombramos los criterios que parecen estar usando los buscadores en sus *rankings* de resultados (Codina; Marcos, 2005):

- Frecuencia (absoluta y relativa) de la expresión o término buscado en la página web (al que llamaremos "palabra clave"), siempre que no se caiga en una repetición abusiva que sea considerada *spam* por los buscadores. En este estudio se han planteado tres posibles palabras clave para cada sitio web.
- Posición: lugar donde se encuentra el término dentro de la página; se tendrá en cuenta que los metadatos y el primer párrafo tienen más peso que otras partes de la página.
- Metadatos: además de los metadatos de la sección *head* (principalmente *title*, *description* y *keywords*), otras etiquetas también proporcionan información descriptiva de utilidad para los buscadores, por ejemplo el *title* de los enlaces y de las imágenes, así como el texto alternativo (alt) de las imágenes.
- Popularidad: número de enlaces externos que recibe la página web. Este criterio está relacionado con el *PageRank* determinado por la barra de herramientas de Google.
- Anclaje: texto que sirve como enlace para llegar a esta página web.
- Tráfico de visitas que recibe la página web, considerando a la vez el tiempo de permanencia de los usuarios en ella. Viene dado por el *TrafficRank* de la barra de herramientas de Alexa.

De estos factores, nuestra investigación está centrada fundamentalmente en los metadatos y la popularidad. Los demás factores también se han estudiado,

si bien no han arrojado resultados que nos ayuden a estimar el grado de importancia que tienen para el posicionamiento web de los sitios analizados.

El estudio ha tenido en cuenta dos tipos de análisis: en primer lugar se ha llevado a cabo un análisis empírico-descriptivo atendiendo a los aspectos que la bibliografía sobre posicionamiento web indica que deben tenerse en cuenta. En base a los resultados obtenidos en esta fase, se han planteado algunas hipótesis que se han puesto a prueba mediante el análisis estadístico ANOVA; se trata de un análisis multivariable en el que una variable dependiente es cruzada con algunas variables independientes.

Para su aplicación se ha utilizado el programa Statgraphics Plus 5.0. La variable dependiente considerada ha sido el posicionamiento web de los sitios web, y las variables independientes con las que se ha cruzado han sido cada uno de los sitios web estudiados, los buscadores sobre los que se han hecho las consultas, las palabras clave de las consultas y los metadatos *title*, *description* y *keywords* . El cruce de datos da como resultado el p-valor, que es el nivel de significación empírico en un contraste de hipótesis. Se considera que una hipótesis es nula en los casos en que el p-valor es inferior a 0'05 y que es alternativa siempre que supera esta cifra.

Elección de palabras clave para el posicionamiento web

Antes de analizar las causas del posicionamiento en estos sitios web, y puesto que este dato constituye la variable dependiente en el análisis estadístico, se ha comprobado cuál es éste para 3 consultas para las que estos sitios web consideramos que deberían aparecer bien posicionados, entendiendo por una buena posición los 10 primeros resultados de los buscadores más utilizados hoy en día. Se ha valorado de forma especialmente positiva (2 puntos) que el sitio web apareciera en el primer o el segundo puesto, de forma parcial (1 punto) si aparece entre el puesto 3 y el 10, y no se ha valorado (0 puntos) si está más allá del resultado número 10.

Las palabras de búsqueda elegidas se han traducido al idioma principal de la interfaz de cada herramienta terminológica siguiendo estos criterios:

Palabra clave 1 (PC1): nombre de la base de datos.

Palabra clave 2 (PC2): sintagma que describe el tipo de recurso terminológico del que se trata en cada caso.

Palabra clave 3 (PC3): sintagma que describe una cualidad específica del recurso terminológico del que se trata en cada caso.

PC1	PC2	PC3
CercaTerm	Base de dades terminològica	Terminologia catalana

WTOTerm	Terminology database	World trade terminology
TerminoBanque	Banque de donées	Terminologie spécialisée
Eurodicautom	Terminology database	Multilingual specialized terminology
EuskalTerm	Banco terminológico	Terminología euskera
TIS	Terminological information system	Consilium terminological database
UBTerm	Base de dades terminològica	Terminologia catalana
UNTerm	Terminology database	United Nations terminology
OncoTerm	Base terminológica de oncología	Terminología oncológica

Tabla 2. Palabras clave utilizadas para buscar en los 6 motores.

En el mes de enero de 2006 se realizaron las tres búsquedas en 6 de los buscadores más utilizados en el 2006: Google, Yahoo! Search, MSN Search, Altavista, Teoma y Vivísimo o Clusty. El resultado obtenido es muy diferente para cada una de las palabras clave (tabla 3):

Las búsquedas por la PC1 obtienen con más frecuencia el sitio web esperado entre los 10 primeros resultados, en especial con Yahoo! Search y Google, aunque no tanto con Teoma y Vivísimo, que sólo consiguen localizar estos sitios web en un 50% de las búsquedas.

El resultado cambia para la PC2, pues dos sitios web no se han encontrado entre los 10 primeros resultados, y otros no ocupan las primeras posiciones. Google y Yahoo! Search muestran estos sitios web en mejores posiciones que los otros buscadores. Teoma en cambio sólo muestra en este primer grupo de resultados 4 de los 10 sitios web.

La búsqueda por la PC3 muestra resultados diferentes a los anteriores: aunque sigue en la línea de la PC2 y algunos buscadores no colocan estos sitios web entre los 10 primeros resultados, el buscador Teoma esta vez ofrece las mejores posiciones, seguido de Vivísimo y Google.

Bases de datos	PC1	PC2	PC3	Promedio de cada base de datos
Eurodicautom	100,0	50,0	100,0	83,3

TIS	33,0	91,6	100,0	74,8
WTOTerm	100,0	25,0	91,6	72,2
EuskalTerm	100,0	83,3	0,0	61,1
CercaTerm	91,6	0,0	50,0	47,2
OncoTerm	41,6	100,0	0,0	47,2
UNTerm	91,6	25,0	16,6	44,4
UBTerm	75,0	33,0	0,0	36,0
TerminoBanque	100,0	0,0	0,0	33,3
Promedio de cada PC	72,3	50,8	45,8	83,3

Tabla 3. Clasificación de las bases de datos en función del promedio del posicionamiento que ocupa cada sitio web al buscar por la PC1, PC2 y PC3. El valor máximo, 100, indica que se encuentra en el primer o el segundo puesto; los valores intermedios, cercanos a 50, indican que se encuentra entre el puesto 3 y el 10, y el valor 0 indica que no está entre los 10 primeros resultados

El análisis estadístico corrobora que existe una diferencia significativa entre el posicionamiento de los sitios web estudiados y las palabras clave para las que se ha probado su posicionamiento, pues para la PC1 el posicionamiento obtenido es mucho mejor que para las PC2 y PC3 (figura 1). Al mismo tiempo, muestra que el uso de un buscador u otro no provoca diferencias significativas en los resultados de posicionamiento, si bien Google, Yahoo! Search y Vivísimo presentan estos sitios web mejor posicionados que los otros 3 buscadores utilizados.

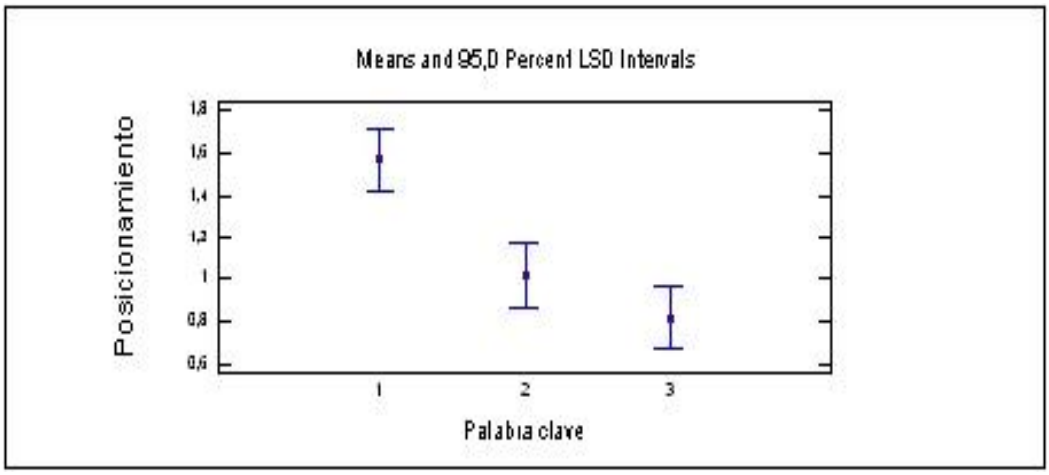


Figura 1. Posicionamiento de los sitios web en relación con PC1, PC2 y PC3.

Análisis de los factores del posicionamiento web

A partir de los aspectos indicados en el apartado 2, presentamos los resultados más relevantes obtenidos en el análisis empírico-descriptivo y en el análisis estadístico.

Frecuencia de aparición y posición de las palabras clave

A pesar de que estamos de acuerdo en la importancia de estos factores para mejorar el posicionamiento web, se ha decidido no tenerlo en consideración para este estudio debido a dos motivos: en primer lugar, las *interfaces* de las herramientas de búsqueda de información terminológica no cuentan con un volumen de texto suficiente para poder determinar valores de frecuencia óptimos; y en segundo lugar, a excepción del nombre de la base de datos, las otras palabras clave establecidas como consultas (PC2 y PC3) no suelen aparecer reflejadas en los sistemas estudiados, por lo que este criterio no nos permitiría establecer comparaciones.

Metadatos

En lo referente a los metadatos de los sitios web estudiados, hemos comprobado que el 90% poseen el campo *title* relleno, el 40% incluyen el campo *keywords* y sólo el 20% presenta el campo *description* (tabla 4). No se han considerado las etiquetas *title* y el *alt* de los enlaces y las imágenes, pues en los sitios estudiados su número era muy bajo o incluso nulo, por lo tanto no afectaría al posicionamiento web de estos sitios.

Bases de datos	Title	Keywords	Description
Base de Terminologie	X	-	-
CercaTerm	X	X	-
Eurodicautom	X	X	-
EuskalTerm	X	X	X
OncoTerm	X	X	-
TerminoBanque	X	-	-
TIS	X	-	X
UBTerm	X	-	-
UNTerm	-	-	-
WTOTerm	X	-	-

Tabla 4. Uso de metadatos en los sitios web de las bases de datos estudiadas.

En el análisis estadístico, el cruce del posicionamiento de los sitios web con la información de metadatos pone de manifiesto que sólo el campo *description* implica diferencias significativas de posicionamiento entre sitios web (figura 2), mientras que la existencia de *keywords* no influye tanto en los resultados obtenidos (figura 3). No se ha podido obtener un resultado fiable sobre la repercusión de la etiqueta *title*, pues 9 de los 10 sitios web contaban con ella, lo que no nos deja margen para establecer comparaciones.

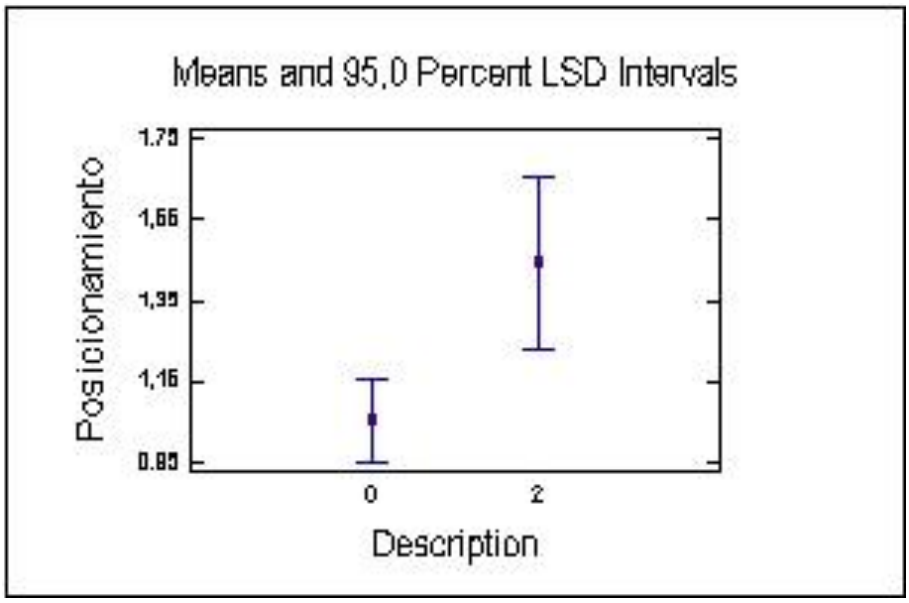


Figura 2. Posicionamiento de los sitios web considerando si tienen la etiqueta meta *description* (2) y si no la tienen (0).

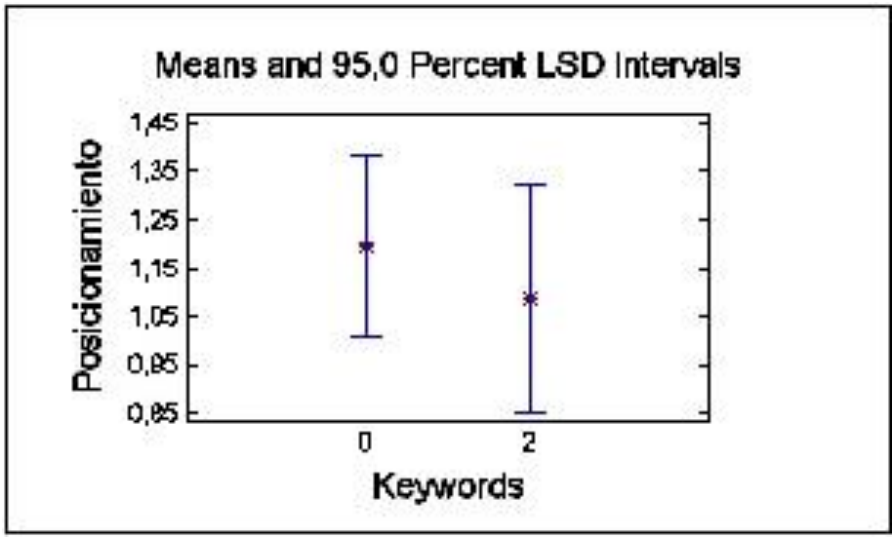


Figura 3. Posicionamiento de los sitios web considerando si tienen la etiqueta meta *keywords* (2) y si no la tienen (2).

Popularidad, textos de anclaje y tráfico de visitas

La popularidad, entendida como el número de enlaces que apuntan hacia un sitio web, puede conocerse parcialmente a través de la búsqueda con el limitador *link:* que ofrecen algunos buscadores. Mostramos los valores que da Yahoo! Search (tabla 5), que son una cifra mucho más alta que la que proporciona Google.

Bases de datos	Enlaces de llegada (Google)
Eurodicautom	2.340
CercaTerm	1.290
Base de Terminologie	1.030
TIS	458
EuskalTerm	243
TerminoBanque	163
WTOTerm	104
UNTerm	79
UBTerm	49
OncoTerm	9

Tabla 5. Clasificación de las bases de datos en función del número de enlaces que apuntan a cada sitio web según Yahoo! Search.

Si comparamos los resultados que obtenemos en las tablas 3 y 5, podemos ver cómo existe una relación bastante directa entre el número de enlaces de llegada de cada sitio web y los resultados medios que han dado para el posicionamiento por las tres palabras clave. Las bases de datos con más citas son al mismo tiempo las mejor posicionadas para estas palabras. En el extremo contrario, UBTerm y UNTerm, que obtienen menos citas también aparecen peor posicionadas. El caso de CercaTerm, que no está tan bien posicionada y en cambio recibe muchas citas, se debe a que la cifra del número de enlaces de llegada que tiene no se refiere a la página principal de la base de datos, sino a la de su institución (TermCat), pues para llegar a la base de datos es necesario pasar por un formulario en la página de inicio de la institución, lo que dificulta el acceso directo a los buscadores. En el caso de este sistema, si no cambia el modo de acceso a la base de datos, deberá optimizarse la página de inicio de la institución para mejorar el posicionamiento en las búsquedas relacionadas con su herramienta terminológica.

En cuanto a los textos usados en los anclajes de esos enlaces, se han revisado los 5 primeros de la lista que ofrece Google y el resultado obtenido ha sido que, a excepción de la PC1 (el nombre de la base de datos), no se han utilizado las palabras clave escogidas en este estudio. Por lo tanto es un criterio que no nos va a servir para establecer comparaciones y detectar la relevancia que tiene en el posicionamiento de sitios web.

También se anotó el valor que otorga Google como *PageRank*, si bien finalmente no lo hemos considerado de interés para este estudio pues se trata de un valor obtenido con un cálculo poco transparente. El mismo criterio nos ha llevado a no tener en cuenta el valor que da Alexa como *TrafficRank*, que

además no es específico para la página web que se analiza sino que toma el valor de su página de inicio.

Conclusiones

Aunque las técnicas de optimización para buscadores suelen enmarcarse dentro del márqueting; nuestra visión de este concepto es amplia, pues no sólo creemos que se trata de una cuestión de publicidad, sino de servicio a los usuarios. En el caso de servicios de uso gratuito, como es el de las bases de datos terminológicas que estudiamos, también deberán preocuparse de poner los medios necesarios (y lícitos) para posicionarse bien, pues es el primer paso para poder ser usado.

En cuanto a la metodología empleada, el análisis empírico-descriptivo ha resultado de utilidad para comprobar que ninguno de los sitios web estudiados ha realizado una campaña de posicionamiento. De hecho, en la mayoría de los casos no se han rellenado los campos de metadatos. De los tres campos básicos que hemos mencionado, *title* es el más frecuente (está en el 90% de los sitios web analizados), seguido de *keywords* (40%) y, de forma muy lejana, por el de *description* (20%). No se ha seguido una política de enlaces para conseguir mayor visibilidad, tanto para los posibles usuarios como para los buscadores. Por otro lado, el análisis estadístico multivariable (ANOVA) ha resultado válido para comprobar las observaciones realizadas y verificar hipótesis. La aplicación conjunta de ambas metodologías nos lleva a afirmar que el metadato *description* juega un papel importante en el posicionamiento de sitios web, mientras que el metadato *keywords* no presenta esta evidencia (figura 4). De todas formas, podemos afirmar que el uso de metadatos no es decisivo para el posicionamiento web, pues se da el caso de sitios bien posicionados que no han hecho uso de metadatos, por lo que no hay que dejar de lado otros factores que influyen en el posicionamiento, como la popularidad.

De hecho, se ha podido observar que existe una correlación entre la popularidad de los sitios web analizados (el número de enlaces que apuntan hacia ellos) y su posicionamiento para las palabras clave escogidas, lo que indica que el criterio de popularidad juega un papel importante en la posición que los sitios web ocupan en los resultados de las búsquedas en motores. Esta popularidad debería venir potenciada por unos textos de anclaje apropiados, de manera que se afiancen las palabras clave que las instituciones escojan para su posicionamiento en la web. En el caso de los sitios estudiados no se ha dado esta política.

En lo que se refiere a los buscadores usados en el estudio, Google es el que ha presentado un comportamiento más regular en relación con el uso de los metadatos en los sitios web, y Google, Yahoo! Search y Vivísimo son los 3 motores que han mostrado mejores posicionamientos para las palabras clave escogidas en estos 10 sitios web. En la mayoría de los buscadores los sitios web aparecen posicionados cuando se realizan consultas por palabras más específicas (PC1), a excepción de Teoma y Vivísimo, que contrariamente a lo

dicho, los sitios web obtienen mejores posicionamientos en búsquedas más generales (PC2 y PC3). Por ejemplo, en el caso de Eurodicautom, la búsqueda por la PC3 no presenta este sitio web entre los primeros resultados de Yahoo! y, en cambio, para Teoma ocupa el segundo puesto.

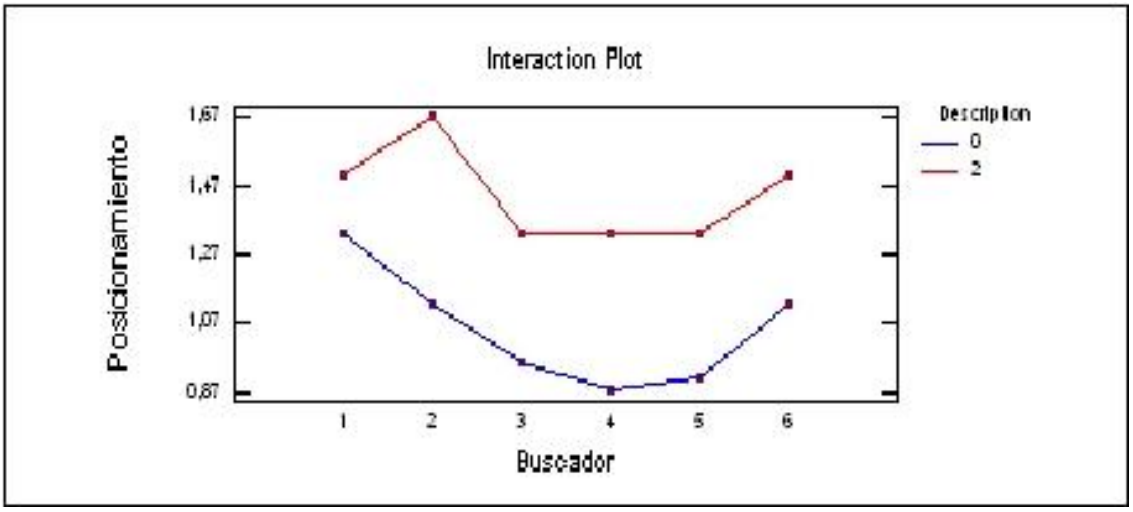


Figura 4. Posicionamiento de los sitios web en 6 buscadores considerando si tienen la etiqueta meta *description* (línea roja) o no la tienen (línea azul): Google (1), Yahoo! Search (2), MSN Search (3), Altavista (4), Teoma (5) y Vivísimo (6).

Taller práctico de indexación

SEO es la tarea de conseguir aparecer en los primeros resultados de los buscadores para determinadas búsquedas, sigla en inglés *de **S**earch **E**ngine **O**ptimization,* optimización para motores de búsqueda.

Para ello se aplican distintas técnicas y estrategias que pueden ser muy diversas. Según estas técnicas podemos clasificar SEO en dos grandes grupos: SEO Black Hat (sombrero negro) y SEO White Hat (sombrero blanco).

Aunque los objetivos son los mismos para ambas, salir en las primeras posiciones de un buscador, sus técnicas son muy distintas. Si bien el *SEO Black Hat* intentará por todos los medios conocidos lograr el posicionamiento, el *SEO White Hat* intentará lo mismo pero de otra manera mucho más sutil y natural para no poner el *website* en peligro. Sobre esto profundizaré mucho más en otros artículos.

Hoy en día el trabajo de un **SEO** es tan amplio y variado que abarca desde sólidos conocimientos de programación hasta la psicología para saber como van a reaccionar los usuarios. Por ejemplo, dependiendo del título que salga en los *serps* de Google, entrarán o no. Un SEO nunca dejará de aprender o dejará de ser SEO, ya que los buscadores intentan evitar la manipulación de los

resultados por parte de éstos y corrigen su algoritmo en base a lo que ven hacer a los SEOs.

Aquí te presentamos un pequeño manual de posicionamiento web en Google. Con él podrás conseguir aparecer en las primeras posiciones de los resultados. Recuerda que solamente podrás ser la primera posición si te esfuerzas mucho.

Tener buenos contenidos es lo fundamental para el éxito seguro de un sitio web.

Por una parte, vas a conseguir atraer a un gran número de visitantes que accederán a tus páginas regularmente.

Por otra parte, si sabes cómo redactar estos contenidos, podrás incluso atraer más visitas gracias a Google. Intenta redactar escogiendo determinadas palabras clave (*keywords*), y aprende dónde situarlas dentro de cada página web.

¿Por qué necesito buenos contenidos?

Los contenidos son lo primordial en un sitio web. Podrás saber todos los trucos y podrás conseguir engañar a Google, pero como realmente vas a conseguir visitas es con unos buenos contenidos.

Además, si los contenidos realmente merecen la pena vas a conseguir más enlaces de los *webmasters* de otros sitios web. Como veremos más adelante, tener muchos enlaces es fundamental para tener un buen posicionamiento en Google.

No dejes de generar contenidos e intenta construir páginas regularmente, con buena información.

¿Debo actualizar constantemente los contenidos?

Es una buena idea actualizar periódicamente los contenidos de tu sitio web por dos motivos:

- A Google le gustan los sitios que renuevan y actualizan sus contenidos. Estima que son sitios “vivos” y que se puede contar con ellos.
- Puedes conseguir que el robot Freshbot pase regularmente por tu sitio web. Este robot pasa por las páginas con los contenidos más “frescos” y actualiza sus contenidos en la Base de Datos de Google al cabo de unas horas. De esta manera, puedes modificar rápidamente los contenidos de tu sitio web (por ejemplo, con un nuevo producto, o nuevas palabras claves), estando seguro que va a aparecer en Google en un par de días.

Con las palabras con las cuales quieres aparecer en la primera posición de los resultados de Google cuando se busca por ellas. Por ejemplo, "coches usados", "abogados en caracas" o "sms gratis".

Planea con antelación cada página web y destina 2 o 3 palabras claves (*keywords*) por página. Es decir, no intentes que la misma página web aparezca en las primeras posiciones de Google buscando por muchas palabras. Será muy difícil conseguirlo.

¿Qué palabras claves (keywords) utilizo?

Puede que tengas un sitio web dedicado al negocio de coches, pero no sepas qué palabras clave escoger.

Una herramienta muy útil nos la proporciona Google en el KeywordSandbox. En realidad se trata de una ayuda para escoger palabras en el programa AdWords, pero nos puede ayudar mucho para escoger nuestras palabras claves.

Por ejemplo, al introducir la palabra "coches", esta herramienta nos sugiere "alquiler de coches", "coches usados" o "coches nuevos", aparte de otras muchas más. A partir de esto, deberíamos plantearnos una estrategia con las palabras clave (*keywords*).

Otras herramientas que sugieren palabras clave:

- http://es.espotting.com/popups/keywordgenbox.asp
- http://inventory.overture.com/d/searchinventory/suggestion/
- http://www.7search.com/scripts/advertiser/sample_get.asp

Tampoco conviene olvidar las páginas que hacen una clasificación de las palabras más buscadas o más populares, como el Zeitgeist de Google. Te podrán servir de ayuda para nuevas palabras claves.

Otras páginas que muestran las palabras más buscadas:

- http://sp.ask.com/docs/about/jeevesiq.html
- http://50.lycos.com/
- http://buzz.yahoo.com/

TITLE: Probablemente el lugar más importante. Intenta que en el título de la página web aparezcan las palabras claves deseadas. Además, haz un esfuerzo para escribir títulos no muy largos (que no superen los 50 caracteres), y no repetir más de 3 veces la misma palabra (Google lo puede considerar *spam*).

ALT: La etiqueta ALT está presente dentro de las etiquetas de imágenes, de la forma:

```
<IMG src="mi_imagen.gif" ALT="Mi comentario">
```

El texto de la etiqueta *ALT* surgió cuando había navegadores que no incluían las imágenes, y este texto era mostrado en lugar de la imagen. Aún hoy en día, algunos navegadores (como MS Internet Explorer) lo muestran cuando pasamos el ratón por encima de la imagen.

Google tiene en cuenta este texto, sobre todo si la imagen es un enlace a otra página web. Por ello es conveniente que dentro de la etiqueta ALT insertemos palabras claves.

META TAGS: Google NO tiene en cuenta los contenidos de los siguientes META TAGS:

META NAME=*keywords*

META NAME=*description*

Este último, sin embargo, es utilizado de vez en cuando por Google en lugar del *snippet* (la pequeña descripción que suele aparecer en los resultados de Google) cuando el contenido del META coincide con la búsqueda realizada.

De todas maneras, es recomendable utilizarlos, ya que algunos buscadores siguen haciendo uso de estos dos METAS. Recuerda que Google no es el único buscador, y podemos conseguir visitas desde otros buscadores.

URL: Se sospecha que Google sí valora que la URL (dirección de la página web) contenga las palabras claves, aunque no le da demasiado peso. Intenta que contenga las *keywords* deseadas, pero no abuses y no intentes que el dominio, subdominio y nombre de la página contenga estas palabras clave. Puedes conseguir que Google te penalice.

En las URLs intenta separar los nombre con guiones “normales” ("-"), y no con un guión bajo ("_"). Intenta escribir "mi-pagina.html" mejor que "mi_pagina.html".

En el resto de tu página web, intenta situar varias veces las palabras clave que intentas optimizar. Tampoco abuses de esto, porque tus textos serán más difíciles de leer (recuerda que diseñas las páginas para los usuarios, no para los buscadores).

Además, Google estima que determinados *TAGS* (etiquetas) reflejan mayor importancia del texto. Por ejemplo, situar un texto entre las etiquetas <H1> y </H1> lo realza en la apariencia que el usuario ve en la página, pero también Google estima que esas palabras son más importantes, y lo tendrá en cuenta. Lo mismo ocurre con las etiquetas <H2> (<H3>, <H4>,...), <B> (negrita) y <I> (itálica o cursiva). Conviene que repases el viejo HTML que has olvidado, o que eches un vistazo a algún tutorial de HTML.

Intenta diseñar las páginas web y sus contenidos para que las palabras clave aparezcan dentro de estas etiquetas, pero tampoco abuses de ello, ya que Google puede considerarlo como *spam*, y te puede penalizar.

Por otra parte, hay herramientas en Internet que obtienen la densidad de palabras claves de tu sitio web. Puedes encontrarlas en esta búsqueda:

http://www.google.com/search?q=keyword+density+analyzer

Cada una de las herramientas te dará un resultado diferente, porque en realidad Google utiliza un algoritmo bastante complicado para estimar en qué grado una página se ajusta a determinadas palabras claves. De todas maneras, puedes utilizar alguna de las herramientas sugeridas, e intentar que la densidad de tus palabras claves en tu página web sea del 5-20%.

Además, intenta que las palabras claves que has seleccionado aparezcan en los *links* que apuntan hacia tus páginas web.

Y, por supuesto, vigila la ortografía de tus palabras claves. Ya sabemos que mucha gente escribe mal las palabras, o que se confunden al escribir. Pero si buscamos “avogados en caracas”, Google nos sugerirá rápidamente “abogados en caracas”.

¿Debería tener mi propio dominio?

Sí. Aparte de la mejor imagen que puedas ofrecer a tus visitantes, puedes optimizar tu posicionamiento en Google gracias a los enlaces.

Tener tu sitio web en “paginas.sitios-gratis.com/mi-empresa/” da una imagen bastante mala, y el precio de un dominio ya no es excusa para que tengas el tuyo propio. Puedes comprarlo por menos de 10 dólares al año. Compara algunos precios de los registrars acreditados (es bastante más barato que otros vendedores de dominios), y elige el tuyo.

¿Qué dominio debo elegir?

Nuestra recomendación es que seas tú mismo y tengas tu propia marca. Como se comenta en este tutorial, debes diseñar tu sitio web para los visitantes, no para los buscadores. Si el sitio es bueno, la gente recordará “tunombre.com”, pero difícilmente “abogados-baratos-en-caracas.com”.

Fíjate en los ejemplos de Google y Yahoo!. Son nombres de marcas no muy sencillas, pero han conseguido que los usuarios las recuerden fácilmente. Incluso la compañía “goto.com” cambió su nombre a “Overture”.

Ahora bien, tener un dominio del tipo “abogados-catalunya.com” te da la opción de que desde otras páginas web te enlacen de la manera:

<A href=http://www.abogados-catalunya.com>abogados-catalunya.com</A>.

Esto te dará la posibilidad de optimizar tu sitio web para las palabras “abogados catalunya”.

Además, si quieres aparecer en los resultados de Google referentes a un determinado país, deberás tener un dominio del tipo –por ejemplo– “midominio.cat” (si quieres aparecer en los resultados de Cataluña) o “midominio.es” (en los de España). Google también te listará dentro de estos resultados si el servidor donde albergas tus páginas web está físicamente en estos países.

Echa un vistazo a la categoría de DMOZ de Registros por países encontrarás el adecuado para el país en el que quieres aparecer.

Las páginas dinámicas son páginas HTML generadas a partir de lenguajes de programación (*scripts*) que son ejecutados en el propio servidor web. A diferencia de otros *scripts*, como el JavaScript, que se ejecutan en el propio navegador del usuario, los 'Server Side' *scripts* generan un código HTML desde el propio servidor web.

Este código HTML puede ser modificado –por ejemplo– en función de una petición realizada por el usuario en una base de datos. Dependiendo de los resultados de la consulta en la base de datos, se generará un código HTML u otro, mostrando diferentes contenidos.

¿Cuáles son los principales tipos de páginas dinámicas?

Las páginas dinámicas se ejecutan en el propio servidor web. Por lo tanto, dependerán del tipo de servidor que dispongamos. Por ejemplo, si disponemos de un servidor con “Microsoft Windows Server”, generalmente encontraremos un servidor web Internet “Information Server” (IIS) que ejecuta *scripts* “Active Server Pages” (ASP). Aunque esto no es siempre así, porque actualmente hay paquetes de *software* que ejecutan todos los *scripts* en todos los servidores, siempre estaremos condicionados por los lenguajes diseñados especialemente para cada Sistema Operativo.

- **CGI**: Abreviatura de *Common Gateway Interface*. Se trata de un estándar para la interacción entre aplicaciones externas y servidores web. Gracias a ello, podríamos adaptar cualquier programa que hayamos realizado en cualquier lenguaje para que interactúe con nuestro servidor. Sin embargo, Perl se ha convertido en el lenguaje más popular para desarrollar aplicaciones CGI, aunque también se suele utilizar C, C++ ó Fortran.
- PHP: Lenguaje *script* de código abierto. Ampliamente utilizado sobre el servidor web Apache.
- ASP: Lenguaje script creado por Microsoft para su servidor web 'Internet Information Server' (IIS), y basado en “Visual Basic Script”. La última versión “ASP.net” forma parte del *Framework* “.net”.
- JSP: Lenguaje *script* creado por Sun, basado en la tecnología Java. No es necesario que el usuario disponga de la máquina virtual de Java ya que ésta se encuentra en el servidor que crea las páginas HTML. Tiene poco que ver que los *applets* de Java, y nada que ver con JavaScript. Los *scripts* JSP son un caso particular de los *servlets*.
- Cold Fusion: Lenguaje *script* creado por la compañía Allaire (adquirida más tarde por Macromedia). Los *scripts* tienen la extensión “.cfm”.

¿En qué me puede beneficiar usar páginas dinámicas?

Las páginas dinámicas nos pueden ayudar a gestionar más fácilmente los contenidos de nuestro sitio web y a interactuar con bases de datos.

Por ejemplo, si tenemos uno o varios menús en nuestras páginas, y queremos modificarlos, no tendremos que ir página por página editándolos, sino que bastará hacerlo una sola vez. En el resto de las páginas, bastará incluir (en PHP, por ejemplo): include 'menu-izquierda.html'.

Además, todos los lenguajes *script* comentados disponen de componentes para la conexión con la mayoría de las bases de datos (mySQL, Oracle, SQL Server, etc.). Esto nos puede servir para almacenar nuestros contenidos dentro de una base de datos, en lugar de realizar cada página web una por una.

Infórmate de las capacidades de cada uno de estos lenguajes *script*, y echa un vistazo a los tutoriales, que puedes encontrar en la Red, de CGI Perl, PHP, ASP, JSP y Cold Fusion.

HTML es un estándar propuesto por el Consorcio W3C, y se pretende conseguir que todos los documentos web sean compatibles en cualquier navegador (no solamente en ordenadores, sino también en cualquier dispositivo).

CSS son las siglas de *Cascade StyleSheet*, y especifica la forma del diseño de los documentos. Una misma página web (un mismo documento HTML, por ejemplo) puede ser vista de diferente forma en un PC que un PDA, gracias a diferentes hojas de estilo CSS.

Utilizar HTML+CSS te puede ayudar a mejorar tu posicionamiento web en Google. Por una parte, conseguirás que el código de tus páginas web sea más limpio y claro a los ojos del robot de Google. Facilitar la labor a este robot siempre es un punto a nuestro favor.

Por otra parte, aumentarás la densidad de las palabras claves dentro de los contenidos (ver "dónde situar las *keywords*"), ya que muchas de las etiquetas te ocuparán muchísimo menos espacio. Esto también supone un menor peso para tus páginas web, lo cual Google agradecerá. Y podrás a su vez cambiar rápidamente los estilos de ciertas palabras, modificando la importancia que les quieres otorgar.

Además, cumplir con el estándar HTML te abrirá la puerta a diseñar páginas web para dispositivos móviles o nuevas tecnologías que vayan surgiendo. Y el uso de CSS te permitirá cambiar el aspecto de estas páginas en cuestión de minutos. En combinación con las páginas dinámicas, puedes conseguir un sitio web realmente eficiente.

No se lo pongas difícil al robot de Google. Si insertas información en los siguientes elementos, ten la seguridad que no serán reconocidos:

- JavaScript
- DHTML
- Flash
- Frames
- Session IDs
- Applets de Java
- Imágenes: no insertes textos dentro de ellas.

Esto no significa que no puedas utilizarlos para el diseño de tu sitio web. Simplemente que la información que aparezca dentro de ellos no aparecerá en las búsquedas de Google.

¿Para qué navegador diseño mi sitio web?

No intentes que tus páginas web se vean mejor con un determinado navegador de Internet u otro. Como consejo, intenta que se vean correctamente con todos pero, sobre todo, con el "navegador" de Google, es decir, con su robot.

Si quieres saber cómo ve el robot de Google tu páginas web, puedes utilizar el navegador Lynx. Se trata de un navegador en modo texto, que no contempla las imágenes ni los elementos superflúos como el JavaScript, Flash, etc.

Si utilzas el Sistema Operativo Linux, probablemente lo tendrás instalado por defecto, y deberá escribir simplemente esto en la *shell*:

lynx http://www.mi-sitio-web.com

Si estás en otro Sistema Operativo (MS Windows, MAC, etc...), lo más sencillo es acceder via web a un emulador de Lynx:

http://www.delorie.com/web/lynxview.html

La caché de Google solamente almacena hasta un límite de 101 kb. Si miras en la información que guarda Google de cualquier página, verás que ésta no supera esa cantidad. Se sospecha que no se indexa más allá de este límite, sin embargo este punto no está demostrado.

Estos 101k son solamente referentes a código HTML (en el que se incluyen todos los textos e información). No se tienen en cuenta imágenes, gráficos Flash, etc.

De todas maneras, y como recomendación, intenta que tus páginas no tengan excesivo “peso”, es decir, que ocupen poco espacio. Si puede ser menos de 30 kb, mejor. Ten en cuenta que a la gente no le gusta esperar demasiado cuando accede a una página web, y muchos se cansan de esperar después de 3-4 segundos, y se van a otras páginas. La conexión por cable o banda ancha no está aún muy extendida, y la mayoría de los usuarios se conectan a Internet vía módem, o compartiendo la conexión de su empresa o universidad.

PageRank (PR) es un valor numérico que representa la importancia que una página web tiene en Internet. Google se hace la idea de que cuando una página coloca un enlace (*link*) a otra, es de hecho un voto para esta última.

Cuantos más votos tenga una página, será considerada más importante por Google. Además, la importancia de la página que emite su voto también determina el peso de este voto. De esta manera, Google calcula la importancia de una página gracias a todos los votos que reciba, teniendo en cuenta también la importancia de cada página que emite el voto.

PageRank (desarrollado por los fundadores Larry Page y Sergey Brin) es la manera que tiene Google de decidir la importancia de una página. Es un dato valioso, porque es uno de los factores que determinan la posición que va a tener una página dentro de los resultados de la búsqueda. No es el único factor que Google utiliza para clasificar las páginas, pero sí es uno de los más importantes.

Hay que tener en cuenta que no todos los *links* son tenidos en cuenta por Google. Por ejemplo, Google filtra y descarta los enlaces de páginas dedicadas exclusivamente a colocar *links* (llamadas *link farms*).

Además, Google admite que una página no puede controlar los *links* que apuntan hacia ella, pero sí que puede controlar los enlaces que esta página coloca hacia otras páginas. Por ello, *links* hacia una página no pueden perjudicarla, pero sí que enlaces que una página coloque hacia sitios penalizados, pueden ser perjudiciales para su *PageRankTM*.

Si un sitio web tiene PR0, generalmente es una web penalizada, y podría ser poco inteligente colocar un *link* hacia ella.

Una manera de conocer el *PageRankTM* de una página es decargándose la barra de búsqueda de Google (solamente disponible para MS IExplorer). Aparece una barra en la que se muestra en color verde el valor de PageRankTM en una escala de 0 a 10. Sitios web con PR10 son Yahoo!, Microsoft, Adobe, Macromedia, o la propia Google. Tenéis una lista completa con los sitios con PR10.

El algoritmo de PageRank fue patentado en Estados Unidos el día 8 de enero de 1998, por Larry Page. El título original es Method for node ranking in a linked database, y le fue asignado el número de patente 6,285,999.

Conseguir enlaces es una de las labores más críticas en el posicionamiento web en Google. En función del número de enlaces que obtengamos, tendremos mayor PageRank o popularidad.

Hay que ser extremadamente cuidadosos con la manera que realizamos el *link* o enlace. En función de la manera en que esté realizado el enlace, promocionaremos unas palabras clave u otras.

Además, hay que saber dónde conseguir enlaces. Hay determinados sitios web donde es más fácil obtenerlos, pero casi siempre hay que esforzarse por conseguirlos y hacer un seguimiento

¿Por qué hay que conseguir enlaces?

Google otorga un valor numérico a cada página web que inserta en su base de datos. Este valor numérico lo denomina *PageRank*. Cuanto mayor sea el *PageRank* de una página, mayor importancia le habrá dado Google.

Y este valor crece cuantos más sitios web enlacen a tu página (ver “transmisión del *PageRank*”). Tienes que pensar que cada *link* es para Google como si fuera un “voto”. Además, si estos sitios web que te enlazan tienen un *PageRank* elevado, el valor crecerá más, porque el “voto” es de mayor calidad.

¿Cómo puedo conocer el PageRank de una página?

Para conocer este valor, puedes decargarte la barra de búsqueda de Google (solamente disponible para MS IExplorer). http://toolbar.google.com/. En ella, hay un espacio para mostrar el *PageRank* (PR) de cada página que visitas. Este valor varía entre 0 y 10.

Sin embargo, no debes centrar todos tus esfuerzos en conseguir un PR elevado. El valor del *PageRank* de una página es importante, pero te habrás dado cuenta de que hay páginas que, teniendo menor PR, están posicionadas por encima de otras de mayor *PageRank* para determinadas búsquedas.

Esto lo han conseguido –entre otras cosas– porque han optimizado mejor el contenido de sus páginas, porque han realizado unos buenos enlaces y porque han conseguido insertar estos enlaces en buenas páginas web.

Aunque Google sea el buscador por excelencia (más del 60% de los usuarios lo utilizan), no es el único. También deberías intentar conseguir tráfico desde otros buscadores.

Además, Google también indexa muchos de los directorios de estos otros buscadores, lo que puede suponer conseguir más enlaces. No van a tener un *PageRank* tan elevado como Yahoo! o DMOZ, pero todos los enlaces son tenidos en cuenta.

Echa un vistazo a las siguientes categorías de DMOZ. Puedes encontrar algunos directorios y buscadores que te pueden resultar útiles:

- ➔ http://dmoz.org/Computers/Internet/Searching/Directories/
- ➔ http://dmoz.org/Computers/Internet/Searching/Search_Engines/
- ➔ http://dmoz.org/World/Español/Referencia/Buscadores_y_directorios/

¿Debo utilizar los sistemas automáticos de envío a buscadores?

Nuestra recomendación es que no utilices los sitemas automáticos que prometen dar de alta tu sitio web en cientos o miles de buscadores. Es mejor que lo hagas personalmente, porque muchos buscadores detectan envíos automáticos, y desestiman las páginas web enviadas por estos métodos.

Además, estos sistemas automáticos son diseñados para un determinado momento. La mayoría de los buscadores modifican los formularios y requerimientos para darse de alta en ellos, y los sistemas automáticos no lo modifican a la vez. Puede ocurrir que simplemente no funcione o que, por ejemplo, no te des de alta en la categoría adecuada.

Recuerda que el posicionamiento web requiere mucho esfuerzo y dedicación. Ten paciencia, y hazlo todo personalmente.

¿Cómo puedo conseguir aparecer en DMOZ?

Hay categorías de dmoz.org cuya página web tiene un *PageRank* de 7 ú 8, por lo que el conseguir un enlace en DMOZ es realmente valioso. Además, Google lo toma como referencia para construir su propio directorio "directory.google.cat". No escojas la categoría con mayor *PageRank*, sino la que más se ajuste a la temática de tu sito web.

Pero no te creas que es una labor fácil aparecer en dmoz.org, ya que los editores solamente incluyen sitios web de calidad y que realmente tienen que ver con la temática de cada categoría.

Navega por las categorías de DMOZ. Descubre cuál es la que más se ajusta a la temática de tu sitio web, y pulsa el enlace "agregar URL". Si tienes un sitio web en español, lo más conveniente es que escojas una categoría dentro de "World > Español".

Debes tener paciencia con tu solicitud. Suelen tardar varias semanas, pero no debes agobiar a los editores. De todas maneras, puedes contactar y debatir con ellos en "www.resource-zone.com". Si en el plazo de seis meses no has conseguido que te enlacen, deberías volver a sugerir tu sitio web.

Conviene que esperes a tener el sitio web realmente listo antes de sugerir a la gente DMOZ un enlace. Si les sugieres un sitio web "en construcción" seguro que no te lo van a aceptar, y posiblemente no te lo revisarán la próxima vez que se lo sugieras. No te precipites y haz las cosas con calma.

¿Tienen algún tipo de relación Google y DMOZ?

No. DMOZ permite la reproducción libre de sus directorios (con una licencia especial), y Google simplemente se limita a recoger sus contenidos desde el año 2000.

Otros muchos sitios hacen lo mismo que Google en su "directory.google.cat", y recogen en sus páginas web un directorio de categorías con los enlaces de los que dispone DMOZ. Google trata estos enlaces de la misma manera que el resto de los *links* pero, al aparecer en más sitios web –debido a estos "clones" de DMOZ–, el número de enlaces se multiplica.

➔ C. La web privada / la web propietaria / la web realmente invisible

Diversos especialistas y entidades académicas se dedican a la tarea de elaborar y mantener páginas concentradoras de recursos web seleccionados por áreas de especialidad, (*subject guides*), que pueden contener recursos que no son recuperables con un buscador común. Estos directorios anotados o guías temáticas suelen tener un alto grado de calidad, ya que comprometen el prestigio de los autores y de las instituciones involucradas. La selección de recursos suele ser muy cuidadosa y su actualización frecuente. En ocasiones, diversas instituciones se asocian formando "circuitos" (*web rings*) para la elaboración cooperativa de estas guías. Un buen ejemplo de ello es The WWW Virtual Library.

Los directorios anotados o guías pueden incluir, además, algún mecanismo de búsqueda en sus páginas o en la Web en general (Moreno Jiménez, 2004). Comúnmente no basta con conocer la variedad de herramientas de búsqueda disponibles en la Web, sino que se requiere una orientación sobre su funcionamiento, sobre qué estrategias seguir para trazar una adecuada ruta de búsqueda y sobre cómo elegir los mejores instrumentos para cada necesidad. De ello se ocupan los tutoriales. *How to Choose a Search Engine or Directory*, de la Universidad de Albany, en Estados Unidos, y las guías de *SearchAbility* y

de la Universidad de Leiden en Holanda *A Collection of Special Search Engines* orientan al usuario en el amplio mundo tanto de los recursos especializados en la Web como de las maquinarias que permiten su localización.

Pero más allá de todas estas herramientas y recursos se encuentra la Web invisible.

Sherman y Price identifican cuatro tipos de contenidos invisibles en la Web:

- La Web opaca (the opaque web).
- La Web privada (the private web).
- La Web propietaria (the proprietary web).
- La Web realmente invisible (the truly invisible web).

La Web opaca se compone de archivos que podrían estar incluidos en los índices de los motores de búsqueda, pero no lo están por alguna de estas razones:

- Extensión de la indización: por economía, no todas las páginas de un sitio son indizadas en los buscadores.
- Frecuencia de la indización: los motores de búsqueda no tienen la capacidad de indexar todas las páginas existentes; diariamente se añaden, modifican o desaparecen muchas y la indización no se realiza al mismo ritmo.
- Número máximo de resultados visibles: aunque los motores de búsqueda arrojan a veces un gran número de resultados de búsqueda, generalmente limitan el número de documentos que se muestran (entre 200 y 1000 documentos).
- URLs desconectados: las generaciones más recientes de buscadores, como Google, presentan los documentos por relevancia basada en el número de veces que aparecen referenciados o ligados en otros. Si un documento no tiene una liga en otro documento será imposible que la página sea descubierta, pues no habrá sido indizada.

La Web privada está compuesta por páginas web que podrían estar indizadas en los motores de búsqueda pero son excluidas deliberadamente por alguna de estas causas:

- Están protegidas por contraseñas (*passwords)*.
- Contienen un archivo “robots.txt” para evitar ser indizadas.
- Contienen un campo “noindex” para evitar que el buscador indice la parte correspondiente al cuerpo de la página.

La Web propietaria incluye aquellas páginas en las que es necesario registrarse para tener acceso al contenido, ya sea de forma gratuita o pagada.

Se dice que al menos el 95% de la Web profunda contiene información de acceso público y gratuito (Turner, 2003).

La Web realmente invisible se compone de páginas que no pueden ser indizadas por limitaciones técnicas de los buscadores, como las siguientes:

- Información almacenada en bases de datos relacionales, que no puede ser extraída a menos que se realice una petición específica. Otra dificultad consiste en la variable estructura y diseño de las bases de datos, así como en los diferentes procedimientos de búsqueda.

Herramientas de búsqueda en la Web profunda

Los motores de búsqueda han mejorado su desempeño en los últimos años, permitiendo un mayor nivel de precisión en las búsquedas y ofreciendo los resultados en formas cada vez más convenientes para el usuario, pero aún son muchos los buscadores que sólo pueden recuperar directamente la información que se encuentra disponible *en* la Web y no aquella que se ofrece a través de la Web. Cuando se tomó conciencia de la magnitud de la Web que resultaba "invisible" por las dificultades que presentan los motores de búsqueda para acceder a ellos, éstos incorporaron funcionalidades adicionales para facilitar la búsqueda en la llamada Web profunda y han surgido buscadores especializados en ese segmento de la Web. Para encarar una búsqueda en la Web profunda se debe tener en cuenta que los metabuscadores pueden presentar limitaciones, respecto a las posibilidades de búsqueda de cada buscador por separado. Por ejemplo, cuando la búsqueda es sobre materiales o formatos especiales, resulta más práctico utilizar las opciones de búsqueda avanzada que presentan los buscadores y, si fuera necesario, realizar búsquedas sucesivas en varios de ellos o recurrir a los directorios concentradores de buscadores. Los mecanismos utilizados para localizar recursos en la Web profunda consisten, mayoritariamente, en directorios de recursos especializados, principalmente bases de datos disponibles de forma gratuita en la red. El patrocinio de las instituciones académicas en la elaboración de los directorios, particularmente de los que son anotados, garantiza la cobertura y calidad de los recursos compilados. Las guías de recursos especializados generalmente están elaboradas por bibliotecarios y son una excelente herramienta de búsqueda y localización de recursos, además de constituir un buen instrumento de aprendizaje en el uso de la información.

Las páginas *How to Choose a Search Engine or Directory* de la Universidad de Albany en Estados Unidos y las guías de *SearchAbility* y de la Universidad de Leiden en Holanda *A Collection of Special Search Engines* incluyen los recursos de información y búsqueda en la Web profunda.

Finalmente, los motores de pregunta dirigida (*directed query engines*) tienen la capacidad de realizar búsquedas simultáneas en varias bases de datos en la Web. Lexibot y su sucesor, Deep Query Manager, así como Distributed Explorer (Warnick y otros) y FeedPoint, son ejemplos de estos motores avanzados de búsqueda.

Estrategias de búsqueda en la Web profunda

Además de las estrategias ya señaladas para la búsqueda en la Web, podemos añadir otras específicas para la búsqueda en la Web profunda o invisible, agrupadas en *rubros* orientativos.

Para la búsqueda de información especializada:

Usar las siguientes herramientas de búsqueda en la Web profunda si buscamos:

información académica de calidad.

- ➔ Usar buscadores regionales especializados para localizar información
- ➔ Originada fuera de los Estados Unidos o en idiomas diferentes al inglés.
- ➔ Usar metabuscadores para realizar búsquedas en varios buscadores especializados a la vez.

Para realizar búsquedas avanzadas:

- ➔ Usar las opciones avanzadas de los buscadores para localizar imágenes o archivos PDF o PostScript.
- ➔ Usar directorios concentradores de buscadores para realizar búsquedas avanzadas sucesivas en varios de ellos.

Para evaluar la información disponible en la Web:

- ➔ Usar directorios Aanotados para evaluar si los recursos disponibles en la Web profunda son útiles para la búsqueda que estamos realizando.
- ➔ Usar directorios de bases de datos para conocer cuáles de ellas pueden ofrecernos información útil para nuestras búsquedas.

Para buscar información en bases de datos:

- ➔ Usar guías, directorios o motores avanzados si la información que buscamos puede estar en una base de datos.

No cabe duda de que los actuales buscadores y directorios de la Web están mejorando su funcionamiento. Más allá de los detalles técnicos que el público no alcanza a ver, la eficiencia de estas tecnologías ha aumentado y esto se aprecia en los resultados de las búsquedas. A medida que estas herramientas se vayan haciendo más poderosas, disminuirá la necesidad de la elaboración manual de guías o concentradores de recursos, y quizás más la de orientación en las estrategias de búsqueda y en el uso y aprovechamiento de los recursos localizados.

Observando los resultados obtenidos por los motores de búsqueda, se puede verificar que persiste aún la práctica de no indexar todas las páginas por parte de los robots de un sitio. Por ejemplo, se puede tener la referencia de una base de datos que está disponible a través de un sitio web, mediante un enlace a ella que contiene una de las páginas del sitio y, en cambio, puede no aparecer la referencia a la página de acceso directo a esa base de datos en ese sitio.

Es evidente que la frecuencia de la indización ha aumentado en algunos buscadores, e incluso ésta se realiza de forma diferenciada para algunos recursos. Aquellas páginas que varían más, por su naturaleza, (la información bursátil, por ejemplo,) serían visitadas con mayor frecuencia por los robots que aquellas que tienden a ser más estables en su contenido.

El número máximo de resultados visibles no es un problema cuando los buscadores presentan los resultados ordenados por relevancia, pues siempre aparecerán primero aquellos que se ajustan más a la búsqueda realizada. En la medida en que se pueda realizar una búsqueda avanzada y los criterios de relevancia combinen el número de ligas con la frecuencia de palabras, la presentación de los resultados no constituirá un obstáculo para encontrar la información.

El usuario siempre debe tener en cuenta que los buscadores son más apropiados cuando la búsqueda es específica, es decir, se conocen datos sobre lo que se busca; mientras que es más adecuado realizar búsquedas temáticas en los directorios. Los URLs desconectados podrían evitarse si existiera la obligación de registrar, aunque fuera de forma muy sencilla, toda página que se colgara en la Web. Pero dada la gran descentralización de Internet, esto no parece vislumbrarse en un futuro inmediato.

El segmento de la Web privada no representa una pérdida de gran valor, en términos de la información que contiene, ya que en general se trata de documentos excluidos deliberadamente del circuito informacional por su escasa utilidad. En cualquier caso, son los dueños de la información los que deciden no hacerla disponible, por lo que difícilmente se podrán encontrar mecanismos legítimos para franquear esa barrera. Además, los archivos robots.txt sirven

para evitar que los robots caigan en “agujeros negros”, que les hagan entrar en procesos circulares interminables, mermando así la eficiencia en su funcionamiento.

En un artículo reciente de la OCLC Office for Research (O'Neill; Lavoie y Bennett) se examinan las tendencias en cuanto a tamaño, crecimiento e internacionalización de la Web pública, es decir, la porción de información más visible y accesible para el usuario promedio. Las principales conclusiones del estudio son:

El crecimiento de la Web pública muestra un estancamiento en los últimos años. Ello se debe a que se crean menos sitios web y otros desaparecen, aunque esto no quiere decir que no aumente el volumen de información, es decir, el número de páginas o el número de terabytes. Otra posibilidad, que no se señala en este estudio pero que puede deducirse de las restricciones para el acceso a ellos, es que algunos sitios web son accesibles mediante el pago de una suscripción u otro medio de registro.

La Web pública está dominada por contenidos originados en los Estados Unidos, escritos en inglés. Esto nos lleva a pensar que probablemente haya más recursos invisibles en páginas originadas en otros países (distintos a los Estados Unidos) y en otros idiomas.

Algunos buscadores tradicionales como Altavista o Google han evolucionado y presentan ahora la posibilidad de realizar búsquedas por materiales o formatos especiales. Así, Google permite realizar búsquedas avanzadas para localizar imágenes. Por su parte, el concentrador HotBot presenta la posibilidad de buscar por distintos formatos, para localizar imágenes, audio, vídeo, archivos PDF, Script y Shockwave/Flash. Estas opciones están activas en HotBot para los buscadores Fast (Altheweb) e Inktomi (Pure Web Search), mientras que no funcionan con Teoma ni Google, aunque como dijimos existe esta posibilidad si se realiza la búsqueda directamente desde el sitio de Google.

Estas búsquedas en materiales especiales, como imágenes, audio y vídeo, son posibles gracias a una catalogación textual de los mismos. Las búsquedas en documentos que presentan formatos PDF, Flash, etc., se pueden realizar porque existen directorios de estos archivos. Así, el principal medio por el cual se pueden efectuar las búsquedas es el texto. Por ejemplo, si queremos recuperar imágenes en blanco y negro, éstas deben estar clasificadas de ese modo en la base de datos. Esto implica, lógicamente, un proceso manual. Una página web que contiene una imagen, sin mayor información textual acerca de su contenido, no podrá ser recuperada automáticamente más que por su extensión (“.jpg”, por ejemplo).

Como hemos visto, la definición más genérica de lo que constituye la Web invisible o profunda apunta a los recursos que no pueden ser recuperados

mediante las herramientas comunes de búsqueda. Para verificar la visibilidad de la Web profunda, que ha sido identificada por los autores de *The Invisible Web*, Moreno Jiménez (2003) ha seleccionado al azar diez recursos de su *The Invisible Web Directory* y realizó la búsqueda en un buscador, un directorio, un metabuscador y un agente metabuscador avanzado en su versión gratuita. Los resultados de esta sencilla prueba aparecen reflejados en el cuadro siguiente:

Recurso	**MSN**	**Yahoo**	**MetaCrawler**	**Copernic**
Artcyclopedia	SI	SI	SI (6 buscadores)	SI (8 buscadores)
CRA Forsythe List	SI	SI	SI (3 buscadores)	SI (5 buscadores)
Current Films in the Work (BoxofficeHollywood Hot Set)	SI	SI	SI (3 buscadores)	SI (4buscadores)
Employee Benefits INFOSOURCE	SI	SI	SI (2 buscadores)	SI (3 buscadores)
Hamnet	SI	SI	SI (4 buscadores)	SI (6 buscadores)
Infonation	SI	SI	SI (5 buscadores)	SI (7 buscadores)
Jourlit	SI	SI	SI (3 buscadores)	SI (7 buscadores)
Scholarly Societies Project	SI	SI	SI (4 buscadores)	SI (6 buscadores)
Vessel Registration Query System	SI	SI	SI (2 buscadores)	SI (6buscadores)
Who's who in American Art (AskArt)	SI	SI	SI (6 buscadores)	SI (8 buscadores)

Cuadro. 15. Resultado de búsqueda de recursos de *The Invisible Web Directory.*

Todos los recursos seleccionados de *The Invisible Web Directory* son localizables con las actuales herramientas de búsqueda. Además, en los resultados se observa que existen múltiples referencias en otras páginas, es decir, que se trata de páginas "conectadas". La única dificultad para encontrarlas consiste, en algunos casos, en las palabras con las cuales se denomina el sitio o el recurso. Por ejemplo, en el *The Invisible WebDirectory* aparece "Vessel Query Registration System", en lugar de "Vessel Registration Query System", lo cual hace que la búsqueda por todas las palabras sea

exitosa, pero la búsqueda por frase no. Igualmente, la denominación de “Who’s who in American Art” para el sitio de “AskArt” dificulta la búsqueda, mientras que si se busca directamente por su nombre aparece en numerosos buscadores. La tabla refleja además cómo el solapamiento entre buscadores es variable.

Puede decirse que el contenido de las bases de datos que están incluidas en este directorio es invisible, ya que es necesario realizar las búsquedas directamente en cada una de ellas. Pero lo cierto es que llegar hasta la “puerta” de estas bases de datos resulta relativamente sencillo. El mismo hecho de que el directorio haya sido colocado en la Web, le confiere mayor visibilidad a los recursos incluidos, ya que los enlaces en el directorio aumentan la posibilidad de indización de esas páginas. Entonces, podemos decir que *The Invisible Web Directory* es un buen directorio de recursos y bases de datos disponibles en la Web, pero no un directorio de recursos “invisibles”. En conclusión, lo que realmente sigue siendo invisible en la Web son:

- Las páginas desconectadas.
- Las páginas no clasificadas que contienen principalmente imágenes,
- audio o vídeo.
- El contenido de las bases de datos relacionales.
- El contenido que se genera en tiempo real.

Pero:

- es relativamente sencillo llegar hasta la “puerta” de las bases de datos con contenido importante;
- existen ya motores avanzados capaces de realizar búsquedas directas
- simultáneas en varias bases de datos a la vez; y aunque la mayoría requieren pago, también ofrecen versiones gratuitas; el contenido que se genera en tiempo real pierde validez con mucha velocidad, salvo para análisis históricos;
- es relativamente sencillo llegar hasta la “puerta” de los servicios que ofrecen información en tiempo real; el contenido que se genera dinámicamente interesa únicamente a ciertos usuarios con características específicas;
- es relativamente sencillo llegar hasta la “puerta” de los servicios que ofrecen contenido generado dinámicamente.

➔ D. La Web realmente invisible

Este contenido no puede ser indexado por los buscadores por razones técnicas. Los documentos pueden estar en un formato que los robots no

reconozcan (música, videos) o por páginas generadas dinámicamente (la página se autodiseña, no hay ningun diseñador humano de por medio o se genera sola). Ese tipo de páginas (foros de discusión, sitios de remates, catálogos, diccionarios, etc.) no son tenidas en cuenta por los buscadores.

A la hora de emprender una investigación, y tras tener delimitado inicialmente el tema, es preciso realizar la pertinente búsqueda de la documentación bibliográfica necesaria. La ciencia es un trabajo colaborativo y acumulativo. Aunque trabajemos solos, necesitamos consultar las teorías, modelos, métodos, resultados, hallazgos, datos, etc. aportados por otros autores sobre el mismo tema que hemos decidido investigar. Se trata de no caer en los mismos errores y/o no duplicar esfuerzos y hacer aportaciones originales a la comunidad científica.

La documentación psicológica es imprescindible en el trabajo de investigación científica en Psicología y se necesita en varias fases del proceso de elaboración de escritos científicos.

En primer lugar, tenemos que saber identificar la documentación científica relevante, sus tipos y formatos (apartado 2). Al respecto, la situación actual es la de una imparable migración de las fuentes documentales en Psicología de publicaciones impresas (formato papel) a publicaciones electrónicas, principalmente en Internet.

Luego entraríamos en la documentación bibliográfica en Psicología (apartado 3), tanto en formato impreso tradicional, como en Internet, ya sea “formal” o “informal” u otras informaciones disponibles en Internet.

Qué buscar: tipos y formatos de la documentación bibliográfica

- ➔ Tipos de documentos: fuentes primarias y secundarias.
- ➔ Formatos de la documentación: impreso y electrónico.

Tipos de documentos: fuentes primarias y secundarias

- ➔ Fuentes primarias: son las que proporcionan directamente información sobre un tema concreto (libros, artículos, diccionarios, etc.), tanto información básica o preliminar en las obras de referencia, como más profunda en manuales o monografías:
- ➔ Obras de referencia: es decir, diccionarios y enciclopedias. En la biblioteca hay buen número de ellas relativas a la Psicología y ciencias afines. No se localizan mediante los ficheros, sino acudiendo directamente al estante donde se agrupan todas ellas (actualmente son los primeros estantes que se encuentran a la izquierda, al entrar).
- ➔ Manuales, tratados y monografías. Mediante la consulta de los ficheros de la Biblioteca (autores, materias, títulos) podemos localizar libros

especializados en el tema. Si no es así, los manuales o tratados más amplios pueden ofrecer una primera visión.

➔ Fuentes secundarias: son las que nos indican cómo y dónde hallar las fuentes primarias. Contienen referencias de otros trabajos, permitiendo así su conocimiento y/o localización. Muchas veces están elaboradas por instituciones (por ejemplo, revistas de resúmenes, catálogos, etc.).

Fuentes primarias y secundarias, sus niveles de complejidad y ejemplos de documentos correspondientes.

Fuentes primarias

NIVEL SUPERFICIAL	NIVEL MEDIO	NIVEL ESPECIALIZADO
Enciclopedias		
Diccionarios		
Tesauros		
Manuales		
Compilaciones		
	Monografías	
	Series	
	Artículos en revistas y boletines	
		Actas de Congresos
		Publicaciones preliminares
		Tesis, tesinas

Fuentes secundarias

NIVEL SUPERFICIAL	NIVEL MEDIO	NIVEL ESPECIALIZADO
Reseñas bibliográficas		
Información sobre tests y audiovisuales		
Revisiones		
	Bibliografías	
	Catálogos	
		Resúmenes
		Índices

Formatos de la documentación: formato impreso y formato electrónico

Fomato impreso:

Tradicionalmente, las fuentes documentales las tenemos disponibles en papel impreso, sobre todo las fuentes primarias. Libros, monografías, tesis, artículos de revistas, etc. siguen encontrándose mayoritariamente en formato papel y, por tanto, para acceder a esas fuentes hay que adquirirlas o consultarlas en los centros de documentación que tengamos más accesibles.

Formato electrónico:

Cada vez son más útiles y se dispone de más fuentes en formato electrónico, tanto fuentes primarias (revistas electrónicas accesibles por Internet) como sobre todo fuentes secundarias organizadas como bases de datos referentes a catálogos de bibliotecas, índices de sumarios o resúmenes de revistas, etc.

El formato electrónico hace referencia a soportes magnéticos accesibles directamente (discos flexibles, duros, discos compactos para CD-ROM, etc.) o vía telemática (Internet, etc.) que contienen bases de datos en las que se almacena la información y es procesada y recuperada por medios informáticos. Tanto fuentes primarias como secundarias se encuentran ya en este soporte, más las segundas que las primeras. Así, cada vez tenemos más artículos de revistas en formato pdf o html en Internet. Y en cuanto a fuentes secundarias, lo más utilizado son las bases de datos sobre:

- resúmenes y referencias de artículos de revistas y capítulos de libros,
- libros,
- resúmenes de tesis doctorales y memorias de licenciatura,
- disposiciones legales,
- catálogos comerciales de editoriales y empresas de *software*, etc.

Hoy día, por lo general, las fuentes secundarias importantes para la investigación psicológica están accesibles en las universidades desde cualquier ordenador dentro de ellas conectados a Internet.

Qué buscar en Internet:

- Documentación bibliográfica, tanto fuentes primarias como secundarias.
- Información temática "informal".
- Información sobre centros y recursos.
- Intercambio de información sobre temas concretos.

Documentación bibliográfica, tanto fuentes primarias como secundarias

Fuentes primarias

- Podemos encontrar artículos de revistas e incluso libros en la Web. El problema es la dificultad para encontrarlos pues a veces se trata de un autor que en la web de su Departamento, en un apartado de profesores, y en su página particular ha puesto su bibliografía más reciente.
- Revistas electrónicas: se trata de revistas que han surgido en la Web, o bien se editaban normalmente en papel y ahora también han pasado a formato electrónico (accesibles gratuitamente desde la red).

Fuentes secundarias

Bases de datos de centros españoles accesibles desde la Web y con documentación de diversas áreas. Hay que distinguir dos tipos:

1) Bases de datos comerciales suscritas por las universidades: se trata de la documentación que sólo se puede conseguir mediante pago o suscripción, sea de modo personal o por una institución o centro de documentación (que si es pública permitirá el acceso a los investigadores).

Esta información suele estar almacenada en bases de datos elaboradas por empresas de documentación y que cobran por acceder a ellas. Las universidades las suscriben con licencia de red y desde sus ordenadores se accede libremente a ellas (no desde ordenadores externos a la universidad).

2) Fuentes secundarias de dominio público en la web:

- ➔ Catálogos de bibliotecas de universidades y centros de investigación.
- ➔ Índices de revistas: algunas revistas que se editan en papel ya exponen sus índices e incluso resúmenes en la Web.
- ➔ Bases de datos de índices o resúmenes de revistas.

Información temática "informal":

En cuanto a las búsquedas por áreas temáticas, sobre todo en directorios temáticos, su ventaja, y también su inconveniente, es que las direcciones web que incluyen están preseleccionadas por el autor o autores de la web, es decir, dependemos de sus criterios de selección, normalmente sesgados en función de su orientación empresarial, teórica (en psicología, orientaciones cognitiva, psicodinámica, etc.) o profesional.

Información sobre centros y recursos:

Es la información típica, la que primero fue surgiendo en la Web. Los centros o instituciones que tenían servidor web, lo primero que hicieron fue colgar información institucional sobre ellos mismos.

Intercambio de información sobre temas concretos:

El sistema más clásico es el de las *News* (ya mencionado). También se han utilizado listas de correo, dentro del sistema de correo electrónico, o accesibles desde la Web. También el sistema chat o IRC (*Internet Relay Chat*).

Cómo buscar los documentos en formato impreso

- ➔ Biblioteca y/o hemeroteca de Facultad o de Universidad.
- ➔ Biblioteca de área o de departamento universitarios.
- ➔ Otros: Centro de Documentación del Colegio Oficial de Psicólogos de Madrid y CINDOC (Centro de Información y Documentación Científica).

Cómo buscar los documentos en Internet

Los buscadores o motores de búsqueda:

La estrategia de búsqueda de información más común es a través de buscadores o motores de búsqueda. Se trata de webs que generan listados de direcciones web tras introducir las palabras clave específicas del tema buscado. El más potente es Google en la dirección http://www.google.com.

Aquí podemos encontrar una amplia variedad de temas. Tanta variedad y tantos nodos web hay ya sobre todo esto que se corre el peligro de perder una gran cantidad de tiempo buscando lo que interesa. De ahí que sea fundamental conocer los dos sistemas de búsqueda disponible:

- ➔ A través de motores de búsqueda (search engine) en la Web, tipo Google, Yahoo, Lycos, etc. Son sistemas que funcionan con palabras clave; el resultado de la búsqueda es un listado de direcciones web en los que se mencionan temas relacionados con las palabras clave buscadas. Hoy día prácticamente el más útil, actualizado y amplio es Google.
- ➔ A través de directorios elaborados por instituciones o personas sobre temas o aspectos concretos.

Las ventajas e inconvenientes de estos sistemas son:

La ventaja de los buscadores generales por palabras clave radica en que sus bases de datos están permanentemente actualizadas, incluyen direcciones web de modo automático –el autor de la web no tiene que solicitar su inclusión, uno o dos meses después de publicar una web ya aparece en alguno de estos buscadores– y además el tiempo de búsqueda es brevísimo (menos de un segundo para el programa de búsqueda y unos pocos segundos más debido a la velocidad de nuestra conexión a Internet). Debe utilizarse este recurso cuando queramos encontrar algo específico y con rapidez y en gran cantidad.

El inconveniente es que hay que afinar mucho en las palabras clave y en el listado que nos devuelve el buscador suelen entrar muchas direcciones web inservibles (lo que se llama “ruido” en Internet). Es decir, los motores de búsqueda recogen páginas web automáticamente y por tanto sin intervención humana y sin control de calidad.

Otro problema es que a veces las recopilaciones de hiperenlaces no están muy actualizadas (actualizar las direcciones web es una tarea muy laboriosa y además hay que comprobar periódicamente si los enlaces funcionan, etc.).

Criterios de calidad de la información de las páginas web

En estos casos, el problema es la sobreabundancia de información y saber discriminar la información de calidad. Para ello, en Romero (2002) se especifican una serie de criterios que sería conveniente tener en cuenta para determinar la calidad de una web sobre psicología:

1) Lo primero es excluir de la recopilación:

- ➔ Páginas primariamente comerciales.
- ➔ Páginas basadas en investigación escasa o poco fundamentada.
- ➔ Ausencia de un patrocinador claramente identificado.
- ➔ Ausencia de contenidos suficientes sobre el tema.

2) Excluidas esas páginas web, se deben utilizar las que se les pueda asignar un mínimo de dos puntos (sobre una escala de 5) en los siguientes criterios:

- ➔ Contenidos, con objetividad, originalidad y citación de las fuentes de los hallazgos de investigación y estadísticas.
- ➔ Autoridad, basada en las credenciales, tanto de la organización que patrocina como de los autores individuales de la información presentada. Las credenciales incluyen factores tales como estatus educativo de directores –*staff* i autores–, número y calidad de las publicaciones de investigación, afiliaciones institucionales, experiencia profesional, etc.
- ➔ Actualización y estabilidad, presencia de fecha de creación o *copyright*, evidencia de mantenimiento de la web tal como fechas de actualización de la web o consistencia de fechas en páginas interiores.
- ➔ Facilidad de uso: accesibilidad del material en el sitio web, facilidad de navegación, formato consistente y coherente de todas las páginas del sitio web, operatividad de los enlaces, tiempo de descarga de la web aceptable.

En una puntuación global del sitio web evaluado, los dos primeros criterios tienen más importancia que los dos últimos (en los criterios antes mencionados, cada uno de los dos criterios supondrían cada uno el 36% de la puntuación global, y los otros dos un 14% cada uno).

3) Por otra parte, a la hora de describir los sitios web de psicología, además de los criterios anteriores, tendríamos que tener en cuenta:

- ➔ País de origen, idioma, enlaces en español y/o en inglés.
- ➔ Temática.
- ➔ Si contienen textos completos (fuentes primarias) y/o referencias (fuentes secundarias).
- ➔ El enfoque o utilidad de la información (práctica, divulgación, profesional, enseñanza/docencia, investigación, etc.).

➔ E. Internet invisible

Definición y retos

Lluís Codina es profesor titular de Ciencias de la Documentación en la Universitat Pompeu Fabra y miembro del Observatorio de la Comunicación Científica. Es sin duda uno de los mejores expertos en Análisis y Métodos en Ciencias de la Documentación. Podéis encontrar mucha documentación en su página web http://www.lluiscodina.com/. Lluis Codina afirma que:

"Internet invisible es un nombre claramente inadecuado para referirse al sector de sitios y de páginas web que no pueden indexar 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 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 indexar páginas HTML. Cualquier otro formato era ilegible, es decir, invisible para tales motores. Todos conocemos 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 este caso un dibujo de un tomate, en formato no HTML (SVG) visto en un navegador con el plugin adecuado.

El formato de gráficos vectoriales SVG está basado en XML y por tanto es indexable por los buscadores. El formato Flash (animaciones) ya es indexado también por Google. Para visualizarlo necesitas instalar en tu navegador un plugin adecuado. Más información en http://es.wikipedia.org/wiki/.svg.

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 esta:

http://www.allmovie.com/search/work/star+trek/results

Los motores de búsqueda no pueden indexar 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 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.

Es decir, podemos acceder a las páginas principales de los sitios web que proporcionan acceso a bases de datos, porque estas 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.

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. Este protocolo consiste en un pequeño número de valores que puede adquirir el atributo c*ontent* 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 este 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 humano que la extranet o la intranet de una corporación sea invisible a los motores de búsqueda. No sólo no es un problema, sino que es deseable que siga siendo así. Nadie quiere que los motores de búsqueda puedan indexar 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 indexar 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 y 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 *interface* 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 interfaz 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 esta base de datos, y cada base de datos, no solamente presenta una interfaz de usuario diferente, sino un conjunto de funciones distintas.

Acceder a los contenidos de Internet invisible

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.

Google tiene capacidad para localizar una gran variedad de documentos en diferentes formatos (pdf, excel, word, access, flash, rtf, postscript, y muchos más). El último formato incorporado más destacable es el de los archivos .swf confeccionados en Flash.

A modo de ejemplo podemos realizar una consulta de contenidos que contengan la palabra museo en formato Flash. La expresión que se debe introducir en la caja de búsqueda es filetype:swf +museo. El formulario de búsqueda avanzada en Google sólo permite restringir las pesquisas a unos

pocos formatos. Es recomendable realizar la consulta desde la página inicial escribiendo en la caja de búsqueda mediante la sentencia filetype:* y el formato de archivo correspondiente obtendremos los resultados deseados.

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 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.

Bases de datos

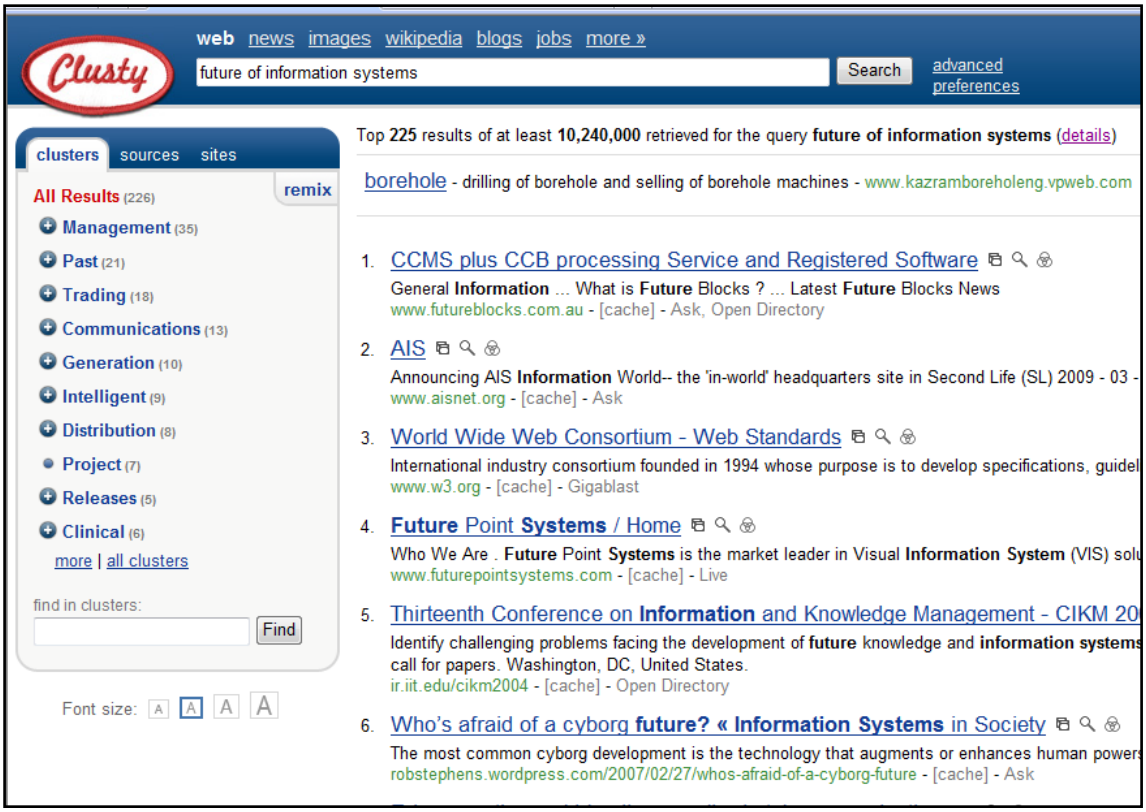
También tenemos indicios de solución al segundo gran "problema" de la Web invisible: el acceso al contenido de las bases de datos, pero desde motores convencionales.

La solución aquí proviene de este enfoque: si bien es difícil o imposible indexar por parte de los motores de búsqueda el contenido de bases de datos ajenas, no debería haber mucha dificultad en generar interfaces 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 http://clusty.com. Una búsqueda en Clusty 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 Clusty:



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

Pero lo que nos interesa aquí examinar es la siguiente idea: Clusty 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.

Multibuscadores de segunda generación

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 siguientes:

- 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 interfaz unificada de información a fuentes diversas.

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 interfaz de consulta universal.

Naturalmente, sindicación de contenidos implica también un modelo de negocio. Implica que los motores de búsqueda como Google 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; y (3) documentos a los que tiene acceso mediante pago con tarjeta de crédito.

La Web semántica

Definiciones

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

*"*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, 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.

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:

…Superar la brecha digital…

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.

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, indexar 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*.

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.

Volcadores, mapeadores y otras herramientas de localización de información

Las herramientas de búsqueda de Segunda Generación son programas cliente que automatizan procesos de localización, búsqueda y recuperación de información. Clasificación:

- Volcadores
- Multibuscadores
- Trazadores
- Indizadores
- Mapeadores
- Contenidos de la Infranet: catálogos de bibliotecas, bases de datos bibliográficas, obras de referencia, estadísticas y bases de datos numéricas, o bases de datos textuales. Los agentes de la Infranet son clientes Z39.50, con mecanismos para la realización automática de búsquedas de forma simultánea y que suele permitir el volcado de los registros. Entre los directorios más interesantes destacan:
- Directorio de recursos Z39.50, a nivel internacional. http://www.ilrt.bris.ac.uk/discovery/z3950/resources/
- Directorio español de recursos Z39.50. http://www.absysnet.com/recursos/recz3950.html. Especial mención merece Bookwhere, aplicación de búsqueda, recuperación y exportación de la información que usa el protocolo Z39.50 de Internet, y que tienen como objetivo facilitar el acceso a registros bibliográficos y a texto completo vía Internet. (Ver demo en http://www.webclarity.info/).
- Contenidos de la Web invisible: páginas huérfanas (sin conexión hipertextual); páginas no textuales (como ficheros multimedia y ejecutables); páginas con acceso mediante pasarelas (como páginas con palabra clave de acceso, ya sean gratuitas o de pago; depósitos de documentos; revistas electrónicas, etc.) o páginas dinámicas. Algunas direcciones que permiten acceso a los contenidos de Internet invisible son: www.internetinvisible.com www.completeplanet.com

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 indexar 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 interfaces 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: 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 en la más gigantesca base de datos que la humanidad pudiera haber soñado jamás.

Iniciativas de patrimonio digital

Proyecto de Carta de la UNESCO para la Preservación del Patrimonio Digital

Dice la UNESCO en su documento "DIRECTRICES PARA LA PRESERVACIÓN DEL PATRIMONIO DIGITAL":

Gran parte de la ingente cantidad de información que se produce en el mundo es de origen digital y existe en una gran variedad de formatos: texto, bases de datos, grabaciones sonoras, películas, imágenes. Para las instituciones culturales que tienen a su cargo el acopio y la preservación del patrimonio cultural, definir qué elementos deben conservarse para las generaciones futuras y cómo proceder en su selección y conservación, se está volviendo un problema apremiante. El enorme tesoro de información digital producida hoy día en prácticamente todas las áreas de las actividades humanas y concebida para ser consultada con computadoras, podría perderse si no se elaboran técnicas y políticas específicas para su conservación.

Podéis leer el texto completo en:

http://unesdoc.unesco.org/images/0013/001300/130071s.pdf

➔ F. Gestión del conocimiento y herramientas colaborativas

Ver la web http://kycabrera.wordpress.com/2008/08/11/gestion-del-conocimiento-y-herramientas-colaborativas/

El conocimiento se da cuando una persona conoce o sabe algo respecto a un tema, pero no cabe duda que los avances de la ciencia han contribuido a la formación del conocimiento. Como todos sabemos, el conocimiento se forma a través de la experiencia que adquirimos con el pasar de los años, ya que el conocimiento se forma a través de la información que tenemos al alcance de nuestras manos gracias a Internet, que nos ha ayudado mucho en el campo de la ciencia y tecnología, y en nuestra vida diaria.

Todos sabemos que la Wikipedia es una enciclopedia que está al alcance de todos en Internet, ya que es gratuita y podemos editarla y redactar algún artículo de algún tema del que tengamos conocimientos, y su principal ventaja es proporcionarnos información de manera más rápida y así podemos adquirir nuevos conocimientos. Esta herramienta permitiría que cada alumno, desde el lugar en que se encuentre, pueda investigar, redactar y publicar la información

que posea y, al mismo tiempo, leer los aportes que hicieron sus compañeros. Finalmente, una posterior edición de los contenidos permitiría crear una definición colectiva y probablemente mucho más rica (bajo el principio de inteligencia colectiva) que la que cada estudiante redactó individualmente.

WordPress comprende los blogs, que también son fuentes de información, ya que en ellos podemos subir información (postear). Esto se a convertido en una bitácora para estudiantes y profesores en el proceso educativo, ya que es un espacio para escribir preguntas, publicar trabajos o registrar enlaces hacia recursos relevantes. Actualmente, existen numerosas comunidades de blog educativas donde se intercambia información y conocimiento entre profesores y alumnos.

Este tipo de página web de estructura cronológica se ha convertido en el sistema de gestión de contenidos más popular de la Web 2.0 y uno de los favoritos de muchos profesores.

Esta herramienta es muy importante ya que podemos tener acceso a cualquier tipo de información, siempre y cuando la sepamos utilizar correctamente.

You Tube: aquí encontramos videos educativos y de entretenimiento.

Flickr: para compartir fotografías o imágenes.

Colaboratorios

Este tipo de plataformas se utilizan como repositorios para la educación, ya que permiten compartir objetos de aprendizaje que luego pueden exportarse a otras plataformas. Son también espacios de cooperación para el desarrollo de investigaciones. Los colaboratorios simplifican de manera notable el acceso e intercambio de insumos entre profesores, académicos y estudiantes, como si fuese una biblioteca o un laboratorio de libre acceso. Aquí se pueden compartir documentos científicos, proyectos, reportes, conferencias, *papers*, clases, tareas, estudios, bases de datos, entre otros.

Ventajas

- Una de las ventajas es que se rompe el modelo de *software* cerrado con derechos de uso y bajo el principio de la obsolencia planificada, para pasar al uso del *software* gratuito o libre.
- Los recursos en línea de la web 2.0 optimizan la gestión de la información, que se convierte en instrumentos que favorecen la conformación de redes de innovación y generación de conocimientos basados en cooperación y reciprocidad.
- El desarrollo de habilidades en los educandos estimula su interés por generar y compartir recursos multimedia de calidad.

Desventajas

- Una desventaja, que talvez puede ser la más importante, es que nosotros mismo nos creamos una dependencia al uso de la Internet.

Además, existe una polémica alrededor de la relevancia y pertinencia del termino Web 2.0, hasta el punto de ser cuestionado por muchos actores del propio entorno, que consideran que Web 2.0 es la denominación más apropiada para describir el nuevo tipo de aplicaciones web dominantes y la fase actual en la que se encuentra la red creada por Berners-Lee. El problema es que las noticias más leídas o votadas no son las más importantes, mientras que las que realmente nos van a servir de gran ayuda no son tomadas en cuenta por la mayor parte de usuarios.

- La propiedad intelectual es un tema importante, pero algunos usuarios de Wikipedia y blogs no saben a que se refiere este término y simplemente buscan y descargan la información que necesitan pero no citan el nombre del autor, y así se pierde el nombre o la identidad del autor de la información que tomamos y esto es perjudicial, ya que, si no citamos al autor original o la fuente de la información que estamos publicando o utilizando, nos pueden someter a sanciones económicas o privarnos de nuestra libertad.

Crowdsourcing

Es un concepto que surge como resultado de aprovechar la arquitectura social de la Web 2.0, los crecientes niveles de participación mediatizada y el poder de la inteligencia colectiva, cuya suma se ha convertido en una fuente de ideas y desarrollos para el sector empresarial e incluso para el campo de la experimentación científica. Además, se refiere a otra forma de emplear mano de obra barata, gracias a la popularidad y ubicuidad de Internet, en la que personas no especializados, resuelven problemas para toda clase de compañías que utilizan el potencial de los millones de cerebros de la multitud que se conecta a través de la red. Crowd es el término en inglés de multitud y sourcing se refiere a la obtención de materia prima (donde source es el término en inglés de fuente, en este caso de un proyecto).

Tipos de Trabajo Masivo

La wiki es la más conocida de crowdsourcing:

- Worthidea es una plataforma multicultural y multilingüe de ideas que sirven como punto de encuentro entre las empresas y los usuarios.
- Procter and Gamble emplea más de 9000 científicos e investigadores en su corporación R&D y aún tienen muchos problemas que no pueden solucionar. Ellos ahora publican sus terribles dolores de cabeza en un sitio llamado InnoCentive, ofreciendo grandes sumas de dinero a más de 90.000 “solucionadores”, quienes tienen su red de científicos de apoyo. P&G también trabaja con NineSigma, YourEncore y Yet2.
- Amazon realiza algo similar con proyectos de sotfware a gran escala.

- iStockphoto es un sitio con más de 22.000 fotógrafos amateur, quienes suben y distribuyen stock's fotográficos. Debido a que no tiene los altos costos de un equipo profesional como Getty Images, es posible comprar imágenes por un bajo costo. Esta empresa fue comprada por Getty Images.
- CambriaHouse es un headquarter de iStockphoto en Calgary, Canadá, y se describe como: código abierto con dinero. Es una incubadora que descubre y comercializa software e ideas a través del crowdsourcing. Los contribuyentes ganan regalías y comparten las ganancias del producto.
- Portucuenta es un portal de programadores freelance en español, que permite desarrollar proyectos de cualquier escala al contar con un *pool* de programadores disponibles a los que se les puede asignar distintas tareas para luego ser integradas.

Open Innovation

Open Innovation es un término promovido por Henry Chesbrough, profesor y director ejecutivo en el Centro de Innovación Abierta en Berkeley.

La idea central detrás de *Open Innovation* es que en un mundo de conocimientos de amplia distribución, las empresas no pueden darse el lujo de depender enteramente de su propia investigación, sino que deben comprar licencias de procesos o invenciones (patentes) de otras empresas. Además, las invenciones internas que no son usadas deberían ser llevadas afuera, por ejemplo a través de la concesión de licencias, las empresas mixtas, las *spin-offs*, etc.

Ventajas

- *Open Innovation* parte de la premisa de que la información y el conocimiento son abundantes y están ampliamente distribuidos.
- En los modelos anteriores el lugar de la innovación era la empresa. *Open Innovation* entiende que los actores internos y externos tienen un papel similar.
- El papel central del modelo de negocio en todo el proceso. En estructuras anteriores el modelo de negocio tenía un papel secundario en el proceso de innovación. En *Open Innovation* el modelo de negocio tiene un papel dual: a) la selección de productos y servicios por los que apostar, y b) la búsqueda y la creación activa de modelos que permitan comercializar aquellas ideas que no encajan en el modelo de negocio actual.
- *Open Innovation* considera proyectos aunque no encajen en el modelo de negocio. Estos proyectos pueden ser relevantes ya sea porque se dirigen al propio mercado o a mercados potenciales donde podemos capturar valor.
- *Open Innovation* entiende la innovación como un proceso global, de esta manera las unidades de negocio no sólo compiten internamente sino también con el exterior.

- Un papel proactivo de la gestión de la IP (sigla inglesa de propiedad intelectual) a través de licencias, licencias cruzadas o incluso donación de patentes.
- Un conjunto de métricas nuevas en la evaluación del proceso de innovación, en consonancia con el cambio de *locus* y la comprensión global del proceso que proporciona el nuevo modelo (actividades de innovación fuera de la empresa, número de partnerships, número de spin-offs, etc.).

Tipologías de herramientas colaborativas

Esta clasificación se hace tomando como referencia la clasificación realizada por McGreal, Gram y Marks, quienes tuvieron en cuenta trabajos realizados con profesionales de la educación:

- Herramientas para la gestión y administración académica: se gestionan asuntos como la gestión de la matrícula e inscripción de los alumnos en los cursos, proporcionar información académica como horarios, fechas de exámenes, notas, planes de estudios, expedición de certificados, concretar reuniones, tutorías, etc. En la Universidad de Antioquia, por ejemplo, esto se logra a través de MARES.
- Herramientas para la creación de materiales de aprendizaje multimedia: aquí se pueden encontrar aquellos programas que son utilizados para la creación de los contenidos de aprendizaje como: los editores de páginas web como HTML o las que facilitan la creación de ejercicios de auto evaluación, simulaciones, o prácticas, como la creación de wikis o la realización de mapas conceptuales en línea.
- Herramientas para la comunicación y el trabajo colaborativo: en este grupo se pueden encontrar aquellas que facilitan la comunicación a través de un ordenador entre alumno-profesor, como el correo electrónico, los chats, las conferencias electrónicas, audio conferencias, las videoconferencias, la pizarra compartida, aplicaciones compartidas o documentos compartidos.
- Herramientas integradas para la creación y distribución de cursos a través de la WWW. Desarrolladas específicamente para propósitos educativos. En este grupo se pueden encontrar todas aquellas plataformas que ofrecen cursos en línea por ejemplo.

Otros tipos:

- E-mail: sirve para recibir y enviar mensajes en forma casi inmediata.
- Backpack: es una aplicación para la organización del material con que se cuenta para elaborar los proyectos compartidos. La información sobre esta aplicación se puede encontrar en http://www.backpackit.com/.
- Vyew: sirve para crear una sala de Chat con un número de sala que no debe sobrepasar los cuatro dígitos, este número se comparte con aquellas personas con las cuales se pretende realizar el proyecto. En esta herramienta se pueden crear iconos de organización y

almacenamiento para la información que se vaya produciendo. Para ejemplos visitar http://vyew.com/.

- ➔ G talk: esta es una herramienta creada por Google que sirve para el envió y recibo de mensajes instantáneos entre dos o más personas. Sólo es necesario contar con una cuenta de Gmail. Para observar más beneficios o aplicaciones de esta herramienta se puede ingresar en http://www.google.cat/talk/.
- ➔ Weblogs o bitácoras: esta es una página web con apuntes fechados en orden cronológico inverso, para que así el usuario pueda encontrar en primer lugar las últimas publicaciones.
- ➔ Wikis: un wiki es un ejemplo claro y preciso de lo que es trabajo en grupo, ya que es una herramienta creada y mantenida por varios autores, lo que los diferencia de los weblogs que no pueden ser modificados sino por el autor original.
- ➔ Redes sociales: las redes sociales también son conocidas como *software* sociales

En la actualidad, todas estas herramientas están siendo incorporadas en las unidades de información de una manera vertiginosa, claro está, cada institución va a su ritmo dependiendo de los recursos económicos y tecnológicos con los que se cuenta. Ahora bien, independiente de que se cuente o no con los recursos suficientes para que las unidades de información trabajen con las herramientas colaborativas, es necesario que se empiece a interactuar con éstas, ya que es una manera muy práctica para que la unidad de información se retroalimente con las inquietudes y aciertos de los usuarios en línea. Además de lo anterior, las herramientas colaborativas les permiten a las unidades de información emprender proyectos con otras y así ampliar más sus perspectivas, lo que redundará en beneficios para sus usuarios tanto reales como potenciales. Personalmente recomendaría los wikis, los weblogs y los backpacks, al menos mientras se empieza a profundizar en otras herramientas.