# Agentes inteligentes en la búsqueda y recuperación de información

**Pablo Lara Navarra**

**José Angel Martínez Usero**

**PLANETA UOC**

Primera edición: julio 2004

Segunda edición, revisada y ampliada: julio 2006



Av. Tibidabo, 39-43, 08035 Barcelona

ISBN 84-9707-571-4

# Indice de contenidos



ISBN 84-9707-571-4

ISBN 84-9707-571-4

# Introducción

Los mayores motores de búsqueda apenas cubren un 20-25% del web, mientras que los principales índices es dudoso que lleguen a un 5%. La desventaja de este ingente volumen de información es que, por razones de celeridad en la respuesta, obliga a limitar las prestaciones de búsqueda de forma que suelen faltar ciertas capacidades avanzadas. Otros problemas importantes derivan de la diferente cobertura de la red (las sedes comerciales y de los países desarrollados están mejor indizadas), el elevado porcentaje de enlaces no activos y la desactualización de los recursos debido a frecuencia de revisión muy baja o inadecuada.

Las herramientas de motor de búsqueda están instaladas en el ordenador remoto y por tanto limitadas por restricciones generalmente ajenas al usuario final. Una nueva generación de herramientas y la adopción de nuevas estrategias pueden ayudar significativamente, así como el reconocimiento de nuevas realidades y el descubrimiento de fuentes ocultas de datos relevantes hasta la fecha frecuentemente infrautilizados.

Las herramientas de segunda generación, instaladas en el ordenador cliente son capaces de tratar con grandes volúmenes de información, automatizando tareas que incrementan la productividad final de los recursos recuperados.

ISBN 84-9707-571-4

# 1. Los motores de búsqueda y la recuperación de la información

La recuperación de la información (RI) es una operación en la que se interpreta una necesidad de información de un usuario y se seleccionan los documentos más relevantes capaces de solucionarla. En el contexto de Internet, se puede definir el objetivo de la recuperación como la identificación de una o más referencias de páginas web que resulten relevantes para satisfacer una necesidad de información.

## 1.1. El lenguaje de interrogación

Un lenguaje de interrogación es el conjunto de opciones (órdenes, operadores y estructuras) que, organizados según normas lógicas, permiten la consulta de los recursos de información mediante una expresión, llamada ecuación de búsqueda.

- Las órdenes son aquellas palabras o abreviaturas que indican al sistema las acciones a ejecutar (buscara la expresión, mostrar los registros resultantes de una búsqueda, ejecutar un perfil de usuario...)

- Los operadores son los encargados de expresar las relaciones que mantienen entre sí los términos que pueden definir las necesidades informativas del usuario.

Si bien inicialmente las ecuaciones de búsqueda se formulaban mediante la formulación textual de expresiones, la implantación de interfaces gráficas a partir de los años 80 llevó al uso de nuevos entornos de selección, donde el usuario sólo debe introducir los términos y guiarse por un sistema de botones y menús desplegables.

### 1.1.1. Operadores lógicos o booleanos

Llamados así en honor a George Boole, matemático del siglo XIX que fue el precursor de la lógica simbólica y el álgebra de Boole (teoría de conjuntos), es uno de los métodos más extendidos de especificar las búsquedas en la mayoría de sistemas. Se basan en tres operaciones lógicas básicas:

- Intersección de conjuntos :AND/ Y Operador que indica que deben estar incluidos en los resultados de la búsqueda los términos unidos por esta partícula. Es un operador restrictivo, puesto que elimina aquellos documentos en los que no aparecen todos los términos de la expresión de búsqueda.

ISBN 84-9707-571-4

Ejemplo: bullseye AND copernic, indica que deben aparecer en el documento las dos palabras si no es así se excluirá el documento.


- Unión o suma de conjuntos : OR / O Indica que cualquiera de las palabras que estén unidos por este operador debe aparecer en el documento, las restantes no tienen que estar presentes. Es un operador de ampliación, pues sólo deberá aparecer uno o alguno de los términos de la expresión de búsqueda.

Ejemplo: bullseye OR copernic, puede aparecer en el documento la palabra bullseye o copernic o ambas


- Exclusión de conjuntos: NO/ AND NOT Operador que excluye de un documento la palabra no deseada. Es un operador de restricción, pues se seleccionan aquellos documentos que contienen el primer término de búsqueda, pero no el segundo.

Ejemplo: Knowbots AND NOT copernic, recupera todos los documentos que contengan la palabra Knowbots pero que no contengan la palabra copernic.


En la elaboración de una ecuación de búsqueda es habitual la combinación de más de uno de estos operadores, por lo que será necesario conocer en profundidad el sistema para saber las prioridades a la hora de su ejecución, puesto que los

ISBN 84-9707-571-4

resultados pueden variar sustancialmente. A menudo, estas prioridades vienen marcadas por el uso de paréntesis, de manera que se ejecuta en primer lugar el operador que une los términos que están entre paréntesis.

> Ejemplo: (bullseye OR copernic OR lexibot) AND (agentes inteligentes), recupera los documentos que contengan los terminos agentes inteligentes y copernic o bullseye o lexibot.

### 1.1.2. Operadores posicionales

Los operadores posicionales toman como partida la posición del término en relación a su contexto, es decir, en relación a los otros términos y al documento. Estos operadores se pueden dividir en dos tipos: los relativos y los absolutos.

#### 1.1.2.1. Operadores posicionales relativos

A menudo llamados operadores de adyacencia o proximidad. Permiten definir al sistema de búsqueda la distancia que puede existir entre un término y otro. Se pueden buscar términos que estén juntas, separadas por varias palabras o caracteres, que se encuentren en una misma frase o un mismo párrafo, e incluso si se debe o no respetar el orden de los términos. Existe una gran variedad de operadores de adyacencia, y expresan diferentes situaciones según los sistemas.

- NEAR, operador que obliga a estar a un número determinado de distancia las palabras claves a recuperar. Este número varía en función de los diferentes programas de recuperación de la información: así, por ejemplo, mientras en Altavista significa un máximo de 10 palabras entre los términos, en WebCrawler significa un máximo de 2 palabras.

  Ejemplo: bullseye NEAR copernic recupera textos con frases como “bullseye es mejor que copernic” o “copernic tiene más motores que bullseye”

- NEAR/N, realiza la misma operación que NEAR, pero N es sustituido por la distancia en palabras que deben estar separados los términos de búsqueda.

  Ejemplo: bullseye NEAR/5 copernic, recupera todos los documentos que aparezcan los dos terminos y cuya separación no sea mayor a cinco palabras.

Otra posibilidad es hacer una búsqueda de una frase exacta. Consiste en la intersección de las palabras de búsqueda que además están adyacentes y en

ISBN 84-9707-571-4

el orden en que se describen.

" ", emplear las comillas expresa, que debe aparecer la frase exacta y en el mismo orden.

Ejemplo: "comparación de agentes inteligentes", tiene que aparecer esta frase en los documentos para que sean recuperados.

#### 1.1.2.2. Operadores posicionales absolutos

Se trata de operadores que permiten buscar el o los términos en un lugar determinado del documento. En general, son operadores delimitadores de un campo.

- Link: recupera todos los links que contenga el término buscado.

Ejemplo: Link:"agentes inteligentes", recupera todos los links que contenga la frase exacta agentes inteligentes.

- Title, recupera en los títulos de web, correos, etc., la palabra/s deseadas.

Ejemplo: Title:"agentes inteligentes", recupera únicamente del título la frase exacta.

- Url, busca url que contengan los términos de la ecuación de búsqueda.

Ejemplo: Url:"ugr.es", presenta todos las páginas web de la Universidad de Granada.

- Body, recupera del cuerpo del documento el conjunto de palabras deseadas.

Ejemplo: Body:"agentes inteligentes", recupera del cuerpo del documento únicamente la frase exacta de la ecuación de búsqueda.

### 1.1.3. Operadores de truncamiento y de límite/comparación

- Operadores de comparación o de rango. Limitan la búsqueda mediante una expresión que establece un rango de valores, especialmente numéricos. Corresponden a formas tipo "igual que"(simbolizado por =, EQ), "mayor que" (simbolizado por >, GT), "menor que"( simbolizado por <, LT) o operadores de intérvalos (simbolizado por un guión, to, ><)

ISBN 84-9707-571-4

- Operadores de truncamiento o máscaras. Los truncamientos ayudan a buscar todas las posibilidades semánticas de un término, por ejemplo sus derivados, fijados por prefjiación o sufijación, las variantes léxicas. Así, los llamados caracteres comodín, como el asterisco (*) o la interrogación (?) sustituyen un carácter o un conjunto de caracteres.

  Ejemplo: document*, recupera todas las palabras que contengan esta raiz por ejemplo, documento, documentos, documentalista, documentado, etc.

## 1.2. Las herramientas de recuperación de información web

En España existen diferentes imprecisiones terminológicas a este respecto. Por un lado, a los motores de búsqueda se les ha denominado con otros términos sinónimos, tales como: buscadores, rastreadores, webcrawlers, agentes, índices, directorios. Por otro, durante cierto tiempo, se han confundido tres tecnologías que ahora tienen autonomía propia: los índices temáticos/directorios, los motores de búsqueda y los agentes inteligentes.

En principio, la diferencia entre motor de búsqueda e índice temático o directorio parece clara. Un índice temático es una página web (sitio web) en donde, las distintas materias se encuentran organizadas en torno a un conjunto de epígrafes. Esta diferencia se tambalea cuando nos encontramos con índices temáticos como Yahoo (www.yahoo.com), que presenta un interfaz similar a los motores e incluso permite realizar búsquedas sobre los recursos que tiene sistematizados. En general, la diferencia radica en el hecho de que los índices temáticos contienen direcciones que son recopiladas, organizadas y clasificadas manualmente y la búsqueda se lleva a cabo exclusivamente sobre los recursos indexados del directorio.

Los agentes inteligentes pueden realizar una serie de tareas sin que los humanos u otros agentes les tengan que decir qué hacer a cada paso que dan en su camino. Se diferencian de los motores de búsqueda en que éstos albergan contenidos estáticos (aunque se actualizan con cierta frecuencia) y responden directamente a las peticiones de los usuarios. Si un motor de búsqueda pudiera almacenar peticiones de los usuarios y notificarles la llegada de información útil, entonces el motor de búsqueda sería un agente. Sin embargo, la diferenciación no es radicalmente clara, puesto que se denominan agentes inteligentes a softwares que realmente no lo son.

ISBN 84-9707-571-4

### 1.2.1. Tipos de herramientas de búsqueda y recuperación

A continuación se presenta una definición estándar de cada una de las herramientas de búsqueda y recuperación de información mencionadas anteriormente.

#### 1.2.1.1. Los directorios o índices temáticos

Los directorios o índices presentan una selección de recursos webs organizados siguiendo una estructura o clasificación jerárquica de materias que va de categorías más amplias a categorías más específicas. Los directorios se exploran mediante la navegación (browsing) de una base de datos de documentos web compilados, recogidos y organizados manualmente por expertos (ayudados por robots de localización automática de recursos en la red). La búsqueda jerárquica sirve al usuario de guía, permitiendo acceder a la información en el contexto temático al que pertenece y en relación a otras áreas temáticas.

Los directorios también presentan un motor de búsqueda interno para localizar directamente recursos de la base de datos, mediante diferentes ecuaciones de búsqueda y palabras clave, obviando de esta manera el uso del directorio temático. Los sistemas de búsqueda por palabras pueden actuar de dos maneras:

- Sobre la clasificación, en una sección de ella (cuando, por ejemplo, sabemos en qué parte del directorio podemos localizar la información que nos interesa)
- Sobre las páginas, pero en este caso se limitan a la información recopilada por el índice (fundamentalmente sitios web, no páginas)

Así pues, la búsqueda de información en los directorios puede hacerse bien de forma guiada, mediante clasificaciones jerárquicas, bien a partir de términos específicos.

Los directorios más comunes son aquellos que ofrecen una navegación por temas, y con una cobertura generalista, como por ejemplo Yahoo! (Yet Another Hierarchical Officious Oracle). Sin embargo, también existen directorios que permiten, por ejemplo, una navegación geográfica (Virtual Tourist http://www.virtualtourist.com ) o directorios especializados.

Los servicios de consulta basados en directorios han ido incorporando prestaciones, y han evolucionado hacia lo que actualmente se llaman portales, un conjunto de

ISBN 84-9707-571-4

servicios que pretende satisfacer todas las necesidades de los usuarios de Internet (cuentas de correo electrónico, chat, páginas amarillas y blancas, información metereológica y de la bolsa, servicio de noticias).

#### 1.2.1.2. Los motores de búsqueda

Los motores de búsqueda o buscadores tienen sus antecedentes en los simples listados de direcciones de recursos y documentos de la red, y son la respuesta al rápido volumen de crecimiento dela red, que supera la capacidad de los recursos humanos de los directorios – que por ello suelen ser selectivos -. Los buscadores son bases de datos creadas por indización automática del texto completo de las páginas web, y realizada por un programa llamado robot. Este robot lógico, o araña (spider), explora de forma automática los servidores, extrayendo las palabras más significativas de cada página y creando un índice de búsqueda. Aun cuando los programas lleguen a ser similares, no existen dos programas de búsqueda exactamente similares en términos de tamaño, velocidad y contenido; no existen dos motores de búsqueda que utilicen coincidentemente el mismo listado de relevancia y tampoco cada motor de búsqueda ofrece las mismas opciones de búsqueda. Por lo tanto, su búsqueda resultará diferente en cada motor utilizado. La diferencia podría no ser mucha, pero sí significativa

Existe una gran porción de la red que las “arañas” de los buscadores no pueden o no alcanzan a indizar. Se las nombra como la "Red Invisible " o la "Red profunda" e incluye, entre otras cosas, sitios protegidos por contraseñas, documentos detrás de “cortinas de fuego”, material archivado, herramientas interactivas, y los contenidos de ciertas bases de datos.

Los servicios de consulta basados en directorios y motores de búsqueda han ido incorporando prestaciones, y han evolucionado hacia lo que actualmente se llaman portales, un conjunto de servicios que pretende satisfacer todas las necesidades de los usuarios de Internet (cuentas de correo electrónico, chat, páginas amarillas y blancas, información metereológica y de la bolsa, servicio de noticias).

#### 1.2.1.3. Los agentes inteligentes

Un agente es una entidad autónoma capaz de almacenar conocimiento sobre sí misma y sobre su entorno, con unos objetivos y capacidad. Asimismo, 8n agente inteligente es un programa que basándose en su propio conocimiento, realiza un conjunto de operaciones para satisfacer las necesidades de un usuario o de otro programa, bien por iniciativa propia o porque alguno de estos se lo requiere.

ISBN 84-9707-571-4

Según las leyes de la inteligencia artificial debe de tener las siguientes características:

- Autonomía: Debe actuar sin ningún tipo de intervención humana directa y tener control sobre sus propios actos.
- Sociabilidad / Comunicatividad: Debe de ser capaz de comunicarse mediante un lenguaje común con otros agentes e incluso con los humanos.
- Capacidad de reacción: Percibe su entorno y reaccionar para adaptarse a él (por ejemplo ante una palabra mal escrita determinar qué es a través del contexto)
- Iniciativa: Emprende las acciones necesarias para resolver un problema.

Tipologías de herramientas de segunda generación (agentes inteligentes)

- Clientes z39.50. Permiten la consulta simultánea de un elevado número de servidores, mediante un único protocolo, es decir, un único interfaz y lenguaje de interrogación. Es especialmente útil en recuperar la información que se encuentra en la llamada "Internet invisible", información que no es indizada por los motores de búsqueda – por ejemplo, las bases de datos -.

- Volcadores. Permiten volcar automáticamente una copia idéntica de sedes, directorios y documentos, manteniendo su estructura y sus elementos – incluso los enlaces - , y creando así un archivo offline. Se puede programar la hora del volcado, reduciendo considerablemente el tiempo y el coste, y permite activar el vuelco de diferentes tipos especiales de documentos ( .html, .doc, .pdf, .gif )

- Mutibuscadores o metabuscadores. Permiten realizar la recuperación de la información en varios motores de búsqueda simultáneamente. A diferencia de los multibuscadores de primera generación, la mayoría de las tareas pueden automatizarse y son muy flexibles en su configuración: traducen expresiones en lenguaje natural, envían los perfiles a varios motores de búsqueda y procesan los resultados, eliminando los duplicados, y ordenando los contenidos según criterios y formatos definibles.

- Trazadores. Permiten la búsqueda en las páginas enlazadas desde una página web determinada o desde una lista de resultados de un buscador. Desde esta primera sede, llamada "semilla", y aprovechando la naturaleza hipertextual de Internet, van comprobándose las páginas que se encuentran enlazadas según una serie de criterios de pertinencia, y así sucesivamente hasta un nivel prefijado. Aunque generan mucho ruido y es una técnica lenta, permite recuperar información que es imposible de localizar para los

ISBN 84-9707-571-4

buscadores.

- Indizadores Permiten indizar y resumir automáticamente diferentes páginas web, y exportar los resultados en diferentes formatos reutilizables por editores web.

- Mapeadores. Describen íntegramente una sede, detallando cada fichero y directorio, y proporcionando un mapa de contenidos. Permiten obtener datos numéricos que ayudan a evaluar dichos contenidos y establecer una comparativa entre diferentes sedes web – en base a valores como el tamaño, la densidad hipermedia de la sede, su estructura de niveles, la tipología de enlaces, etc.

### 1.2.2. Funcionamiento de los motores de búsqueda

Un motor de búsqueda está formado por cuatro elementos básicos:

1. Un programa (también denominado robot, rastreador o webcrawler) que recorre el WWW buscando recursos de información y sus respectivas URLs.
2. Un sistema automático de análisis de contenidos e indexación de los documentos localizados por el robot.
3. Un sistema de interrogación, generalmente basado en la lógica booleana, que permite al usuario expresar su necesidad de información.
4. Un programa que actúa de pasarela entre el servidor de documentos html y la base de datos.

Funcionamiento: el motor de búsqueda recibe la consulta del usuario (query), formada por uno o más términos, realiza una consulta interna en la base de datos que contiene los recursos web indexados y ofrece una lista de aquellos recursos que cumplen una parte o el total de los requisitos establecidos en la consulta. Generalmente, los resultados aparecen ordenados según una puntuación (score) que el programa asocia automáticamente a cada recurso.

Para realizar una consulta es necesario tener en cuenta un conjunto de variables:

1. Lenguaje de interrogación, que debe ofrecer diferentes tipos de operadores: lógicos, de comparación, de truncamiento, de proximidad, de especificación de campo.

ISBN 84-9707-571-4

2. Posibilidad de refinar (refine) una búsqueda inicial.
3. Campos limitadores que nos permitan reducir la búsqueda: dominios, lenguas, países, fecha de creación del recurso.
4. Búsquedas alternativas: búsqueda simple, búsqueda avanzada, búsquedas combinando operadores e índices temáticos, etc.
5. Opciones avanzadas: buscar diferentes recursos (texto, sonido, imagen), guardar y reutilizar búsquedas, diferentes formatos en los resultados de búsqueda (estándard, detallado, compacto, etc.), búsqueda de conceptos relacionados (related topics), consulta directa en bases de datos (infranets), etc.

### 1.2.3. Los metabuscadores

La gran cantidad de información y el notable aumento de motores de búsqueda accesibles desemboca en la necesidad de realizar consultas simultáneas en diferentes motores de búsqueda y con una sola estrategia (query). De esta necesidad surgen los denominados "metabuscadores", que ofrecen nuevas prestaciones y mejores y más exhaustivos resultados de búsqueda.

Los metabuscadores permiten automatizar el proceso de realizar una misma consulta en diversos motores de búsqueda, lo cual no significa que sea totalmente exhaustivo, puesto que el metabuscador envía la consulta solamente a aquellos motores de búsqueda con los que ha establecido un acuerdo previo.

En el funcionamiento de los metabuscadores cabe destacar algunas variables interesantes. Por una lado, la exhaustividad no está garantizada (desde el momento en el que sabemos que un motor de búsqueda es capaz de indexar a lo sumo un 30% de los recursos web disponibles). Por otro, los tiempos de respuesta pueden ser mucho más largos dada la necesidad de realizar múltiples búsquedas simultáneas (Metacrawler permite delimitar el tiempo máximo de espera de 1 a 10 minutos), además, la recuperación de recursos duplicados suele ser muy elevada, por ello, algunos metabuscadores ya han implementado la utilidad que permite eliminar los duplicados.

### 1.2.4. Tendencia actual de los motores de búsqueda

Partiendo de los problemas actuales que presentan los motores de búsqueda en cuanto a su funcionamiento y los resultados ofrecidos se pueden adelantar algunas vías de solución futura que marcan la tendencia en la evolución de los motores de búsqueda.

ISBN 84-9707-571-4

Los resultados de la búsqueda pueden ser satisfactorios o no tanto. Los motores de búsqueda ofrecen resultados muy diferentes ante una misma cuestión inicial; este hecho demuestra la poca exhaustividad de los motores en la indexación de los recursos web y pone de manifiesto los problemas derivados de la escasez de control lingüístico.

Actualmente se aboga por la incorporación de herramientas de análisis lingüístico y control terminológico en los motores de búsqueda, de forma que sea posible efectuar una recuperación menos ligada a la comparación de cadenas de caracteres y más vinculada a la comparación de conceptos.

La escasa calidad de la información recuperada es otro inconveniente de los actuales motores de búsqueda. Los mecanismos para aumentar la precisión en la búsqueda (refinamientos, búsquedas avanzadas, acotación por dominios, etc.), a veces, no funcionan como cabría esperar. A ello, hay que añadir el mínimo valor de algunos de los sitios web recuperados, el porcentaje de recursos repetidos y el porcentaje de recursos inactivos (que ya no existen físicamente en la red aunque continúan indexados).

En este sentido se empieza a hablar de una Internet para el gran público y una Internet de los recursos culturales, científicos y técnicos. La especialización de los motores de búsqueda es una buena vía para conseguir mejores servicios de información, la especialización conduce a la concentración del conocimiento en ciertos lugares donde los usuarios pueden encontrar fácilmente los recursos relacionados con su ámbito de conocimiento.

## 1.3. La Infranet o Internet invisible

La infranet o Internet invisible es el conjunto de recursos accesibles únicamente a través de algún tipo de pasarela o formulario web y que, por tanto, no pueden ser indizados de forma estructural por los robots de los motores de búsqueda.

Muchos de estos recursos son de gran calidad y, desde el punto de vista del gestor de información, tienen una importancia clave en la recuperación de información de alto valor añadido. Su invisibilidad para los motores de búsqueda implica una dificultad considerable para la recuperación efectiva de estos recursos y requiere aproximaciones novedosas por parte de los profesionales.

### 1.3.1. Los recursos de la Internet invisible

La propia heterogeneidad formal de la información en Internet puede plantear dificultades a la hora de entender qué recursos están incluido bajo la denominación

ISBN 84-9707-571-4

de Internet Invisible. Con el fin de clarificar qué contenidos pueden resultar invisibles, se ha considerado una clasificación que atiende a criterios documentales.

**Bases de datos bibliográficas**: se incluyen en este grupo los catálogos de biblioteca accesibles a través de una pasarela (OPAC) web, otras bases de datos de referencias bibliográficas (de acceso público o restringido –registro previo gratuito o de pago-) y entidades similares tales como los catálogos de librerías (ej.: Amazon.com)

**Bases de datos alfanuméricas**: definidas por exclusión del grupo anterior, son todas las bases de datos que no tienen carácter bibliográfico. Comprendería, además, los recursos llamados de referencia que requiren algún tipo de pasarela de acceso para su consulta (ej.: Encyclopaedia Britannica).

Una situación ligeramente distinta es la planteada por las páginas generadas dinámicamente (asp, jsp, php o similares). Dichas páginas sólo existen en virtud de una consulta puntual, imposible de realizar por los robots de los motores de búsqueda y cuyo contenido puede alcanzar un considerable grado de personalización. Desde un punto de vista documental, los contenidos que explotan están en una base de datos y, por tanto, se consideran dentro de esta categoría.

**Archivos y revistas electrónicas**: se trata de bases de datos que incluyen documentos a texto completo y que sólo se pueden recuperar previa identificación de la referencia, labor para la que se requiere utilizar una pasarela web simple (formulario de consulta) o doble (palabra de acceso y formulario de consulta).

**Ficheros no HTML o textuales**: el (relativo) fracaso de HTML original a la hora de generar páginas con una maquetación muy rica, ha permitido que algunos formatos de diseño más elaborado hayan adquirido popularidad en la web (pdf, ps, ppt, doc). Estos formatos no textuales no son indizados correctamente por los robots de los motores de búsqueda (excepción: Google.com ya indiza documentos pdf y los convierte en HTML) y, por tanto, constituyen una parte del webespacio invisible que ha ido adquiriendo cada vez más importancia.

### 1.3.2. La recuperación de la información en la Internet invisible

La recuperación de la información en la Internet invisible se apoya fundamentalmente en la disponibilidad de directorios e índices que identifiquen y organicen su principales recursos. El más importante en el mercado español es la sede española de Internet Invisible http://www.internetinvisible.com .

ISBN 84-9707-571-4

Las herramientas más importantes para la recuperación de estos recursos de información invisibles son:

**Clientes Z39.50**: La progresiva adopción del protocolo Z39.50 por la mayoría de las bibliotecas está incrementando el valor de los clientes Z39.50, que ya pueden acceder a una masa crítica de recursos.

| | |
|---|---|
| Bookwhere 2000 | Sea Change (www.bookwhere.com) |
| EzCat | BookSystems (www.booksys.com/ezcat) |
| ZNavigator | EnWare (www.enware.es) |
| ZSearch | Infoworks Technology (www.itcompany.com) |
| ZPista | Ifgenia Plus (www.ifigenia.es/zeta) |

**Agentes inteligentes**: estos programas se han convertido en herramientas muy populares en Internet, que permiten superar algunos de los problemas tradicionalmente asociados a los motores de búsqueda. No todos los agentes ofrecen acceso a recursos de la Internet invisible e incluso aquellos que así lo hacen lo presentan como opciones avanzadas, normalmente no disponibles en las versiones "shareware".

| | |
|---|---|
| BullsEye | Intelliseek (www.intelliseek.com) |
| Copernic | Copernic (www.copernic.com) |
| EZSearch | American Systems (www.americansys.com) |
| LexiBot | BrightPlanet (www.lexibot.com) |
| WebSeeker | Blue Squirrel (www.bluesquirrel.com) |

## 1.4. Bibliografía

Aguillo Caño, Isidro (1999). Del multibuscador al metabuscador: las agentes

ISBN 84-9707-571-4

trazadores de Internet. En*: Congreso ISKO* (IV. Granada. 1999). La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de información. Granada: Isko: Universidad de Granada, p.239-245

Aguillo Caño, Isidro (2001). Internet invisible o Infranet: definición, clasificación y evaluación. En: Maldonado Martínez, Angeles. La información especializada en Internet. Madrid: CSIC, p. 161-178

Codina, L (1997). Cómo funcionan los servicios de búsqueda en Internet: un informe especial para navegantes y creadores de información. Part I. *Information World en Español*, vol. 6, nº 5, p. 22-26

Codina, Lluis (2003). Internet invisible y web semántica: ¿el futuro de los sistemas de información en línea?. *Revista tradumática*, nº 2, 2003.

Cornella, Alfons (2001). *Mensaje 364 de Extra!-Net, la revista de infonomía* La infranet, ¿dónde esta el valor?.

Fernández de las Heras, José Manuel; Díaz de Cerio, Pako (2000). La Intranet del conocimiento. VII Jornadas Españolas de Documentación: La Gestión del Conocimiento, retos y soluciones de los profesionales de la información. Bilbao 19-20-21 octubre 2000. p. 567-574

Gómez Díaz, Raquel (2003). La evaluación en recuperación de la información». Hipertext.net, nº 1 (mayo 2003). http://www.hipertext.net/web/pag238.htm

Marcos Mora, María del Carmen (2005). Elementos visuales en sistemas de búsqueda y recuperación de la información. *Hipertext.net*, nº 3 (mayo 2005). http://www.hipertext.net/web/pag257.htm

Martínez, Francisco J.; Rodríguez Muñoz, José Vicente (2004). Reflexiones sobre la evaluación de los sistemas de recuperación de la información : necesidad, utilidad y viabilidad. *Anales de documentación*, núm. 7 (2004), p. 153-170. http://www.um.es/fccd/anales/ad07/ad0710.pdf

Tramullas, J (1999). Agentes y ontologías para el tratamiento de la información: clasificación y recuperación en Internet. En: *Congreso ISKO* (IV. Granada. 1999). La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de información. Granada: Isko: Universidad de Granada, p. 247-248

Vaquero, J.R (1997). Motores de búsqueda. *Information World en Español*, vol. 6, nº 7-8, p. 31-32

Vidal Bordés, Francisco Javier; Salvador Oliván, José Antonio (2000). *La*

ISBN 84-9707-571-4

*implementación de metadatos y Dublin Core en sedes y páginas web de bibliotecas y centros de documentación de universidades y centros de investigación de la Red IRIS*. VII Jornadas Españolas de Documentación: La Gestión del Conocimiento, retos y soluciones de los profesionales de la información. Bilbao 19-20-21 octubre 2000. p. 197-210.

ISBN 84-9707-571-4

## 1.5. Casos prácticos

### 1.5.1. Caso práctico 1. Evaluación de motores de búsqueda

PRESENTACIÓN:

La recuperación del conocimiento mediante la utilización de motores de búsqueda es muy heterogénea. Cada motor indexa y recupera de una forma diferente. Por tanto, es importante tener unos parámetros para poder evaluar qué motores de búsqueda son los más adecuados ante una necesidad de información.

OBJETIVOS:

- Conocer el funcionamiento de los motores de búsqueda
- Utilizar una metodología para la evaluación de motores de búsqueda

ENUNCIADO:

El estudiante deberá evaluar 3 motores de búsqueda de ámbito internacional, para ello usará una estrategia de búsqueda que aplicará en los 3 motores preseleccionados, de forma que puedan apreciarse claramente las diferencias entre éstos.

- Motores de búsqueda: Google.com, Yahoo.com, Altavista.com
- Estrategia de búsqueda: digital libraries

Las características que pueden presentar los diferentes motores de búsqueda se van a agrupar en tres apartados: recogida de la información, búsqueda y recuperación de la información y presentación de los resultados.

- **Recogida de información.** En este apartado hay que determinar si el robot es capaz de identificar las etiquetas "META" de los documentos HTML y extraer información de las mismas para ser usada en la búsqueda o presentación de resultados.
- **Búsqueda.** Las características básicas que un motor de búsqueda debe cumplir desde el punto de vista de la recuperación de la información son las siguientes:
    - Formularios de búsqueda: posibilidad de elegir entre simple o avanzado.

ISBN 84-9707-571-4

- Herramientas de búsqueda: posibilidad de utilizar operadores booleanos (AND, OR, NOT), paréntesis, comillas para los términos compuestos o frases y, finalmente, posibilidad de truncado en palabras derivadas.
- Clasificación temática: existencia de un índice general para aquellos que no saben concretar su tema de búsqueda.
- Campos de búsqueda: posibilidad de limitar la búsqueda a campos determinados, tales como: Título, URL, Descripción, Palabras Clave, Localización e Idioma.
- Control del vocabulario: descubrir si el motor dispone de alguna herramienta para eliminar sinonimias y polisemias, etc.
- Detección de novedades: posibilidad de diferenciar los nuevos recursos incorporados.

- **Resultados.** Hay que tener en cuenta si el motor de búsqueda permite elegir entre diferentes formatos de presentación de los resultados así como diversos criterios de ordenación.

FORMATO:

El establecido en la tabla anexa.

ISBN 84-9707-571-4

CUADRO PARA LA EVALUACIÓN DE MOTORES DE BÚSQUEDA

| **MOTORES** | **Recogida de información** | | | **Búsqueda y Recuperación de Información** | | | | | | | | | | | | | | | | **Resultados** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Metadata | No Metadata | ? | Formularios | | Herramientas de búsqueda | | | | | | Clasif. | Campos de búsqueda | | | | | | Control Vocab. | Nov. | Format. | Orden |
| | | | | Simple | Avanz. | OR | AND | NOT | ( ) | " " | * | | Tít. | URL | Desc. | Keyw. | Loc. | Leng. | | | | |
| | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | |



ISBN 84-9707-571-4

### 1.5.2. Caso práctico 2. Selección de un motor de búsqueda

PRESENTACIÓN:

Una organización ha destinado una partida presupuestaria para aumentar la presencia web del ayuntamiento e implementar servicios y productos de información interactivos para los ciudadanos.

Los departamentos de Informática, Documentación y Comunicación están colaborando en el proceso de compra de nuevo software que permita implementar estas prestaciones y sea compatible con el back-office de la organización.

En la próxima reunión se va a decidir qué software para la implementación de un motor de búsqueda en la Intranet y sitio web del ayuntamiento se va a adquirir.

OBJETIVOS:

- Conocer las características de los principales software del mercado para la implentación de tecnología pull para la recuperación de la información.
- Presentar una aplicación práctica de la gestión del conocimiento (recuperación) en un entorno web.
- Evaluar un paquete de software con relación a unos criterios técnicos y metodológicos preestablecidos.

ENUNCIADO:

El estudiante es la persona que representa al Departamento de Documentación en esta reunión y debes presentar tu propuesta del paquete/s de software que más se adapta a los requerimientos de la organización

Los requerimientos establecidos en la reunión anterior son:

- Software compatible con el Sistema de Información del ayuntamiento: Oracle portal sobre UNIX, Bases de datos Access y SQL Server, JSP y Dreamweaver Ultradev.
- Software compatible con cualquier plataforma web: Intranet, sitio web, CD-ROM, DVD.
- Indexación de materiales diversos: HTML, XML, asp, php, pdf, doc, etc.
- Entorno usable y familiar para el usuario interno y externo.
- Opciones de búsqueda avanzada: operadores booleanos, truncamiento, limitaciones de campo, y otros.
- Indexación de páginas HTML y de texto en varios idiomas.

ISBN 84-9707-571-4

FORMATO:

El formato debe ceñirse a los siguientes apartados:

1. Metodología de la evaluación: [el estudiante debe enunciar las fuentes consultadas y el conjunto de los paquetes de software analizados]

2. Evaluación del software: [el estudiante debe recopilar la siguiente información del paquete/s de software seleccionado/s]

| Nombre del SW:<br>Precio:<br>Requerimientos que cumple: | |
|---|---|
| Compatibilidad con el sistema de información | |
| Compatibilidad con diferentes plataformas web | |
| Indexación de materiales diversos | |
| Usabilidad del entorno | |
| Opciones de búsqueda | |
| Idiomas | |

3. Propuesta del Departamento de Documentación: [el estudiante debe comentar algunos puntos a favor y/o en contra del software propuesto como opción 1]

RECURSOS:

Sullivan, Danny. Search Engine Software for your web site. *SearchEngineWatch.com*, Sept. 16, 2002. (Consultado el 23-05-2006). http://searchenginewatch.com/resources/software.html

ISBN 84-9707-571-4

## 1.6. Anexo. Introducción a Google

**Presentación**

Internet es una extensa red de documentos de múltiples formatos, desde páginas web en html a documentos de distintos tipos como puedan ser textos, imágenes, archivos de sonido o de vídeo, etc. Cuando necesitas buscar cierta información en Internet lo más eficaz es utilizar un buscador, ya que estas herramientas disponen de robots que rastrean la web en busca de información, e indexan los documentos según sus formatos y contenidos, de tal forma que muestran a sus usuarios los documentos relevantes con los criterios de búsqueda elegidos.

Conocer adecuadamente las propiedades y herramientas de cada buscador nos puede ayudar a restringir las búsquedas a lo que realmente es relevante para nosotros, sin embargo, generalmente, estos criterios son desconocidos por el internauta medio, ya que se basan en criterios documentales.

Google es el buscador más famoso y utilizado, realizar una búsqueda bien acotada es necesario incluir el máximo de palabras relevantes para el usuario, prestando especial atención a los diferentes significados de una palabra. Google determina los resultados por las coincidencias y cercanía de las palabras entre sí. Google tampoco distingue entre mayúsculas y minúsculas, ni acentos. Entrecomillar frases obliga al buscador a encontrar páginas que tengan esa frase completa.

**Búsqueda sencilla**

http://www.google.es/intl/es/help/basics.html

¿Qué operador booleano utilizan por defecto las búsquedas sencillas? Pon un ejemplo.

¿Google permite la utilización de comodines (*) para limitar la búsqueda? Pon un ejemplo.

¿Google diferencia los acentos y las mayúsculas y minúsculas? Pon un ejemplo.

**Restricciones en la búsqueda**

http://www.google.es/intl/es/help/refinesearch.html

¿Cómo se excluye una palabra de una estrategia de búsqueda? Por ejemplo, de la búsqueda [biblioteca arte moderno], qué debo hacer para excluir la palabra moderno

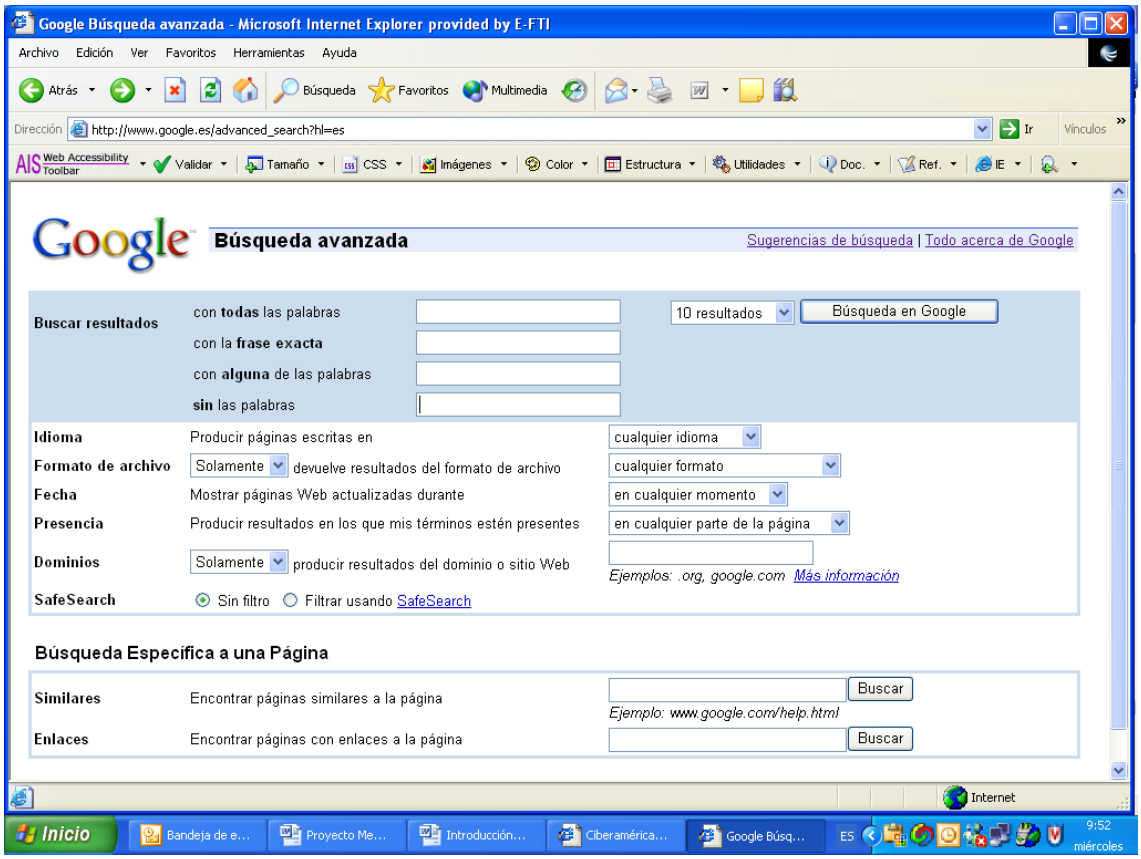
ISBN 84-9707-571-4

¿Qué debo hacer para buscar una frase, por ejemplo [gran hermano]?

¿Cómo puedo buscar noticias sobre la guerra de Irak sólo en el sitio web del Elmundo.es?. Especifica la estrategia de búsqueda.

**Búsqueda avanzada**

http://www.google.es/advanced_search?hl=es




Los buscadores de internet han superado ampliamente las clásicas posibilidades de filtrado de preguntas basadas en el álgebra booleana (operadores Y, NO, O, en inglés AND, OR y NOT) Por ejemplo, google descarta por defecto el operador booleano OR (historia O roma), asumiendo que su base de datos es tan grande que, o intenta ser muy específico, o el resultado obtenido en una consulta de este tipo tendrá demasiado ruido, esto, saldrán demasiadas respuestas. Aun así, nos permite utilizarlo a voluntad en su búsqueda avanzada.

Así, google busca por defecto con el operador AND (historia Y roma), y es más, aplica por defecto un operador NEAR decreciente. NEAR busca páginas en las que aparezca "historia" y también "roma", con una o más palabras de separación entre ambos conceptos. El número de palabras se regula por 1,2,3, etc.

ISBN 84-9707-571-4

De este modo, si tecleamos en la cajetilla de búsqueda nuestras dos palabras ( historia - roma), intenta encontrar primero las palabras adyacentes y en el orden en que se han escrito. Después aumenta NEAR de NEAR 1 a NEAR 2,3, etc, independientemente del orden en que fueron escritas en la búsqueda ("query") aunque esta característica se combina con los otros parámetros de asignación del ránking de google.

¿Cuál sería la estrategia de búsqueda para localizar información en formato pdf sobre accesibilidad de sitios web y que los términos se encuentre en el título de la página?

¿Cuál sería la estrategia de búsqueda para localiza información sobre opacs en catalán y publicada en los últimos 3 meses?

¿Cómo buscarías páginas similares a www.boe.es?

¿Cómo buscarías las páginas que tienen un enlace a http://www.eubd.ucm.es/?

**Aplicaciones tecnológicas de google**

http://labs.google.com/

¿Para buscar bibliotecas en Orlando, que aplicación se debe utilizar?

¿Para recibir noticias sobre motores de búsqueda una vez a la semana, que aplicación se debe usar?

**Google Page Rank**

http://trucosdegoogle.blogspot.com/2002_12_01_trucosdegoogle_archive.html#85791235

http://www.webtaller.com/google/pagerank.php

http://www.hipertext.net/web/pag216.htm

¿Qué es Google Page Rank y cómo funciona?

**Prestaciones especiales de Google**

http://www.google.es/intl/es/features.html

ISBN 84-9707-571-4

¿Google permite prestaciones de calculadora y realizar operaciones básicas? Pon algunos ejemplos de éxito y error.

¿Cómo se pueden conocer las páginas que apuntan o tienen un enlace a una determinada url?. Por ejemplo las páginas que apuntan a http://www.eubd.ucm.es/

¿Qué es y cómo funciona el botón "Voy a tener suerte"?

**Google para empresas**

http://www.google.es/enterprise/index.html

Google Mini permite realizar búsquedas rentables y de alta calidad en el sitio web público o la intranet de su empresa, y se instala y se pone en marcha en menos de una hora

Google Search Appliance indexa todo tipo de contenido de su intranet y sitios web, por lo que constituye una solución sólida, escalable y rentable para las necesidades de búsqueda de su empresa

**Servicios específicos de Google para documentalistas**

La estrategia de motores de búsqueda como Google que comienza a realizar tareas que tradicionalmente han realizado organizaciones intermediarias de información, como las bibliotecas o los servicios de información científica (ISI).

Algunos ejemplos, recogidos en http://www.google.com/services/librarian_center.html son:

**1. Google Scholar**: Un motor de búsqueda dirigido a docentes y científicos que se constituye como un verdadero servicio de información y vigilancia científica con información a texto completo.

Más información:

El mundo.es. 'Google Scholar', la versión beta de un buscador para docentes y científicos. http://www.el-mundo.es/navegante/2004/11/19/empresas/1100856475.html

**2. Google Print**: digitalización e indexación de millones de libros, con acceso libre y gratuito.

Más información:

ISBN 84-9707-571-4

20Minutos. Google Print llega a España y a otros siete países europeos http://www.20minutos.es/noticia/57685/0/Google/Print/libros/

Noticiasdot.com. La "Biblioteca Google" despierta entusiasmo y temores entre profesionales.http://www.noticiasdot.com/publicaciones/2004/1204/1612/noticias161204/noticias161204-15.htm

El mundo.es. Google digitalizará los libros de cinco de las mejores universidades del mundo. http://www.el-mundo.es/navegante/2004/12/14/cultura/1103031183.html

ISBN 84-9707-571-4

# 2. El posicionamiento en los motores de búsqueda

## 2.1. Concepto de posicionamiento web

Posicionamiento se puede definir como el conjunto de procedimientos que permiten colocar un sitio o una página web en un lugar óptimo entre los resultados proporcionados por un motor de búsqueda. Por extensión: Optimizar una página web de cara a los resultados proporcionados por los motores de búsqueda. En este sentido, esta disciplina a veces se denomina también "optimización en motores de búsqueda". Esto es, el conjunto de procedimientos para mejorar la posición de un recurso electrónico en los resultados de un motor de búsqueda se denomina posicionamiento web (web positioning) o bien optimización en motores de búsqueda (search engine optimization, SEO). La posición relativa de un recurso electrónico depende de más de 100 parámetros que recogen los algoritmos que utilizan los robots de los buscadores para encontrar este recurso entre los millones que tienen indexados. Además, los parámetros que utilizan los diferentes motores de búsqueda no son conocidos, ya que forman parte de su ventaja competitiva y son objeto de secreto industrial.

El posicionamiento se puede alcanzar mediante una planificación o bien de forma natural:

- **Posicionamiento planificado:** el posicionamiento que consigue una página o un sitio web debido a una campaña consciente y planificada. El posicionamiento planificado puede ser ético o fraudulento.
- **Posicionamiento natural:** el posicionamiento que consigue una página o un sitio de modo espontáneo, es decir, sin que sea consecuencia de una campaña consciente o planificada.

## 2.2. Criterios básicos para el posicionamiento

Los motores de busca mantienen en secreto el detalle último de sus procedimientos, ya que se trata de una información susceptible de conferirles ventaja competitiva y, por tanto, la consideran un secreto industrial. Por este motivo, todo aquello que los estudiosos y profesionales afirman sobre el tema, en realidad es, o bien simple especulación, o bien resultado de inferencias indirectas. Es decir, a partir de la observación y del análisis de los resultados, existe un cierto consenso entre los analistas sobre qué clase de criterios usan los motores de búsqueda para ordenar los resultados.

ISBN 84-9707-571-4

A continuación se especifican algunos criterios que, a juicio de la mayor parte de los analistas siguen los tres o cuatro mayores motores de búsqueda generalistas (Google, Yahoo, HotBot y MSN, entre otros). Ahora bien, aunque toda la evidencia apunta hacia el hecho de que los criterios señalados a continuación son los más importantes, se ignora como combinan, en cada momento, la importancia de cada uno de ellos. Además, tales criterios pueden variar a lo largo del tiempo.

A modo de síntesis, los motores de búsqueda combinan dos grupos de criterios: internos a la página web y externos a la página web.

### 2.2.1. Criterios de optimización internos a la página web

Se trata de criterios intrínsecos a la página web, que forman parte de su contenido, tanto mediante la codificación realizada en el lenguaje de marcado correspondiente, como en los metadatos utilizados para la descripción del recurso electrónico o página web.

**Optimización de palabras clave**. La selección óptima de las palabras clave es la base de toda estrategia de posicionamiento web, por tanto, se profundizará más adelante sobre este criterio.

**Optimización de los títulos.** Dado que la etiqueta <title>título</title> situada en el <head> de una página web es lo que los buscadores muestran en la lista de resultados, el objetivo de la optimización es doble. Por un lado debe ser un reclamo para que los usuarios entren en la web, y por otro debe estar configurado de tal forma que los buscadores otorguen una buena posición a la web. Para alcanzar buenos rankings de relevancia se aconseja redactar títulos de entre 5 y 10 palabras en los que se mencione por lo menos una vez las keywords que optimizan la web. Además, se debe especificar la estructura fundamental en la que la página se encuentra enmarcada, por ejemplo “Ayuda para la consulta – Catálogo general – Biblioteca Nacional”.

**Las metaetiquetas o metadatos**. Los lenguajes de marcado (html, xhtml, dhtml, etc.) permiten utilizar una serie de etiquetas, denominadas Meta, a través de las cuales se puede añadir una serie de informaciones sobre una página. Principalmente se suelen utilizar para describir el contenido a través de pequeños resúmenes y palabras-clave. Hay robots que son capaces de identificar las etiquetas META y extraer la información de las mismas para ser usada en la búsqueda o en la presentación de resultados.

Los metadatos están incluidos dentro de la etiqueta <head>, tienen como objetivo ofrecer a los buscadores información acerca del recurso electrónico. Las más

ISBN 84-9707-571-4

importantes son: (además del título comentado más arriba), la etiqueta META DESCRIPTION (describe el contenido de la página y sirve de descripción en los resultados de algunos buscadores), la etiqueta META KEYWORDS (actualmente casi todos los buscadores la ignoran debido a su manipulación para conseguir mayor relevancia). También tienen especial relevancia la etiqueta META LANGUAGE (indica el idioma de la página) y la etiqueta META ROBOTS (indica al buscador si se desea indexar la página y/o se desean seguir los links). A pesar de que algunos motores de búsqueda no los tienen en cuenta, se recomienda continuar creándolas para cada una de las páginas de la web, puesto que suelen ser un factor relacionado con la calidad del recurso y se pueden utilizar para otros propósitos relacionados con la gestión de recursos electrónicos.

**Elementos de descripción contextual**: además de los metadatos de la sección *head* (principalmente *title* , *description* y *keywords* ), otras etiquetas también proporcionan información descriptiva de utilidad para los buscadores y donde se deben colocar las palabras clave secundarias:

- Los encabezados, que se codifican mediante <Hn> (donde n es un número del 1 al 6) se utilizan para estructurar jerárquicamente los contenidos principales de una página web. Por tanto, aquellas palabras clave destacadas deberían situarse en los niveles de encabezado más altos, esto es, H1 y H2.
- El texto de los enlaces, que es la parte activa codificada mediante <a href="url">texto</a> y donde se establecen enlaces internos o externos a los contenidos del sitio web, contenidos cuyos textos se consideran importantes como palabra clave para la indexación por parte de los motores de búsqueda.
- Los "alt" de las imágenes, según las Pautas de accesibilidad a los contenidos web, se prevé que todas las imágenes deben contener un alt (alternative text o texto alternativo), que debe describir el contenido fundamental de la imagen. Si la imagen no transmite información el atributo alt debe ir vacío.
- Los "title" de los enlaces y las imágenes, en principio la utilización del atributo title para enlaces e imágenes, aunque es valorado por algunos motores de búsqueda puede entrar en contradicción con los criterios fundamentales de la accesibilidad web. En accesibilidad web sólo se debe incluir el atributo "title" en un enlace, cuando no es posible activar alguna palabra o palabras significativas. Además, aunque el atributo title se acepta para las imágenes, las Pautas de accesibilidad a los contenidos web

ISBN 84-9707-571-4

recomiendan la utilización de “alt” y no mencionan, de forma positiva o negativa la utilización del atributo “title”.

- La etiqueta “strong”, que se codifica mediante texto y denota que el texto destacado como “strong” tiene cierta importancia. La etiqueta “strong” se utiliza como equivalente a <b> (bold), que es una etiqueta desaconsejada en HTML. La etiqueta “strong” tiene mucho peso semántico y se muestra en los navegadores como negrita.
- La etiqueta “em”, que se codifica mediante texto, se utiliza para dar énfasis a una palabra o frase, marcando de forma distintiva los puntos más importantes de un texto. La etiqueta “em” tiene menos peso semántico que “strong” y se muestra en los navegadores como cursiva.

**Palabras clave secundarias**. Las keywords secundarias se deben distribuir correctamente en el texto de una página. Se recomienda una densidad (porcentaje de palabras clave sobre el total de palabras) de entre el 5% y el 8%. Según el principio del *keyword proximity* se recomienda situarlas lo más cerca posible del principio de la página. Para darles más importancia existen varias etiquetas: <Hn>, <i>, <b>,  y <li>. La keyword principal se resalta con un <H1> y debe colocarse cerca del principio de la página.

### 2.2.2. Criterios de optimización externos a la página web

**El PageRank.** Es un valor entre 1 y 10 que depende de la cantidad y calidad de las webs que tengan links hacia la web de referencia, así como de sus links internos. El PR transmitido por las webs depende a su vez del PR propio y del número de links salientes que tenga esa página [2]. La fórmula del PR es la siguiente: **PR(A) = (1-d) + d (PR(T1)/C(T1) +...+ PR(Tn)/C(Tn)) PR(A)** es el PageRank de la página de referencia. **d** es un factor de debilitación. **(1-d)** asegura que cualquier página, aunque no reciba ningún enlace, tendrá un PR mínimo de 0'15. **PR(Ti)/C(Ti)** es el PageRank (PR) de la página i-ésima que enlaza a la web de referencia, (Ti), dividido por todos los enlaces (C) que también salen de esa página Ti, es decir, el PR que transmite. **i = 1,...,n** ya que se suponen n páginas que enlacen a la de referencia.

**El texto de los links.** El texto de los enlaces que apuntan a una página es considerado por algunos como el recurso número uno de la optimización. Las palabras clave se deben colocar en el texto que actúa como anchor de la etiqueta de un link.

ISBN 84-9707-571-4

**Los links externos.** Dado que los motores de búsqueda usan medidas de popularidad de enlaces (cantidad de sitios "relacionados" o "autoritativos" que enlazan a una determinada web) para determinar los rangos de los resultados de las búsquedas, se recomienda la implementación de estrategias de "Popularidad de Enlaces", como pueden ser el alta en Directorios o el intercambio de links.

## 2.3. Los metadatos y el posicionamiento web

### 2.3.1. Concepto de metadatos

Los metadatos o datos representacionales son definidos como el dato sobre los datos, es un conjunto de elementos que poseen una semántica comúnmente aceptada, o sea tratan de representar la información electrónica tan dispersa y representan a la descripción bibliográfica de recursos electrónicos. Estos datos abarcan ámbitos tanto individuales como colectivos, también documentos, recursos de Internet e incluso objetos reales. Nace de la necesidad de recuperar la información electrónica tan dispersa. Los metadatos tratan, principalmente, de describir el contenido y la localización del objeto de la información en Internet.

Una de las características mas importantes de los metadatos va a ser su capacidad de relación o de establecer enlaces. De esta forma se han hecho imprescindibles en la recuperación global de la información en Internet, puesto que se trata de indizar y clasificar inconmensurables cantidades de información de diversos tipos. Se tratará de integrar de forma heterogénea fuentes de información muy diversas, así como integrar diferentes formatos de bases de datos. O sea, se emplean metadatos para organizar el contenido de la información en Internet. De esta forma en la definición de metadatos podemos incluir y de forma mas específica, se refiere a la información accesible por Internet. Por lo que los metadatos tienen el objetivo primordial de que los documentos introducidos en la Red incluyan todos los datos necesarios para su posterior búsqueda, localización y recuperación. Ya que introducir o publicar dentro de Internet es una tarea sencilla, sin embargo la localización, control y uso de la información es una tarea mas compleja. Por tanto, será una tarea primordial establecer las normas y elementos que ha de contener cualquier descripción y catalogación de recursos en Internet.

Un catálogo de biblioteca o un repertorio bibliográfico son tipos de metadatos. Estos tipos de metadatos emplean, fundamentalmente, reglas de catalogación y formatos para transmitir la información como los formatos MARC. Por lo que nuestra primera idea de metadatos van a ser los catálogos bibliotecarios y bibliográficos. O sea, cada ficha es un metadato de un libro o bien de un autor y los metadatos

ISBN 84-9707-571-4

proporcionan una información básica sobre las obras de un autor y lo relaciona con otras obras del mismo autor u otras obras de similar contenido. Lo que hasta ahora venía denominándose descripciones bibliográficas o registros bibliográficos hoy día van a ser denominados metadatos o sea que tienen como objetivo la descripción de los recursos de Internet.

Existen varios gestores de metadatos que tratan de unificar el mapa representado para cada documento, los elementos de los datos y la conversión de varias sintaxis en una sola. El movimiento de metadatos en Internet trata de integrar distintos formatos de metadatos de las bases de datos para ser integrados conjuntamente aunandose el legado de los catálogos automatizados de las bibliotecas y una estructura de catálogos electrónicos o también denominados catálogos hipertextuales, donde su idiosincrasia radica, no solo en las formas tradicionales de acceso, sino en la propia estructura del hipertexto con enlaces tanto para la clasificación sistemática como para la alfabética e incluso para toda la descripción bibliográfica. Ya que se trata de una estructura articulada con distintos tipos de enlaces. De esta forma los servicios y fuentes de los catálogos electrónicos van a estar accesibles también a través de los denominados buscadores y de las propias fuentes de las páginas Web. Puesto que la nueva estructura de las bases de datos es accesible a través de las páginas Web, supone que las usuales bases de datos transformen su propia estructura.

Así la arquitectura navegable y jerárquica reporta a los diferentes metadatos de distintos formatos para que converjan en uno, y además posibilita establecer una estructura de enlaces que los haga accesibles. Existen varios modelos de metadatos, pero en la aplicación bibliográfica y bibliotecaria se ha extendido e implantado, de forma mas mayoritaria, el formato denominado "Dublin Core" o Círculo de Dublin, creado por las iniciativas de las asociaciones de bibliotecarios americanos, y en concreto por OCLC (On Line Computer Library Catalog ). Se trata de un formato bastante standard para las fuentes de Internet, originariasmente bibligráficas y bibliotecarias. Es un formato de metadatos, basado en la asociación de superenlaces, y estableciendo mapas semánticos similares a los elementos y estructuras con metadatos standares. O sea, se trata de un sistema de conversión de metadatos que abarca y contiene metainformación, esta conversión necesita todavía de la intervención humana e identifica y enlaza las páginas Web. En definitiva, es un formato muy simple que incluso puede ser aplicado por catalogadores no muy expertos.

Los metadatos Dublin Core tratan de ubicar, en el entramado de Internet, los datos necesarios para describir, identificar, procesar, encontrar y recuperar un documento introducido en la Red. Si este conjunto de elementos Dublin Core se lograra aceptar internacionalmente supondría que todos los robots que indizan documentos en

ISBN 84-9707-571-4

Internet encontrarían, en la cabecera de los mismos, todos los datos necesarios para su indización y además estos datos serían uniformes. La eficacia de estos robots cono Google, Altavista, Yahoo y otros mejoraría notablemente.

### 2.3.2. La función de los metadatos en la recuperación de información

La aplicación de metadatos supone una mejora en la organización y recuperación de la información, tanto de forma humana como automatizada. La gran incógnita en este sentido consiste en determinar los beneficios específicos que aportan los metadatos en la búsqueda y recuperación de la información web, sobre todo cuando muchos motores de búsqueda no utilizan los metadatos como un criterio en la indexación de los recursos electrónicos y, por tanto no se utilizan para la búsqueda. Si embargo, existe una amplia gama de software de motor de búsqueda para la indexación de los recursos del sitio web, la Intranet y los productos electrónicos (CD-ROMs, DVDs y otros productos que utilicen tecnología web) de las organizaciones que indexan y gestionan metadatos.

Algunos motores de búsqueda (como Convera, Harvest, Blue Angel, Microsoft Site Index, etc.) son capaces de utilizar los metadatos y otras herramientas de representación del conocimiento (como ontologías y topic maps) para obtener mejores resultados en la recuperación. Por tanto, aunque las organizaciones públicas pueden sentirse reticentes ante la incorporación de metadatos en sus recursos de información debido al esfuerzo económico que ello supone (coste de personal y tecnología), es importante que la Administración Pública en su conjunto tome conciencia de la importancia de los metadatos para mejorar la relevancia de los sistemas de recuperación de información, así como para facilitar la integración y combinación de recursos heterogéneos en el desarrollo de servicios electrónicos y mejorar el acceso de los usuarios a los recursos.

Los sistemas de recuperación de la información en Internet de propósito general (motores de búsqueda) se basan en la extracción automática de la información y utilizan sencillas técnicas para representar el conocimiento contenido en los recursos electrónicos. Por tanto, no pueden dar una respuesta precisa a una pregunta concreta sobre el contenido semántico de los documentos y por ello, recuperan mucho ruido. Sin embargo, los sistemas de recuperación en sectores específicos, como la información pública, dado que todos los recursos de información son objeto de descripción, organización y control del vocabulario, ofrecen mayor relevancia en la recuperación.

ISBN 84-9707-571-4

La existencia de un compromiso para que la información de carácter público se adapte a unos estándares, y contemple el uso de metadatos en todos los recursos electrónicos y digitales, favorece la recuperación de la información en este ámbito de conocimiento. La clave esencial reside en la aplicación de metadatos de forma sistemática, normalizada y coherente. Con este proceso se facilita la descripción de todos los recursos de la organización (aplicación sistemática), el intercambio de información (mediante la normalización), y su adaptación a nuevas formas tecnológicas (aplicación coherente).

#### 2.3.2.1. La iniciativa Dublin Core

La Iniciativa de Metadatos Dublin Core DCMI es una organización dedicada a la promoción y difusión de normas interoperables sobre metadatos y el desarrollo de vocabularios especializados en metadatos para la descripción de recursos que permitan sistemas de recuperación mas inteligentes.

Uno de los esfuerzos de los participantes de DCMI es el desarrollo en colaboración y el continuo perfeccionamiento de convenciones sobre metadatos basados en la investigación y la opinión entre los Grupos de Trabajo DCMI.

#### 2.3.2.2. Los elementos Dublin Core

Los elementos poseen nombres descriptivos que pretenden transmitir un significado semántico a los mismos. Para promover una interoperabilidad global, una descripción del valor de algunos elementos podrá ser asociada a vocabularios controlados. Se asume que otros vocabularios controlados serán desarrollados para asegurar esta interoperabilidad en dominios específicos.

Cada elemento es opcional y puede repetirse. Además, los elementos pueden aparecer en cualquier orden.

Aunque algunos entornos, como HTML, no diferencian entre mayúsculas y minúsculas, es recomendable escribir correctamente cada metadata según su definición para evitar conflictos con otros entornos, como XML (Extensible Markup Language) http://www.w3.org/TR/PR-xml

Podemos clasificar estos elementos en tres grupos que indican la clase o el ámbito de la información que se guarda en ellos:

- Elementos relacionados principalmente con el contenido del recurso
- Elementos relacionados principalmente con el recurso cuando es visto como una propiedad intelectual

ISBN 84-9707-571-4

- Elementos relacionados principalmente con la instanciación del recurso

Tabla. Clasificación de los elementos DC

| Contenido | Propiedad Intelectual | Instanciación |
|---|---|---|
| Title | Creator | Date |
| Subject | Publisher | Type |
| Description | Contributor | Format |
| Source | Rights | Identifier |
| Language | | |
| Relation | | |
| Coverage | | |

Más información: http://es.dublincore.org/documents/dces/

## 2.4. La optimización de las palabras clave

Una palabra clave en el ámbito del posicionamiento en Internet ser refiere al término respecto al cual se persigue la optimización de una página web. Puede ser una palabra única, como "Arte" o una frase, como "Subastas de Arte". Optimizar para una frase siempre será más fácil que para una palabra clave; a su vez, posicionar para una combinación de dos palabras clave siempre será más fácil que posicionar para cada palabra clave aislada. Por ejemplo, siempre será más fácil posicionar para una pregunta del tipo <"Arte"AND "Barcelona">, que para cada una por separado. Al mismo tiempo, se ha comprobado empíricamente que es muy difícil posicionar a la vez un mismo sitio para más de tres o cuatro palabras clave (cada una de ellas por separado). También es mucho más difícil optimizar un sitio por una palabra clave cuanto más competitiva sea la palabra clave, es decir, cuando muchas páginas y muchos sitios web contienen esa palabra. Por ejemplo, probablemente es mucho más difícil posicionar un sitio para la palabra clave "hardware"que para la palabra clave "arañas". La primera tiene mucha presión

ISBN 84-9707-571-4

comercial, con muchos sitios queriendo posicionarse, a diferencia de la segunda. En general, la palabra clave ideal sería alguna que fuera muy buscada, pero con pocos sitios web que la contengan. La peor palabra clave sería una muy poco buscada y con muchos sitios que la contengan. La mayor parte de las veces tendremos que conformarnos con una palabra clave o grupo de palabras clave que mantengan un cierto equilibrio entre ambos elementos.

La selección de palabras clave es considerada como el factor más importante en el posicionamiento en buscadores y es la base de toda estrategia SEO. Para una selección óptima de keywords se deben tener en cuenta los siguientes puntos:

Popularidad de las Keywords.

Competencia por las keywords.

Relevancia desde el punto de vista del Marketing.

Relación de las keywords con el contenido de la web.

**Popularidad de las Keyword** . Dado que las palabras individuales son en general muy buscadas, suelen atraer mucha competencia. La solución a este problema es elegir "frases clave", compuestas de entre dos a cinco palabras, para la optimización de cada página. También se recomienda optimizar las páginas para palabras clave de gran popularidad mal escritas (p.ej. palavras por palabras). Para estas variantes la competencia será mucho menos intensa.

**Competencia por las keyword** . Al optar por palabras clave de gran competencia existe el riesgo de que la web se pierda entre la multitud de resultados del buscador. Debido a esta alta demanda por algunas palabras, se aconseja identificar un nicho o conjunto de palabras clave que describan claramente la web y que sólo unos pocos hayan elegido antes.

**Relevancia desde el punto de vista del Marketing.** En el diseño de una buena estrategia de marketing para buscadores se consideran multitud de aspectos. Por un lado, es necesario tener un excelente conocimiento de nuestro público objetivo, es decir, qué buscadores suelen utilizar, y qué palabras clave usan para encontrar un producto o servicio. En este contexto, es importante dotar a los productos de la web de palabras clave específicas y precisas, ya que éstas tienen la ventaja de tener menos competencia y de asegurar visitas más cualificadas. Por otro lado, se debe tener en cuenta que los usuarios de un buscador no piensan como un director de marketing, y que lo que suena bien para un eslogan promocional no corresponde a las frases que probablemente se utilizarán en un buscador.

**Relación de las keyword con el contenido del site** . Se deben elegir no sólo las palabras clave que encajan con los contenidos de la web, sino además las palabras clave que los usuarios usan para encontrar sitios como el de la web a optimizar. Por regla general, para cualquier página con contenidos escritos superiores a las 250 palabras se recomienda utilizar de una a tres palabras clave.

ISBN 84-9707-571-4

Teniendo en cuenta los puntos anteriores, y tras elegir las keywords que identifiquen la web, se debe optimizar la página principal para las keywords que mejor se ajusten al site. El siguiente paso es optimizar las páginas secundarias para las keywords que identifiquen esas páginas. No obstante, no hay que olvidar agregar siempre en todas ellas la keyword de la página principal. Así, se destaca esa palabra clave representativa de la web y se consiguen también buenas posiciones para las keywords secundarias.

## 2.5. La planificación de un proyecto de posicionamiento

### 2.5.1. Plan de posicionamiento

El posicionamiento web se refiere a la posición relativa de su sitio web frente a los demás sitios como resultado de una búsqueda de términos relacionados con su negocio en los buscadores. La posición relativa de su web depende de más de 100 parámetros que recogen los algoritmos que utilizan los robots de los buscadores para encontrar su página entre las millones que tienen indexadas.

Nuestros servicios de posicionamiento en buscadores o SEO (Search Engine Optimization) están diseñados para optimizar, sobre tecnología punta, los parámetros más significativos para que su sitio web logre posicionarse entre las primeras posiciones en los principales buscadores (Google, Yahoo, MSN, Hispanista y otros) cuando un usuario busque los términos más relacionados con su negocio.

Tareas:

- Auditoria de estructura de navegación, diseño, tecnologías empleadas y contenidos de su web. (Validación de lenguajes de marcado, encabezados de página, alt de las imágenes, textos de los links, metadatos normalizados, etc.)
- Informe de recomendaciones de cambios y ajustes en base a parámetros críticos.
- Ejecución de cambios y ajustes.
- Evaluación final de visibilidad por términos relacionados y por marca.

### 2.5.2. Alta en los principales buscadores

El alta en buscadores se refiere al proceso de registrar (indexar) un sitio web en los motores de búsqueda. La finalidad de este proceso es que los robots o spiders de los motores de búsqueda consideren a su sitio web como resultado potencial de las

ISBN 84-9707-571-4

búsquedas que realizan. Si su sitio web no está indexado será invisible para los mismos por lo que representa un paso obligatorio y fundamental en el proceso de posicionamiento.

### 2.5.3. Enlaces patrocinados

Las **campañas de enlaces patrocinados** permiten una flexibilidad absoluta y casi instantánea en su gestión, pudiendo manejar rápida y eficazmente y hacer cambios en sus campañas en tiempo real.

Los e**nlaces patrocinados** son altamente recomendados para la promoción de nuevos productos o para dar a conocer su página web de forma rápida y efectiva. Además, son el complemento perfecto a las acciones publicitarias convencionales ya que los usuarios tienden a buscar en Internet información adicional sobre los productos que desean adquirir.

La inversión necesaria para optimizar una campaña de e**nlaces patrocinados** depende de factores tanto cuantitativos como cualitativos.

- Factores cuantitativos, como la competencia. En sectores de gran presencia en la red, las palabras clave pueden superar los 3 euros el clic. Sin embargo, en mercados poco maduros, se puede comenzar a pujar desde 0,05€ el clic.
- Factores cualitativos, como la calidad de redacción de los textos que figuran en los enlaces patrocinados. Un anuncio más atractivo, editorialmente hablando, tendrá una frecuencia de click superior, por tanto, recibirá un mayor número de clicks, por lo que el sistema de gestión de pujas tendrá este factor muy en cuenta a la hora de colocar el anuncio en una determinada posición.

Tareas:

- Análisis previo y planificación de la estrategia de e-marketing
- Elaboración de listados de palabras clave
- Creación de contenidos publicitarios (banners y enlaces patrocinados)
- Lanzamiento y supervisión de la campaña de e-marketing
- Informe de resultados

Para la realización de las tareas anterioes de una forma eficiente se recomienta utilizar algún tipo de software especializado, como: Atlas Search o Dart Search para

ISBN 84-9707-571-4

gestionar de forma integrada las campañas de enlaces patrocinados en Google adwords, Yahoo Search Marketin, MSN adcenter, MIVA y otros.

### 2.5.4. Servicios de consultoría

Algunos servicios de consultoría que complementan un plan de posicionamiento son los siguientes:

**1. Análisis de la audiencia.** Se realiza la comparativa de los las palabras clave más significativas para un sector o materia determinados, y por otro, el potencial de tráfico y clicks estimados. Se emplea para conocer las palabras clave más relevantes y más útiles para nuestros servicios

**2. Análisis de la competencia.** Se realiza una comparativa de las audiencias de las 10 palabras clave más significativas para un sector o materia determinados, comparando con un número determinado de competidores. Este estudio se emplea para conocer con todo detalle la fortaleza respecto a la competencia.

**3. Análisis de visibilidad.** Se realiza una comparativa de los posicionamientos logrados por unas palabras clave en los 8 principales buscadores (99% del mercado de búsquedas español). Se emplea para medir y conocer en que situación nos encontramos actualmente en buscadores y lo fácil o difícilmente que un site es visible para su target.

ISBN 84-9707-571-4

## 2.6. Bibliografía

Azlor, S. (2003). Posicionamiento en buscadores: guía básica. http://www.guia-buscadores.com/posicionamiento.

Codina, Lluís (2004). "Posicionamiento web: conceptos y ciclo de vida". Anuario Hipertext.net. http://www.hipertext.net

Codina, Lluís; Marcos, Mari Carmen (2005). "Posicionamiento web: conceptos y herramientas". El profesional de la información, v. 14, n. 2, pp. 84-99.

Dublin Core Metadata Initiative. http://es.dublincore.org/index.shtml

Gonzalo, Carlos (2004). "La selección de palabras clave para el posicionamiento en buscadores: conceptos y herramientas de estudio" Anuario Hipertext.net. http://www.hipertext.net

Mari-Carmen Marcos et al. (2006). *Evaluación del posicionamiento web en sistemas de información terminológicos online* [on line]. "Hipertext.net", núm. 4, 2006. http://www.hipertext.net

Martínez Usero, José Angel (2006). El uso de metadatos para mejorar la interoperabilidad del conocimiento en los servicios de administración electrónica. En: *El profesional de la información*, 2006, vol.15, n. 2, pp. 114 -126.

MOEN, Willian E (2001). The metadata approach to accessing government information. *Government Information Quaterly*, 18 (2001), p. 155-165

Proyectos Dublin Core. http://es.dublincore.org/projects/index.shtml

Red Iris. Metainformación - Dublin Core. Elementos del conjunto de metadatas de Dublin Core: Descripción de Referencia. http://www.rediris.es/metadata/

SAN SEGUNDO MANUEL, Rosa (1998). *Organización del conocimiento en Internet. Metadatos bibliotecarios Dublin Core*. En: VI JORNADAS Españolas de Documentación, Valencia 1998. --Valencia : FESABID, 1998; P.805-817.http://fesabid98.florida-uni.es/Comunicaciones/r_sansegundo.htm

ISBN 84-9707-571-4

## 2.7. Caso práctico: Plan de posicionamiento web

PRESENTACIÓN:

El departamento de informática de una Administración local debe realizar un informe sobre las posibilidades de posicionamiento web del sitio web corporativo. Tu eres el documentalista de la Administración local y se te encarga la redacción de dicho informe que se habrá de presentar al Pleno de la Administración Local como respuesta a una consulta del Pleno anterior sobre cómo realizar el proceso de identificación, publicación y posicionamiento del sitio web de la administración local XYZ.

El informe consta de tres apartados:

- Creación de los metadata-metatags de nuestra administración local según las directrices de Dublín Core – Government Application Profile y The e-Government Metadata Framework (e-GMF).
- Selección y justificación de la opción que vamos a adoptar para el alta en buscadores y el posterior posicionamiento web.
- Creación de una estrategia de posicionamiento web para mejorar las relaciones económicas entre nuestra administración local y otras administraciones locales de los países de cola de la UE (Irlanda, Grecia y Portugal).

OBJETIVOS:

- Conocer la importancia de la estructuración interna de la información en páginas HTML.
- Conocer cómo funcionan los metadata-metatags y sus principales aplicaciones.
- Conocer la importancia de la organización de los recursos de información para su correcta difusión y recuperación.
- Conocer la metodología para el alta en buscadores.
- Conocer las bases del posicionamiento web.

ISBN 84-9707-571-4

ENUNCIADO:

Uno de los aspectos más importantes a la hora de publicar en la web consiste en asegurarse de que los usuarios encontrarán los recursos de información. Para ello, vamos a seguir los siguientes pasos.

Creación de Metadata-Metatags

Alta en Buscadores

Posicionamiento web

**1. Creación de Metadatos**

Los metadatos o metatag tienen dos usos principales:

Facilitar y mejorar el alta en buscadores. Un sitio web identificado adecuadamente con metadata es más accesible a través de las diferentes opciones de búsqueda disponibles en la web**.**

Favorecer los procesos administrativos y técnicos. Por ejemplo, podemos controlar la memoria caché asegurándonos de que el usuario siempre ve la ultima versión de la pagina HTML con la siguiente etiqueta: <meta http-equiv=”pragma” content=”no cache”>

Los metadata se desarrollan en el encabezamiento de una pagina HTML, dentro de la etiqueta <HEAD> Metadata </HEAD>.

Debemos tener en cuenta que cualquier organización puede utilizar las etiquetas del Dublin Core, pero al tratarse de un organismo público está sometido a nuevos procedimientos y directrices relacionados con la implementación de la Administración electrónica o e-Government, deberíamos usar una adaptación de Dublín Core denominada “Government Application Profile”.

http://dublincore.org/documents/2001/09/17/gov-application-profile/

Ejemplo

ISBN 84-9707-571-4

```
<html>
<head>
<title>Department X Home Page</title>

<meta name="DC.Identifier" scheme="URI"
content="http://www.departmentx.gov.uk">
<meta name="DC.Creator" lang="en" content="Mr AN Other">
<meta name="DC.Publisher" lang="en" content="Department X Media">
<meta name="DC.Rights" lang="en" content="Copyright Department X">
<meta name="DC.Title" lang="en" content="Department X Home Page">
<meta name="DC.Subject" lang="en" content="UK Public Sector; UK online; Tax;
Health; Defence; Civil Service;">
<meta name="DC.Description" lang="en" content="Department X is a UK public
sector body with wide ranging powers, covering Health, Tax and Defence">
<meta name="DC.language" scheme="RFC1766" content="en">
<meta name="DC.Date.created" scheme="ISO8601" content="2000-08-15">
<meta name="DC.Type.category" lang="en" content="document">
<meta name="DC.Format" scheme="IMT" content="text/html">
<meta name="keywords" content="="UK Public Sector, UK online, Tax, Health,
Defence, Civil Service">
<meta name="description" content=" Department X is a UK public sector body with
wide ranging powers, covering Health, Tax and Defence">

</head>
```

## 2. Alta en Buscadores

El proceso de alta en buscadores consiste en remitir un conjunto de datos de nuestro sitio web a un gran número de motores de búsqueda para que sean objeto de indización y, posteriormente, los usuarios puedan recuperar nuestro sitio web en una búsqueda.

Existen tres metodologías para llevar a cabo el alta en buscadores (con sus ventajas e inconvenientes)

La organización lleva a cabo el proceso de alta en buscadores de forma interna, con los recursos y el personal de que dispone.

ISBN 84-9707-571-4

La organización contrata un servicio externo genérico (una empresa que ofrece servicios TI: un proveedor de internet, generalmente) [Ejemplo http://www.arsys.es/productos/monline/ sección alta en buscadores]

La organización contrata un servicio externo especializado en registro y alta en buscadores. [Ejemplo http://www.altaenbuscadores.com ]

## 3. Posicionamiento web

Un plan de posicionamiento permite planificar en qué situación de los resultados de un motor de búsqueda queremos estar.

Para ello hay que:

3.1. Seleccionar los motores de búsqueda en los que queremos aparecer en una situación privilegiada. Ej.: Google y Lycos.

Es conveniente utilizar estadísticas de Internet para conocer cuáles son los "buscadores" más utilizados en cada país, cada sector de actividad, etc.

Algunos recursos útiles son:

http://www.searchenginewatch.com

http://www.nua.ie/surveys/

http://www.nielsen-netratings.com/

3.2. Seleccionar las palabras clave para las que queremos obtener una posición determinada.

Para hacer un estudio detallado de cuáles son las palabras clave más adecuadas para describir un recurso web, así como estimar cuáles son las palabras clave más representativas en un ámbito de conocimiento web específico, podemos usar algunas de las siguientes herramientas:

ISBN 84-9707-571-4

Keyword Report

http://www.wordtracker.com

3.3. Determinar nuestro objetivo de posicionamiento web y enunciar una estrategia.

Objetivo: Aumentar el turismo proveniente de Inglaterra, Alemania y Francia en nuestra área (la de la administración local que estamos tratando)

Estrategia: Estar entre el puesto 1 y el 5 en los dos buscadores principales de Inglaterra, Alemania y Francia para las palabras clave “España” y “vacaciones” en inglés, alemán y francés. (Ejemplo: “spain”, “holidays”)

3.4. Contratar un servicio que nos permita cumplir nuestros objetivo de posicionamiento.

ISBN 84-9707-571-4

# 3. Los agentes inteligentes de información

## 3.1. Concepto de agente inteligente

El concepto de agente inteligente es tan amplio y posee tantas aplicaciones que no es sencillo aportar una definición exacta. Bradshaw señala que la definición de agente depende del punto de vista del investigador o de los atributos propios del agente. En la misma línea, Nwana afirma que el concepto de agente ya se puede encontrar en la investigación en Inteligencia Artificial de la década de los 70s, pero que continúa siendo un término difuso, un meta-término o paraguas que da cobertura a diferentes enfoques. Finalmente, Tramullas comenta que la complejidad que rodea al ámbito de los agentes, donde intervienen la Inteligencia Artificial, la Sociología, la Lógica, las Telecomunicaciones y otras disciplinas, hace necesario el estudio de las situaciones en las que puede encontrarse un agente para definirlo de forma adecuada.

Stenmark ofrece una definición genérica: un agente inteligente es un software que asiste al cliente y actúa en su nombre.

Hipola y Vargas-Quesada lo definen como una entidad software que, basándose en su propio conocimiento, realiza un conjunto de operaciones destinadas a satisfacer las necesidades de un usuario o de otro programa, bien por iniciativa propia o porque alguno de éstos se lo requiere.

Una definición bastante apropiada, sería la que define agente inteligente como programas de ordenador capaces de efectuar una tarea o actividad sin la manipulación directa de un usuario humano. Los agentes inteligentes han cambiado sustancialmente la forma de interacción hombre-máquina. El usuario delega diferentes tareas a los agentes que son capaces de actuar en su nombre. Además, los agentes tienen la característica esencial de aprender de diferentes formas:

- Observando e imitando el comportamiento del usuario
- Recibiendo un feedback positivo o negativo del usuario
- Recibiendo instrucciones explícitas del usuario
- Pidiendo consejo a otros agentes

ISBN 84-9707-571-4

Desde el punto de vista de la gestión y recuperación de la información, una definición de agente sería: una entidad software que recoge, filtra y procesa información contenida en la Web, realiza inferencias sobre dicha información e interactúa con el entorno sin necesidad de supervisión o control constante por parte del usuario. Estas tareas son realizadas en representación del usuario o de otro agente.

## 3.2. Características de los agentes

Las características que un programa debe poseer para ser considerado un agente inteligente, en opinión de los expertos, son:

- **Autonomía**: el agente debe tener control sobre sus propias acciones y ser capaz de lanzar acciones independientemente del usuario.
- **Capacidad de reacción**: los agentes pueden detectar cambios en su entorno y reaccionar en función de éstos.
- **Comunicatividad**: el agente es capaz de interactuar con los usuarios y otros agentes.
- **Consecución de metas**: los agentes tienen un propósito determinado y actúan en consecuencia hasta conseguirlo.

Otras características de los agentes inteligentes reseñadas en la mayoría de la literatura en este ámbito son: dinamismo (los agentes deberían ser capaces de funcionar independientemente del espacio y el tiempo), adaptabilidad (los agentes aprenden y cambian su conducta basándose en las experiencias previas), continuidad temporal (los agentes no deberían parar o reanudar su actividad para ciertas tareas, más bien su funcionamiento debería ser un proceso continuo) y movilidad (los agentes se pueden transportar de una máquina a otra e, incluso, entre diferentes arquitecturas y plataformas)

## 3.3. Aplicaciones de los agentes

Existen muchos más ejemplos donde podríamos encontrarnos sistemas o áreas de aplicación donde la orientación basada en agentes resulta especialmente prometedora ofreciendo nuevas perspectivas y posibilidades. Numerosas aplicaciones basadas en este nuevo paradigma vienen ya siendo empleadas en infinidad de áreas. Podemos destacar dos áreas como serían las aplicaciones industriales y las comerciales:

ISBN 84-9707-571-4

Dentro del marco de las **aplicaciones industriales**, la tecnología basada en agentes es considerada muy apropiada para el desarrollo de sistemas industriales distribuidos. Dentro de esta línea podríamos destacar aquellas aplicaciones

que se encargan de:

· **Control de procesos**: gestión autónoma de edificios inteligentes en cuanto a su seguridad y consumo de recursos, gestión del transporte de electricidad (ARCHON), control de un acelerador de partículas, monitorización y diagnóstico de fallos en plantas industriales, como por ejemplo nucleares o refinerías, control en el proceso de bobinado del acero y robótica. En otro tipo de área se han desarrollado aplicaciones para el control del tráfico aéreo en aeropuertos como el de Sidney en Australia.

· **Producción**: aspectos como la planificación y *scheduling* de la producción o fabricación de productos serían tratados desde la perspectiva de agencia. Se ha aplicado con éxito, por ejemplo, a sistemas encargados de las fases de ensamblaje, pintado, almacenamiento de productos, etc. Algunos ejemplos serían AARIA, ABACUS, CORTES, MASCOT, Sensible Agents, YAMS, etc.

· Por otro lado, también está siendo empleado en **aplicaciones comerciales**, sobre todo a nivel de aplicaciones de red, tanto en Internet como en redes corporativas, podemos distinguir entre:

· **Gestión de información**: como por ejemplo el filtrado inteligente de correo electrónico (Agentware e InfoMagnet), de grupos de noticias o la recopilación automática de información disponible en la red (Letizia, AT1, BullsEye, Go-Get-It, Got-It, Surfbot y WebCompass). Tareas para las cuales el agente necesita ser capaz de almacenar, aprender y manipular las preferencias y gustos de cada usuario, así como sus cambios. La imposibilidad en ocasiones de gestionar todo tipo de información suministrada por la red ha provocado que el agente se especialice en la búsqueda de determinados tipos de documentos (CiteSeer). Otra posible línea sería la planificación de la agenda personal, en otras palabras, disponer de una secretaría virtual o asistente personal.

· **Comercio electrónico**: en este caso la tecnología se emplea para proporcionar el entorno virtual donde realizar posibles operaciones comerciales (compra-venta de productos) o también para realizar tareas de búsqueda de productos (comparando precios, consultando disponibilidad) todo ello de manera automatizada (Jango, BargainFinder, Kasbah). En este caso, el agente debe poder comunicarse con las tiendas en línea utilizando protocolos que permitan trabajar con las interfaces de estas tiendas; actualmente, los usuarios pueden comprar y vender artículos

ISBN 84-9707-571-4

comunes como libros y CD's de música. El empleo de agentes aumentará el impacto del comercio electrónico en un futuro muy cercano, revelando asimismo cómo los agentes basados en la web pueden proporcionar un enorme poder añadido a los consumidores.

· **Monitorización**: proporcionan al usuario la información cuando sucede un determinado acontecimiento; por ejemplo cuando la información ha sido actualizada, trasladada de lugar o borrada (*WBI* de *IBM*, *BullsEye* y *Smart Bookmarks*). Este tipo de agentes permite tener alerta a un usuario frente a eventos en la red interesantes para el mismo. La forma en que este tipo de agentes sirve la información a su usuario puede ser el indicar únicamente qué página o páginas han cambiado y desde cuando ha sucedido esto o llegar a bajarse el texto de las páginas actualizadas, filtrando en este caso imágenes, gráficos y demás.

· **Mediador de diferentes fuentes de información**: se están realizando esfuerzos en la línea de desarrollar agentes que permitan interoperar a diferentes fuentes de información independientemente del sistema en que se hayan desarrollado.

## 3.4. Clasificación de los agentes inteligentes

Como consecuencia de una definición diáfana de agente inteligente, diferentes autores han propuesto una gran variedad de taxonomías o clasificaciones. A continuación se presenta una doble clasificación que pretende aclarar la tipología de agentes inteligentes basándonos tanto en el ámbito en el que actúan como en las tareas que llevan a cabo, para finalmente demarcar nuestro ámbito de actuación en los agentes de Internet que realizan tareas de recuperación de información.

En cuanto a su ámbito de actuación:

- Agentes de escritorio (agentes de sitema operativo, agentes de aplicaciones, etc.)
- Agentes de Internet (agentes de búsqueda, filtrado, recuperación de información, agentes de notificación, agentes móviles, etc.)
- Agentes de Intranet (agentes de customización cooperativa, agentes de bases de datos, agentes de automatización de procesos, etc.)

ISBN 84-9707-571-4

En cuanto a su función:

Stenmark clasifica los agentes en las siguientes tipologías: Interface agents, System agents, Advisory agents, Filtering agents, Retrieval agents, Navigation agents, Monitoring agents, Recommender agents, Profiling agents y otros que están surgiendo continuamente.

En el ámbito de la gestión eficiente del conocimiento, podemos destacar tres tipos:

1. **Filtering agents**: agentes que se usan para reducir la sobreabundancia de información mediante el borrado de los datos no deseados (por ejemplo, los datos que no satisfacen completamente el perfil de usuario). Muchos clientes de e-mail, así como los productos Agentware e InfoMagnet proporcionan prestaciones básicas de filtering agents.

2. **Retrieval agents**: agentes que buscan, recuperan y proporcionan la información como si fueran auténticos gestores de información y documentación ("information brokers"). Muchos productos se autoproclaman como retrieval agents, tanto aplicaciones cliente: AT1, BullsEye, Go-Get-It, Got-It, Surfbot y WebCompass; como aplicaciones servidor: Agentware e InfoMagnet.

3. **Monitoring agents**: proporcionan al usuario la información cuando sucede un determinado acontecimiento; por ejemplo cuando la información ha sido actualizada, trasladada de lugar o borrada. Algunos productos ejemplo son: WBI de IBM, BullsEye y SmartBookmarks.

En nuestro caso nos interesan los denominados "retrieval agents", esto es, agentes para la recuperación de información. Otros autores denominan a este tipo de software como agentes de información. Al fin y al cabo su función no se basa en la mera recuperación de información sino que disponen de un conjunto de utilidades conexas que nos permitirían denominarlos agentes para gestión del conocimiento.

## 3.5. Los agentes de recuperación semántica de la información

El papel del agente inteligente en el proceso de recuperación "semántica" de información no debe confundirse con la de un buscador inteligente. Un buscador

ISBN 84-9707-571-4

inteligente se aprovechará del enriquecimiento semántico de los recursos web para mejorar (principalmente en la precisión) la recuperación de información, aunque su funcionamiento se basará, como los actuales buscadores, en la previa indización de todos aquellos recursos susceptibles de ser recuperados.

En cambio, un agente inteligente recorrerá la Web a través de los enlaces entre recursos (hiperdocumentos, ontologías, ...) en busca de aquella información que le sea solicitada, pudiendo además interactuar con el entorno para el cumplimiento de tareas encomendadas. Por ejemplo, un agente inteligente, ante una consulta dada, podría consultar autónomamente un buscador, y a partir de sus resultados, explorar la Web hasta encontrar la información solicitada, pudiendo finalmente llevar a cabo una acción sobre dicho recurso, como podría ser la reserva de una habitación en un hotel.

## 3.6. Bibliografía

Adam, N. Y Yesha, Y (1996). Electronic Commerce and Digital Libraries: towards a digital agora. *ACM Computing Survey*, vol. 4, nº 28, diciembre de 1996.

Aguillo, I. F (1999). Del multibuscador al metabuscador: las agentes trazadores de Internet. En*: Congreso ISKO* (IV. Granada. 1999). La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de información. Granada: Isko: Universidad de Granada, 1999, p.239-245

Alonso Berrocal, J.L.; Figuerola, C.G. Y Zazo Rodríguez, A.F (1999). Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información. *IV Encuentros Internacionales sobre Sistemas de Información y Documentación*: IBERSID 99, Zaragoza, 15-18 de Marzo de 1999.

Brashaw, J (1997). An introduction to software agents. En: Brashaw, J. *Software agents*, AAAI Press, 1997, p. 4-7

Caglayan, A.; Harrison, C (1997). *Agent Sourcebook*. New York, etc.: Jonh Wiley & Sons, 1997.

Chaves, A... et al (1997). A Real-Life Experiment in creating an agent marketplace. En: *Proceeding of PAAM'97*. Practical Applications Company, 1997.

Codina, L (1997). Cómo funcionan los servicios de búsqueda en Internet: un informe especial para navegantes y creadores de información. Part I. *Information World en Español*, vol. 6, nº 5, 1997, p. 22-26

Eriksson, J.; Finne, N. Y Janson, S (1999). To each and everyone an agent:

ISBN 84-9707-571-4

augmenting web-based commerce with agents. *Intelligent Systems Laboratory. Final Report, 1999.*

Giles, C.L.; Bollacker, K.D.; Lawrence, S (1998). Citeseer: an autonomus web agent for automatic retrieval and identification of interesting publications. In: *2nd International ACM Conference on Autonomus Agents,* ACM Press, May, 1998.

Hípola, P.; Vargas-Quesada, B (1999). Agentes inteligentes: definición u tipología. Los agentes de información. *El profesional de la información*, vol. 8, nº 4, abril de 1999, p. 13-21

Klusch, M. (Ed.) (1999). *Intelligent information agents: agent-based information discovery and management on the Internet*. Berlin: Springer, 1999.

Leloup, C (1998). *Motores de búsqueda e indexación: entornos cliente servidor, Internet e Intranet*. Barcelona: Gestión 2000, 1998.

Maes, P (1994). Agents that reduce work and information overload. *Communications of the ACM*, vol. 7, nº 37, 1994, p. 31-44. Disponible en http://pattie.www.media.mit.edu/people/pattie.

Maldonado Martínez, A.; Fernández Sánchez, E (1998). Evaluación de los principales "buscadores" desde un punto de vista documental: recogida, análisis y recuperación de recursos de información. *Actas VI Jornadas Españolas de Documentación*, 1998, p. 529-551

Matthew, C (1998). Bridging intranet profit and value. *Datamation*, diciembre/enero 1998, p. 120-124

Nwana, J (1996). Software agents: an overview. *Knowledge Engineering Review*, 11(3), 1996, p. 205-244

Peis, E.; Herrera-Viedma,E.; Hassan Y. and Herrera, J.C (2003). Ontologías, taxonomías y agentes: recuperación "semántica" de la información. JOTRI 2003: II Jornadas de Tratamiento y Recuperación de Información, 8 y 9 de septiembre de 2003

Snyder, H.; Rosenbaum, H (1999). Can search engines be used as tools for web-link analysis? A critical review. *Journal of Documentation*, vol. 55, nº 4, 1999, p. 375-384

Stenmark, D (1998). *Intelligent Software Agents: a attempt to do a classification*, 1998 . http://w3.informatik.gu.se/~dixi/agent/agent.htm

Tramullas, J (1999). Agentes y ontologías para el tratamiento de la información:

ISBN 84-9707-571-4

clasificación y recuperación en Internet. En: *Congreso ISKO* (IV. Granada. 1999). La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de información. Granada: Isko: Universidad de Granada, 1999, p. 247-248

ISBN 84-9707-571-4

## 3.7. Caso práctico. Comparación Google versus Copernic

PRESENTACIÓN:

Ante una necesidad de información, el usuario establece qué fuentes de información van a ser consultadas para solucionar esta determinada necesidad. Mayoritariamente las fuentes de información serán los motores de búsqueda y los agentes inteligentes para la recuperación de información.

Para orientarnos sobre qué herramienta es más útil para satisfacer una necesidad de información debemos tener en cuenta el conjunto de prestaciones que un software de recuperación de recursos web puede ofrecer. En esta actividad se presenta una tabla de comparación que puede ayudar a determinar las prestaciones básicas de una software de búsqueda y recuperación de recursos web.

OBJETIVOS:

- Conocer en profundidad las funcionalidades de Google y Copernic
- Establecer una serie de criterios objetivos de comparación de motores de búsqueda y agentes inteligentes.

ENUNCIADO:

Atendiendo a la siguiente lista de valoración para recuperar información en Internet comparar el motor de búsqueda Google y el agente Copernic (versión avanzada) , a partir de vuestra experiencia en la realización de estrategias de búsqueda y aprendizaje de su funcionamiento.

Además, el estudiante debe realizar en una extensión de una página (verdana 12) un análisis crítico de los resultados obtenidos en la comparación de estas herramientas de búsqueda.

Para cumplimentar las columnas de Google y Copernic, debéis rellenar con un 1 si la herramienta cumple esa prestación y con un 0 si no lo cumple y con una X si no se dispone de información suficiente. Al final tendremos una suma total de prestaciones que nos informará sobre el nivel de sofisticación de cada una de las herramientas.

FORMATO:

El especificado en la tabla anexa

ISBN 84-9707-571-4

**Tabla para la comparación.**

| | | Google | Copernic |
|---|---|---|---|
| PERSONALIZACION | Búsqueda simple/avanzada | | |
| | Búsqueda booleana | | |
| | Truncamiento | | |
| | Lenguaje natural | | |
| | Filtros | | |
| | Métodos de ordenación | | |
| | Búsquedas por campos | | |
| | Configuración de presentación | | |
| CALIDAD RESULTADOS | Rating/Ordenación | | |
| | Exhaustividad | | |
| | Precisión | | |
| USABILIDAD | User friendliness/Facilidad de uso | | |
| | Limpieza de la pantalla | | |
| | Respuesta mínima (en espacio) | | |
| | Ayudas al usuario | | |
| | Inclusión de la url | | |
| | Información añadida | | |

ISBN 84-9707-571-4

| | | | |
|---|---|---|---|
| RENDIMIENTO | Tiempo de respuesta | | |
| | Periodicidad indexación | | |
| | Conceptos añadidos | | |
| | Páginas similares | | |
| FUNCIONES AÑADIDAS | Enlaces patrocinados | | |
| | Productos relacionados | | |
| | Retroalimentación/refinado | | |

**TOTAL**

ISBN 84-9707-571-4

## 3.7. Caso práctico. Vigilancia tecnológica con agentes

PRESENTACIÓN:

Un conjunto de empresas y centros de investigación europeos están preparando un proyecto de investigación con el objetivo de solicitar financiación del VI Programa Marco de I+D. El proyecto de investigación consiste en el desarrollo de nuevos materiales de plástico para envases y embalajes para su uso en horno y microondas.

El Departamento de Documentación del Instituto Tecnológico del Envase y Embalaje debe establecer un servicio de vigilancia tecnológica durante el periodo que dure la preparación del proyecto.

OBJETIVOS:

- Conocer los principales recursos para realizar una vigilancia del conocimiento tecnológico.
- Descubrir la utilidad de los agentes inteligentes como herramienta para la vigilancia del entorno y la gestión del conocimiento.
- Desarrollar un sistema de vigilancia tecnológica ad hoc.

ENUNCIADO:

Se trata de generar un informe en formato electrónico con los recursos web existentes sobre un tema muy especializado, para ello se utilizará un agente inteligente para la recuperación y monitorización de la información.

Utilizar el agente inteligente Copernic [versión permanente en español]

http://www.copernic.com/en/products/agent/download.html

Con el informe creado en html/xml, hay que grabarlo como “informe.html/xml” y pergarlo en un documento word.

Para tener conocimiento de los sistemas de vigilancia de estrategias de búsqueda y monitorización de sitios web hay que testear el sitio web http://www.cdetracker.com y suscribirse a uno o más canales.

ISBN 84-9707-571-4