La web invisible: búsqueda y recuperación de información oculta por medio de los buscadores convencionales

Miguel Ángel Amaya Ramírez

Introducción

En la actualidad, hablar de información en internet implica una enorme gama de posibilidades: desde cualquier tipo de información, como la que se encuentra en páginas personales, comerciales, de grupos de amigos, etcétera, hasta documentos académicos, como son artículos de revistas, tesis, reportes de investigación, libros y revistas en texto completo, diccionarios, enciclopedias, entre otras posibilidades.

La búsqueda de información por medio de internet tiene como principal recurso los buscadores de internet. Para su conformación, en algunos se realiza una labor por parte de personal experto, que busca, evalúa y registra la información que será almacenada en la base de datos del buscador; es el caso de Yahoo, en su modalidad de directorio, aunque la modalidad más utilizada en la actualidad es la de buscador.

Las bases de datos de los otros buscadores se incrementan por medio de robots o arañas (*spiders*) (Amaya, 2006), que rastrean la red en busca de información y automáticamente la dan de alta en sus bases de datos. Ejemplos de esto lo son Yahoo, en su modalidad de buscador, Google, Altavista, etcétera. Una característica de estos buscadores es que indexan principalmente páginas *web*, en código html, por lo tanto, cuando se realiza una búsqueda en internet, los buscadores rastrean en su base de datos y envían como resultados los vínculos a sitios *web* que contienen la información de la solicitud.

Asimismo, existen en internet documentos que se encuentran en diferentes formatos como pdf, ppt, doc, etcétera (Codina, 2003), formatos que para algunos motores de búsqueda no son visibles y que se encuentran dentro de la misma página *web*, o bien son recupera-

dos mediante bases de datos. A esta información que los buscadores no pueden recuperar de primera instancia, algunos autores le han llamado de diferente forma, por ejemplo, internet invisible, *web* profunda, *web* oscura, *web* invisible, etcétera.

En tal contexto, el objetivo del presente capítulo es mostrar métodos, técnicas y herramientas con las cuales se pueda recuperar información de calidad, y no únicamente navegar, sino sumergirse en las profundidades de la web *invisible*.

Breve semblanza del uso de internet

Uno de los fenómenos más importantes del siglo pasado es sin duda alguna la aparición de la red conocida como internet, que algunos autores han denominado red de redes o súper carretera de la información. Esta red es utilizada por una gran cantidad de personas de diferentes edades que se dedican a diversas actividades. El uso que las personas hacen de internet es heterogéneo, de acuerdo con sus necesidades. Por ejemplo, los estudiantes utilizan internet para buscar información y resolver tareas escolares, mientras que los empleados lo utilizan para actividades profesionales, los científicos para realizar proyectos colaborativos en tiempo real, algunas personas para realizar compras y otras más para enviar correos electrónicos.

Como se menciona en el estudio *Global internet statistics*, realizado por Global Reach (2004), en Estados Unidos el 90% de la población que utiliza internet lo hace para buscar información mediante los buscadores. Esto hace suponer que la mayoría de las personas que utilizan la red, lo hacen para cubrir alguna necesidad de información de diferente tipo.

Crecimiento de internet

Como es sabido, internet fue creado en 1969 por un proyecto del Departamento de Defensa de Estados Unidos llamado DARPANET (Defense Advanced Research Project Network). En esa época, internet apenas contaba con cuatro servidores interconectados entre ellos (Evolución de internet, 2001). Para 1971 contaba ya con 11 servidores; en 1975 se sumaron 24 servidores más y pasados 10 años acumulaba 188 servidores vinculados. Como se puede apreciar, en sus

primeros diez años de vida, internet tuvo un crecimiento paulatino pero es evidente que su mayor crecimiento todavía estaba por llegar.

En la década de los ochentas internet inició un crecimiento exponencial. Para 1983 había 562 servidores interconectados, mientras que para 1986, 1988 y 1989 los servidores vinculados fueron 5 089, 33 000 y 159 000 respectivamente. Pero sin duda alguna, el mayor crecimiento de la red se ha dado en los últimos quince años, puesto que, para el año 2000, internet tenía conectados 95 millones de servidores y actualmente cuenta con más de 260 millones de servidores.

Antecedentes de la recuperación de información en internet

Con este crecimiento de internet se vislumbró un problema: al existir una gran cantidad de información almacenada en todos los servidores integrados de forma constante a la red, era necesario contar con herramientas que permitieran su recuperación.

En 1990, Alan Emtage, estudiante de la Universidad McGill de Montreal, creó la primera herramienta para buscar información en internet, y la denominó Archie (Martínez y Oña, 1997), en honor al famoso personaje de historieta. En esa época no existía la *web* y la modalidad más generalizada para el intercambio de archivos era el *file transfer protocol* (ftp). *Archie* consistía de una base de datos de servidores ftp y un motor de búsqueda sencillo. Éste buscaba en los servidores los archivos que coincidieran con los términos de la búsqueda realizada por el usuario.

Archie fue el único medio para la recuperación de información en la red hasta 1993, en que surgió *Verónica* (también nombrada así en honor de la famosa personaje de historieta), desarrollado por la Universidad de Nevada en Estados Unidos. Fue concebido como una herramienta similar a la de *Archie*, pero para servidores *Gopher*, de búsqueda en internet, que para 1993 eran la aplicación más popular.

Inicios de la web

En la invención y desarrollo de la *web*, Tim Berners-Lee ha jugado un papel fundamental, tan es así que se le reconoce como el padre de la *web*. La genialidad de Berners-Lee, al inventar la *web*, radicó, sobre todo, en saber unir las piezas tecnológicas que en cierto momento

histórico del desarrollo de la computación, internet y las telecomunicaciones existían. Si bien, desde inicios de los ochentas, Berners-Lee empezó a experimentar sobre la idea de crear un programa que permitiera vincular documentos alojados en diferentes servidores, es hasta la finales de la misma década que logra su propósito: la invención de la *web*, así como las especificaciones lógicas y técnicas de sus principales elementos: direcciones url, el protocolo http y el código html. La visión de la *web* fue, desde sus inicios, como el mismo Berners-Lee lo narra: *la de cualquier cosa potencialmente conectada con cualquier cosa. Es una visión que nos proporciona una nueva libertad...* (Berners-Lee, 2000). Hacia 1993-1994, el modelo *web* empieza a crecer exponencialmente y miles de personas e instituciones empiezan a tener presencia en el sistema hipertextual y a subir contenidos de diversa índole, generándose todo un conjunto de repercusiones en todas las actividades humanas.

Berners-Lee utilizó la tecnología del hipertexto para enlazar de manera conjunta documentos, como si se tratase de una telaraña, que podían ser utilizados de cualquier manera para buscar información. Gracias a esto, surgió la *web*, además de iniciarse un desarrollo verdaderamente impresionante de recursos de información de todo tipo, disponibles a través de la *web*.

La *web*, mediante el modelo hipertextual, permite muchas relaciones posibles entre cualquier recurso digital y otros recursos existentes en la red. Berners-Lee implementó un sistema de navegación de hipertexto, el cual permite a los usuarios moverse libremente entre documentos sobre la red, sin importar el lugar de origen.

La *web*, con su extraordinario desarrollo e impacto, se ha convertido rápidamente en la principal herramienta de internet. Por ello, surgió también la necesidad de contar con sistemas de búsqueda y recuperación de sitios y páginas *web*, así como diversos recursos de información digital.

En tal entorno, en la segunda mitad de la década de los noventas, prevalecía el escenario idóneo para el surgimiento y desarrollo de la mayoría de los actuales motores de búsqueda. A la fecha, se estima que existen alrededor de 5 300 buscadores en la red, de los cuales 5 000 son internacionales y unos 300 son iberoamericanos. Esto implica una competencia muy fuerte entre las empresas o instituciones que han creado estos sistemas para posicionarse en la *web* y ganar la preferencia de los usuarios de la red.

Web invisible

Como se ha mencionado anteriormente, existe una gran cantidad de información en internet, a la cual se tiene acceso por medio de la *web*, pero uno de los problemas que se presenta tiene que ver con la búsqueda y recuperación de información, pues encontramos tanta información que en ocasiones es difícil discernir entre lo que sirve y lo que no (Amaya, 2006).

Asimismo, hemos señalado que diferentes autores han denominado con diferentes nombres a la información que no puede ser recuperada en primera instancia por los buscadores convencionales, por ejemplo: internet invisible, *web* profunda, *web* invisible, etcétera, pero para efectos de este documento se utilizará principalmente el de *web* invisible, en especial, porque la *web* es un servicio de internet que nos permite recuperar principalmente información.

Por lo tanto, a lo largo de este apartado se definirá lo que es la *web* invisible y se presentarán aquellos recursos que contienen información de calidad, así como las diferentes herramientas útiles para su uso y recuperación

Definición

El término *web* invisible fue utilizado por primera vez por Jill Ellsworth para denominar a la información que resultaba *invisible* para las máquinas de búsqueda convencionales en la *web* (Ellsworth, 1995).

Para Isidro Aguillo, la *web* invisible es "una fracción muy voluminosa de información que, aunque accesible a través de la red, por distintas razones no es indexada por los motores de búsqueda" (Aguillo, 2002).

Bergman (2000) señala que el hecho de buscar información en la *web* es como pescar en el océano con una red, donde se captura mucha información de la superficie, pero los contenidos de las *profundidades del océano*, que es donde se encuentra frecuentemente la información valiosa y de índole académica, no se recuperan tan fácilmente o difícilmente son arrastrados por esa red. Para entender mejor lo antes mencionado se reproduce una de las figuras del estudio de Bergman (figura 1).

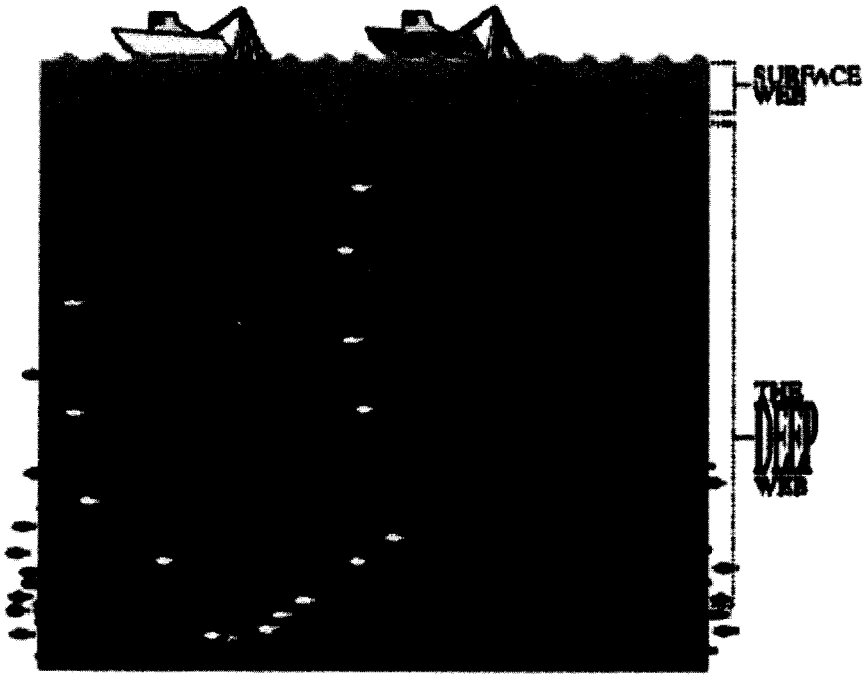

Figura 1. Metáfora visual de la *web* invisible. Fuente: Bergman, M. (2000). The deep web: surfacing hidden value. Documento en línea. Recuperado el 20 de julio, 2006 de: *http://www.brightplanet.com/images/stories/pdf/deepweb whitepaper.pdf.*

El mismo autor también la ha denominado *web profunda* (*deep web*), para diferenciarla de la *web superficial* (*surface web*), cuya información puede recuperarse por medio de los buscadores de internet (Bergman, 2000).

Los buscadores son un conjunto de programas instalados en un servidor conectado a la red, cuyo propósito principal consiste en que los usuarios puedan encontrar información. Estos mecanismos y el software que los apoya tratan de indexar toda la *web*, por lo que generan y mantienen enormes bases de datos. A pesar de su pretendida exhaustividad, se calcula que los motores de búsqueda más potentes indexan sólo entre un tercio y la mitad de la información y documentos disponibles en internet (Turner, 2005).

Otros autores visualizan a la *web* invisible como un gran *iceberg* en la inmensidad de un océano, es decir, comparan a la parte del

iceberg que sobresale del mar con la parte de la *web* que es indexada por los buscadores, y en la cual encontramos información de todo tipo, en ocasiones información no académica. Con respecto a la parte que no sale del mar, que es la proporción de hielo más grande, la comparan con la información de la web invisible, esa que sólo es recuperada por otros métodos, pero que cuenta con una gran cantidad de información de calidad o altamente académica.

Tamaño de la web *invisible*

Un estudio realizado en el año 2000 por Michael K. Bergman arrojó una serie de datos que pueden dar una idea del tamaño de la *web* invisible. En este estudio se menciona que la información pública contenida en la *web* invisible es de 400 a 550 veces más grande que el segmento común de la *web*, y crece a mucha mayor velocidad (Bergman, 2000).

Otros estudios calculan que el tamaño de la *web* profunda es 275 veces mayor que el de la *web* visible. En cambio, estimaciones posteriores señalan que el tamaño de la *web* invisible es sólo entre dos y 50 veces mayor que el de la *web* visible (Sherman y Price, 2001). Las diferencias en las cifras se deben a las diferentes metodologías utilizadas por los autores. En cualquier caso, el valor de la información contenida en la llamada *web* invisible justifica su estudio y el análisis de sus formas de acceso.

La *web* invisible contiene 7 500 *terabytes* de información, comparados con los 19 *terabytes* de información en la *web* superficial. La cifra de la *web* invisible equivale a 550 000 millones de documentos individuales, comparados con los 8 000 millones de la *web* visible que reporta Google, uno de los buscadores más grandes de la red y el más frecuentemente utilizado.

En el mismo estudio de Bergman se menciona que existen más de 200 000 sitios conocidos de *web* invisible. Como un ejemplo se menciona que 60 de los sitios más grandes de la *web* invisible contienen, en conjunto, cerca de 750 terabytes, lo que significa 40 veces más que la contenida en la *web* visible. Esto se debe a que los sitios de la *web* invisible tienden a ser más densos, en cuanto información, con un contenido más profundo que los sitios indizados en la *web* convencional.

Esto nos permite entender por qué algunos autores mencionan que la cantidad del contenido total de la *web invisible* es de 1 000 a 2 000 veces mayor que la de *web* superficial.

Contenido de la web *invisible*

Hasta este momento se ha definido qué es la *web* invisible y se ha tratado el debate acerca del tamaño de la misma, pero es necesario mencionar cuáles son los recursos de información que contiene. Al respecto, Isidro F. Aguillo agrupa dentro de la *web* invisible los siguientes recursos:

- Catálogos de bibliotecas y bases de datos bibliográficas.
- Bases de datos no bibliográficas.
- Revistas electrónicas.
- Documentos en formatos no indexables por todos los buscadores, como documentos pdf, doc, ppt, etcétera.
- Obras de referencia: diccionarios, enciclopedias, etcétera.

Con base en la tipología anterior, a continuación se presenta una explicación sobre los diversos recursos de información contenidos en la *web* invisible.

Catálogos de bibliotecas y bases de datos bibliográficas

El contenido de los catálogos de bibliotecas y las bases de datos bibliográficas normalmente queda aislado de la indexación y por lo tanto de la búsqueda y recuperación de información por parte de los buscadores convencionales, no obstante que tales catálogos y bases de datos estén integrados en plataformas *web*.

Como ejemplo de estos recursos podemos mencionar a LIBRUNAM, el catálogo electrónico creado y desarrollado desde 1978 por la Dirección General de Bibliotecas de la Universidad Nacional Autónoma de México, que contiene los registros bibliográficos de los libros que adquieren las 143 bibliotecas que integran su sistema bibliotecario. En la actualidad, es un catálogo de acceso público en línea (figura 2).

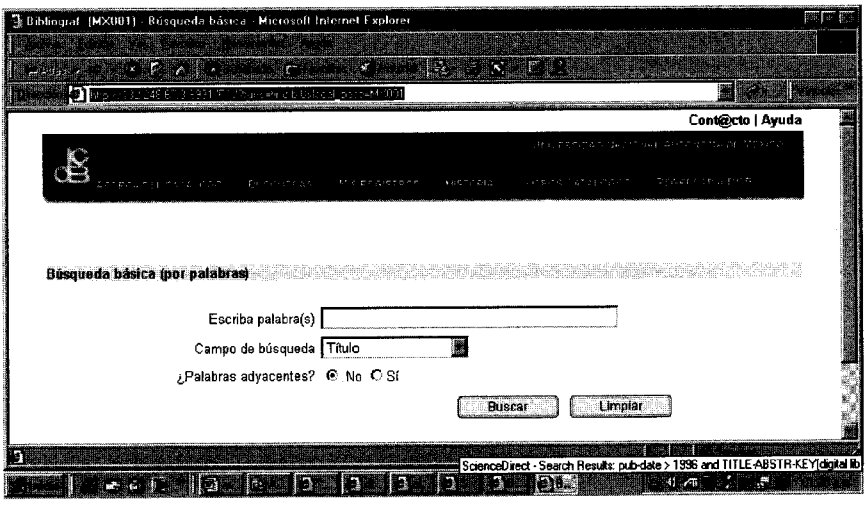

Figura 2. Página *web* de LIBRUNAM.

Bases de datos no bibliográficas

Se incluyen como bases de datos no bibliográficas desde las bases de datos alfanuméricas o de texto completo hasta las obras de referencia, tipo diccionario o enciclopedia. En este caso se presentan dos ejemplos: Scirus y Alphadictionary, que son considerados como bases de datos, pero cada cual presenta diferentes formas de recuperar información en la *web* invisible.

Scirus es un sistema de búsqueda y recuperación de información desarrollado por Elsevier, una de las más importantes casas editoriales a nivel mundial, especializada en la publicación de libros y revistas de carácter académico y científico. Este potente motor de búsqueda permite enviar las solicitudes de búsqueda de los usuarios a diferentes bases de datos integradas en la *web* invisible, incluidas las bases de datos generadas por Elsevier, relacionadas con los artículos y otros textos académicos, disponibles en texto completo, de las revistas y libros de la casa editorial (figura 3).

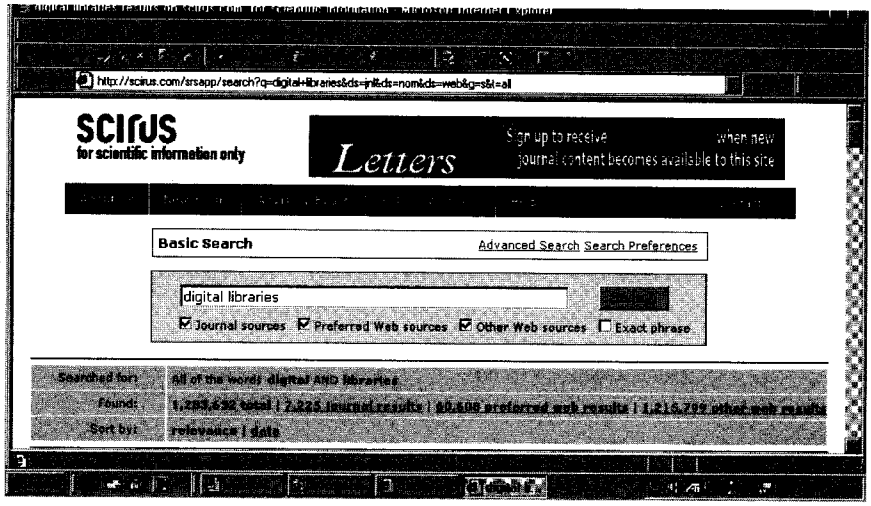

Figura 3. Página *web*.

Alphadictionary, por otra parte, hace búsquedas en aproximadamente 992 diccionarios y enciclopedias que se encuentran en la red (figura 4).

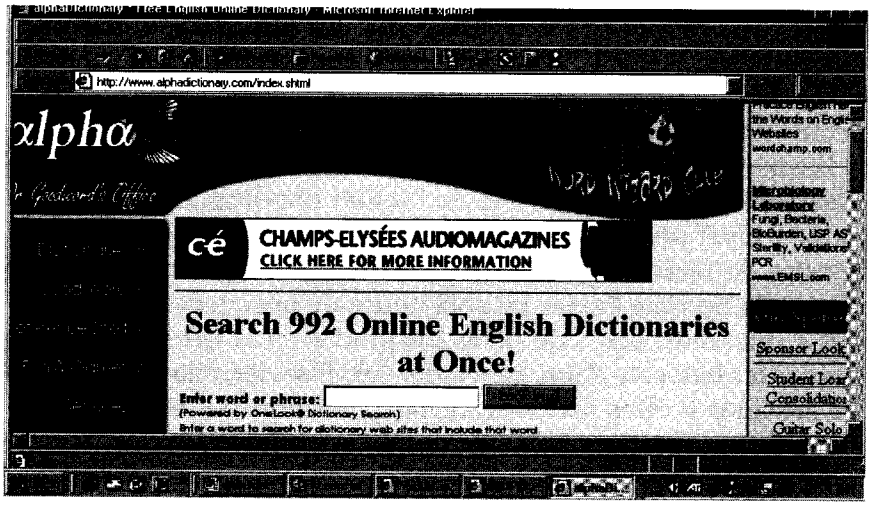

Figura 4. Página *web* de Alphadictionary.

Revistas electrónicas

Tanto las revistas de acceso gratuito que exigen registro previo como las de pago, protegidas por clave, son invisibles a los motores de búsqueda. A continuación se presentan dos ejemplos: ScienceDirect y DOAJ, directorio de revistas de acceso abierto.

ScienceDirect es un servicio que ofrece la editorial Elsevier para el acceso a los textos completos de los materiales que publica, sobre todo artículos de revistas académicas y científicas. La única condición para acceder a los textos en versión completa es contar con una suscripción (figura 5).

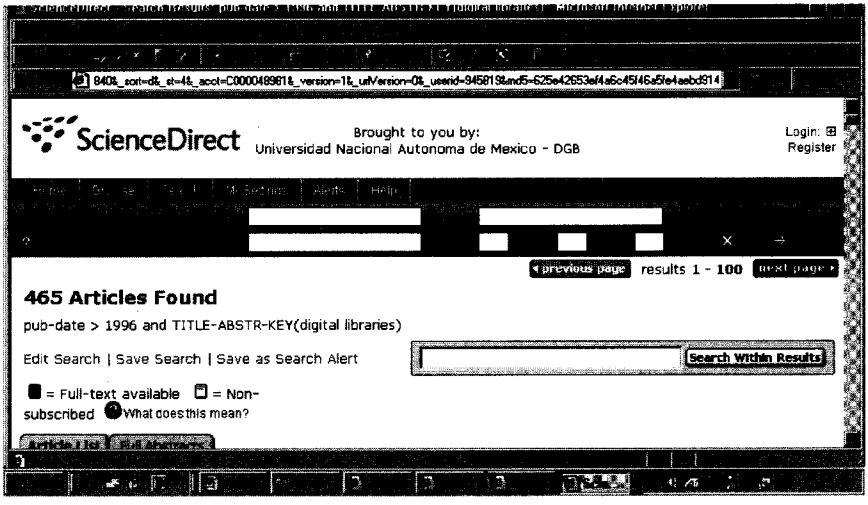

Figura 5. Página *web* de ScienceDirect.

DOAJ, el directorio de revistas de acceso abierto, promovido por la Sweden's Lund University, surgió en mayo de 2003. Ofrece el acceso gratuito al texto completo de cerca de 2 587 revistas académicas y científicas, en su mayoría, disponibles en la *web* mediante el modelo de acceso abierto (*open access*) y que contienen aproximadamente 127 423 artículos de diferentes áreas del conocimiento (figura 6).

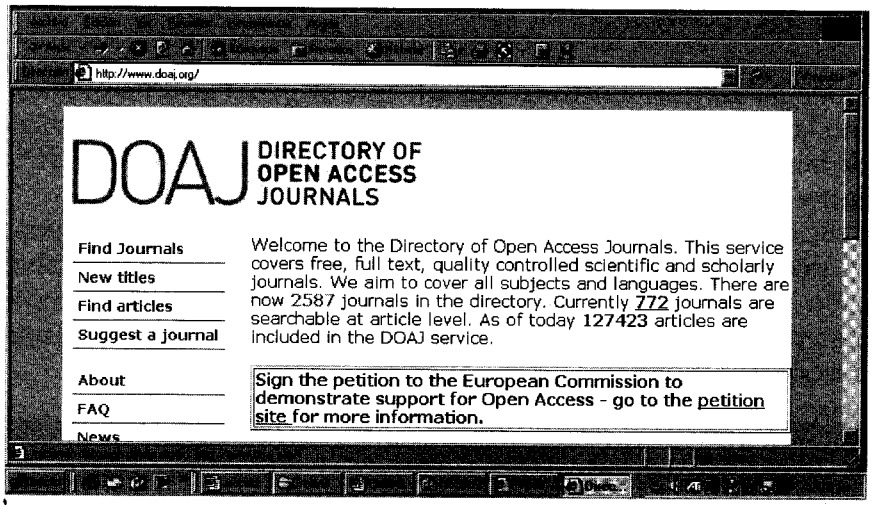

Figura 6. Página *web* de DOAJ.

Documentos en formatos no indexables por todos los buscadores, como documentos pdf, doc, ppt, etcétera

A continuación se ejemplifica la impresionante cantidad de documentos, en formatos diferentes a html, que se encuentran disponibles en la *web*, por ejemplo documentos pdf, postscript, doc, ppt, etcétera. Para ello se utilizó el buscador Google (cuadro 1):

Cuadro 1

Formato	*Extensión*	*Cantidad de documentos en Google*
Acrobat	*pdf*	*289 000 000*
Potscript	*ps*	*27 900 000*
Word	*doc*	*78 400 000*
Exel	*xls*	*16 900 000*
PowerPoint	*ppt*	*21 200 000*
Texto enriquecido	*rtf*	*6 990 000*
Flash	*swf*	*51 600 000*
Total		*491 990 000*

Cuadro 1. Documentos en formatos distintos a html, disponibles a través de Google.

Como se puede apreciar en este cuadro, la cantidad de documentos que existen con diferentes formatos en el buscador Google es de 491 990 000, los cuales, si no se sabe buscar en la *web*, resultan invisibles para los buscadores.

A continuación se presentan dos ejemplos de búsqueda por formato (figuras 7 y 8).

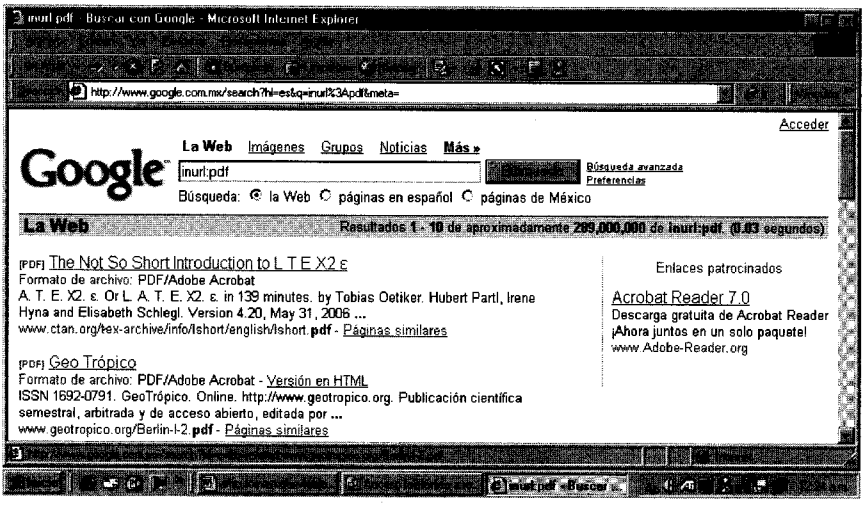

Figura 7. Búsqueda en Google de documentos pdf.

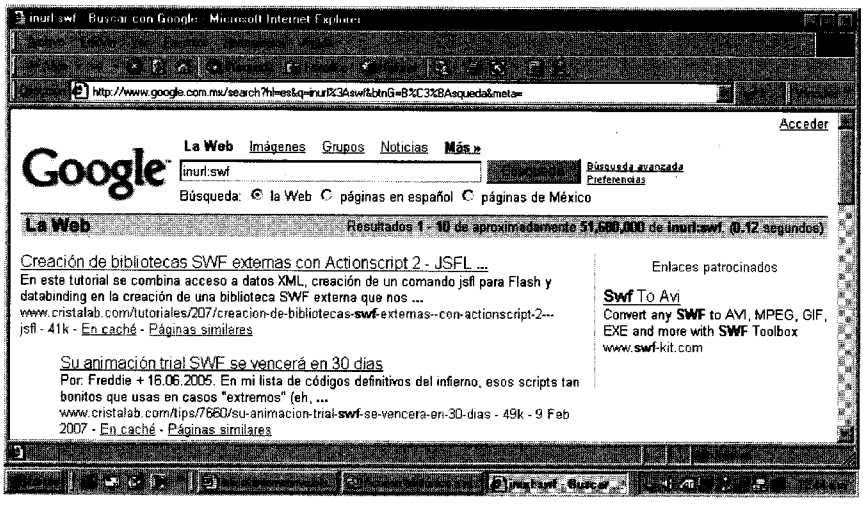

Figura 8. Búsqueda en Google de documentos flash, extensión swf.

La calidad del contenido en la web *invisible*

Como podemos observar, el contenido de la *web* invisible es altamente relevante para diversas necesidades de información, pues la información que la integra está respaldada principalmente por instituciones académicas, casas editoriales u organizaciones de mucho prestigio, que ofrecen sus recursos de información a los usuarios de la *web*. Pero esto rinde los frutos adecuados siempre y cuando se tengan habilidades y se manejen estrategias adecuadas para localizar los sitios y documentos en la *web*. Para ejemplificar esto, se reproduce una gráfica del estudio de Bergman, en la cual se presentan, con porcentajes, los tipos de recursos de información que integran la *web invisible* (figura 9).

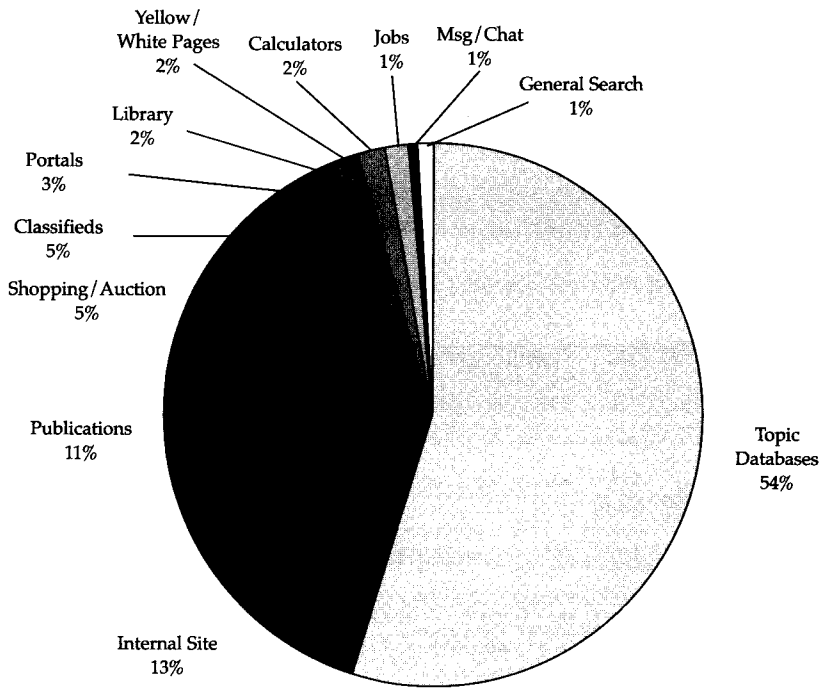

Figura 9. Distribución de recursos de información en la *web* invisible por tipo de contenido. Fuente: Bergman, M. (2000). The deep web: surfacing hidden value. Documento en línea. Recuperado el 20 de julio, 2006 de: *http://www.brightplanet.com/images/stories/pdf/deepwebwhite paper.pdf.*

Como podemos observar en la gráfica, existen tres rubros importantes que en conjunto suman 78% de los contenidos de la *web invisible:* bases de datos temáticas, sitios *web* internos y publicaciones en texto completo gestionadas por bases de datos.

A continuación se presentan algunos de los sitios que integran estos tres recursos de información, de los cuales el porcentaje mayor lo ocupan las bases de datos temáticas con un total de 54 %, de acuerdo con el análisis de Bergman. Entre los sitios que podemos mencionar se encuentran:

- Scirus: *http://www.scirus.com.*
- Ingenta: *http://www.ingenta.com.*
- ERIC: *http://www.eric.ed.gov.*
- Medline PubMed: *http://www.pubmed.gov.*
- Agricola: *http://agricola.nal.usda.gov/.*

En segundo lugar aparecen los sitios web internos de carácter restringido o privado, con un porcentaje del 13%. Son sitios que carecen de acceso libre al público en general, pues más bien están destinados únicamente a una audiencia específica, como es el caso de las intranets que, aunque se encuentran en línea, solamente pueden acceder a ellas las personas que cuentan con una clave.

Otro rubro importante lo conforman las publicaciones en texto completo, principalmente revistas académicas y científicas electrónicas, con un 11% del total del contenido de la *web invisible.* Ejemplos de directorios, etcétera, que permiten acceder a estos recursos de información son:

- Ejournal SiteGuide: *http://www.library.ubc.ca/ejour/.*
- Yale Medical Library: *http://www.med.yale.edu/library/journal.*
- Scholarly Journals: *http://info.lib.uh.edu/wj/webjour.html.*
- E-journals.org: *http://www.e-journals.org.*

La invisibilidad de información en la web

Como hemos podido darnos cuenta a lo largo de este documento, el proceso normal de búsqueda en la *web* consiste en lo siguiente: la información que se encuentra en los buscadores se almacena en

grandes bases de datos, los cuales al recibir una orden de búsqueda, consultan las bases de datos y ofrecen una respuesta a los solicitantes. Pero desgraciadamente esta información en ocasiones no es relevante o carece de un contenido académico. Esto ha ocasionado que en muchas ocasiones escuchemos a algunas personas mencionar que internet está llena información que consideran basura.

En tal contexto, si hemos afirmado que existe mucha más información valiosa en internet y que se puede recuperar por medio de los buscadores, ¿por qué buena parte de dicha información relevante es invisible?

Existen dos tipos de páginas que pasan desapercibidas para los motores de búsqueda:

- Páginas inaccesibles para los robots de búsqueda.
- Páginas excluidas por los robots de búsqueda.

Páginas inaccesibles para los robots de búsqueda

Como ya se indicó, las bases de datos de los buscadores son generadas por robots que navegan por internet y escudriñan en el contenido de páginas web estáticas, que para ser indexadas, en la mayor parte de los casos, deben estar vinculadas desde otras páginas *web*; si no existe un enlace a una página determinada, difícilmente el robot la puede detectar, pues al navegar, el robot va *saltando* de enlace en enlace. Estos robots de las bases de datos tampoco pueden registrar el contenido de las páginas a las que no pueden entrar, cuando es el caso de que se solicita clave de acceso o es necesario teclear varias opciones antes de que se despliegue el contenido.

Páginas excluidas por los robots de búsqueda

Algunas páginas estáticas son visibles y fácilmente clasificables para las arañas de búsqueda, aunque en ocasiones puedan ser invisibles por razones de autocensura de los buscadores (por motivos técnicos o políticos, como ha sucedido en el caso de sitios *web* de origen chino). Por otra parte, los directorios seleccionan, clasifican y jerarquizan una pequeña parte del contenido de internet, pero también los mo-

tores de búsqueda, que rastrean la red periódicamente, excluyen ciertas páginas *web*, de contenido efímero, por ejemplo, para no saturar sus enormes bases de datos y así hacer las búsquedas más rápidas y eficientes.

Los robots de los buscadores están diseñados, primordialmente, para trabajar con páginas *web* que contienen recursos de información en hipertexto, y por lo tanto, codificadas en html. Por esta razón algunos buscadores excluyen de sus resultados los recursos digitales que se encuentran en otros formatos. Hay que considerar, también, que si los objetos digitales tales como videos, imágenes, animaciones, etcétera, no contienen datos textuales asociados que los hagan indexables y recuperables, no podrán ser cubiertos por los buscadores, además de que éstos deberán contar con recursos técnicos potentes y secciones especiales para manejar tales recursos.

Sitios que contienen información de la web *invisible*

A continuación se presentan los sitios más relevantes para la recuperación de información contenida en la *web* invisible, organizados por el tipo de recurso de información.

Buscadores generales

En primera instancia se presentan buscadores generales que intentan recuperar información y documentos representados en diferentes formatos. Entre ellos podemos mencionar a:

- Google: *http://www.google.com.*
- Altavista: *http://www.altavista.com.*
- Alltheweb: *http://www.alltheweb.com.*

Este tipo de buscadores han logrado desarrollar, dentro de sus motores de búsqueda, métodos y técnicas para recuperar recursos de información normalmente invisibles, por ejemplo imágenes, grabaciones sonoras, videos, animaciones, y documentos que se encuentran en formatos tipo pdf, doc, ppt, etcétera. En la figura 10 se presenta el ejemplo de Google que, como puede observarse, permite buscar

en la *web* en general, o bien de manera especial: imágenes, grupos de discusión, noticias, etcétera. Mediante la búsqueda avanzada también es posible recuperar por tipo especial de documento: pdf, doc, ppt, etcétera.

Figura 10. Página *web*.

Otro aporte de Google para recuperar información relevante es la creación de Google Académico (figura 11), que permite buscar bibliografía especializada de una manera sencilla. Desde esta modalidad se pueden realizar búsquedas para un gran número de disciplinas y en fuentes como: estudios revisados por especialistas, tesis, libros, resúmenes y artículos provenientes de editoriales académicas, sociedades profesionales, universidades y otras organizaciones académicas, o que están en la red bajo el paradigma de acceso abierto, así como depósitos de artículos científicos en versión preimpresa, etcétera.

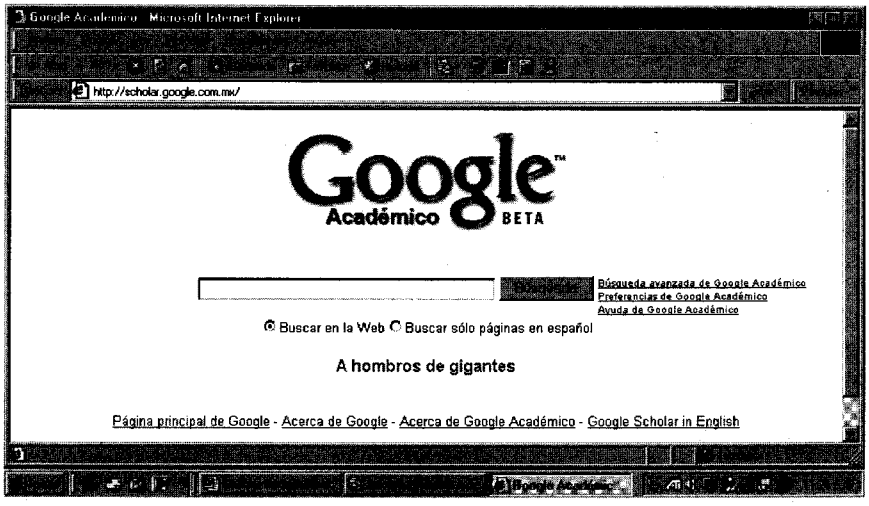

Figura 11. Página *web* de Google Académico.

Metabuscadores

También contamos con otras herramientas de recuperación de información, tales como los metabuscadores, que nos permiten buscar información, simultáneamente, en diferentes buscadores y que por lo tanto amplían notablemente las posibilidades de recuperar información que se encuentra en el espectro de la *web* invisible. Algunos ejemplos son:

- Vivísimo: *http://vivisimo.com.*
- Mamma: *http://www.mamma.com.*
- Turbo 10: *http://turbo10.com.*
- Copernic Agent Basic: *http://www.copernic.com/en/index. html.*
- SurfWax: *http://www.surfwax.com.*

Vivísimo

Vivísimo es un metabuscador que utiliza un software de categorización automática llamado *clustering*. Una vez realizada la búsqueda, el sistema agrupa los resultados por categorías y subcategorías. Proporciona el título y dirección del sitio o página *web*, un fragmento

del texto que contiene la información buscada y las fuentes donde se realizó la búsqueda, con indicación del número que ocupa en los resultados de dicho recurso (figura 12).

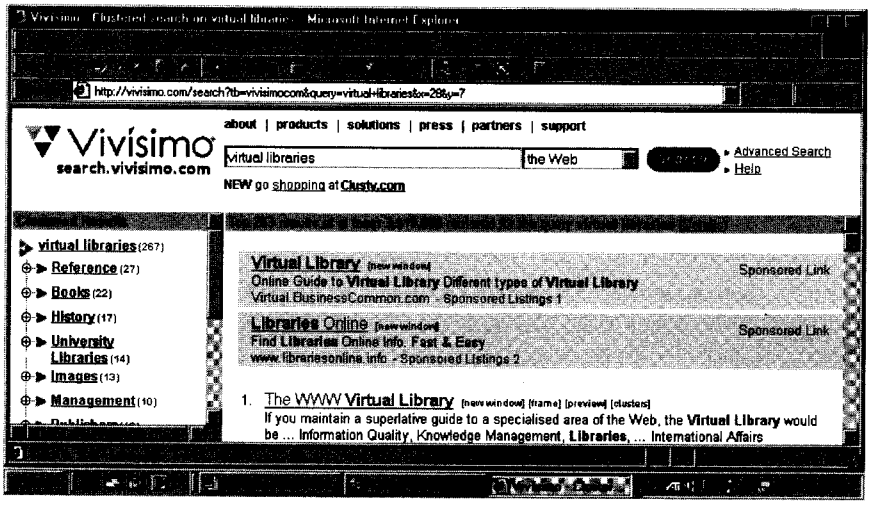

Figura 12. Página *web* de Vivísimo.

Mamma

Mama es un metabuscador que se autodenomina la madre de todos los buscadores. Contiene una sección que busca en bases de datos consideradas parte de la *web* invisible. Entre los recursos donde puede realizar búsquedas se encuentran: eMedicine Health, Health AtoZ, MayoClinic.com, Medem, MedicineNet.com, MedlinePlus, NHSDirect Online, etcétera (figura 13).

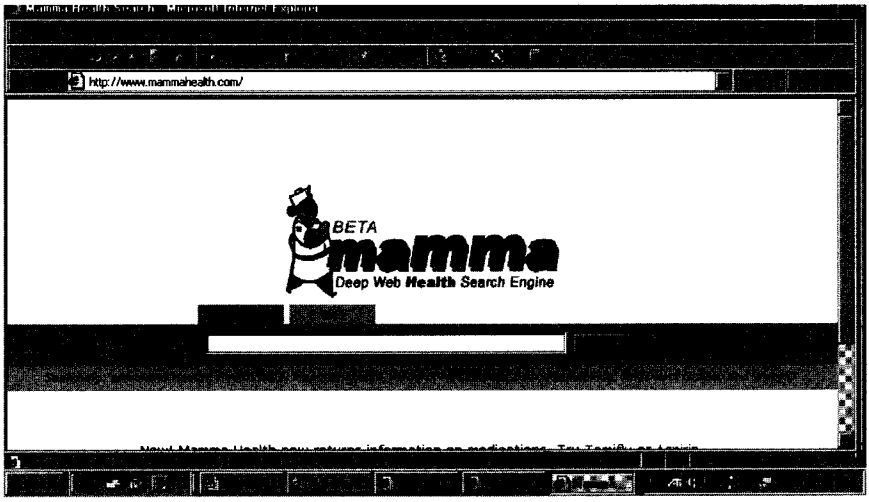

Figura 13. Página *web* de Mamma.

Turbo 10

Turbo 10 es un metabuscador que realiza consultas en unos 2 200 motores de búsqueda de la *web invisible,* aunque en la pantalla de búsqueda solamente hace referencia a 804 motores. Los recursos donde busca en primera instancia son: dmoz, mirago, wisenut y yahoo; también permite crear nuevas colecciones personalizadas, al seleccionar del listado un máximo de diez motores por colección. Los resultados muestran el título, dirección *web,* la fuente, texto donde aparece la palabra o palabras buscadas, etcétera. Además contiene una ventana que agrupa los recursos por temas y por motores de búsqueda (figura 14).

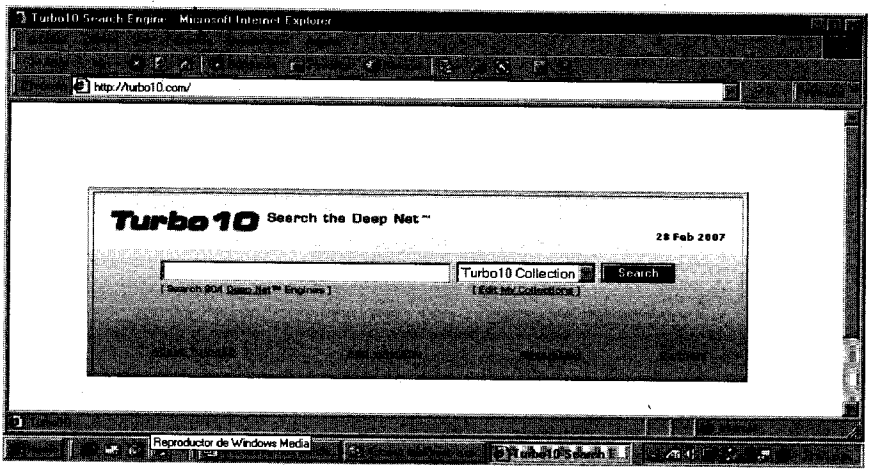

Figura 14. Página *web* de Turbo10.

Copernic Agent Basic

Copernic Agent Basic es un metabuscador que realiza búsquedas en una serie de motores, lo que depende de la selección geográfica seleccionada. Permite verificar y eliminar vínculos rotos y duplicados, utilizar filtros, navegar y buscar en los resultados, así como importarlos y exportarlos, guardar las páginas, conservar las búsquedas, etcétera. Algunas de estas opciones no están disponibles en esta versión básica pero sí en la versión de pago por suscripción (figura 15).

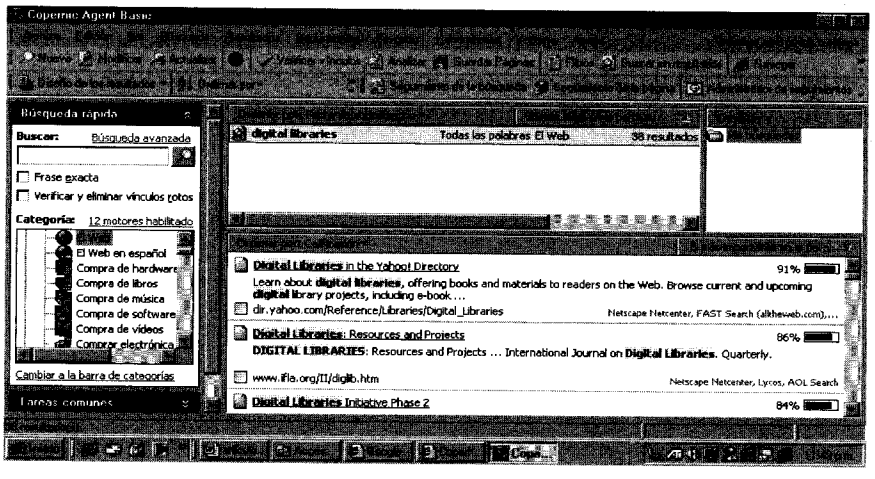

Figura 15. Página *web*.

SurfWax

SurfWax es un metabuscador que realiza consultas en las siguientes fuentes: Yahoo, Encarta, WiseNut, Yahoo News, CNN, AOL, LookSmart y MSN. Los resultados se pueden ordenar por relevancia, de forma alfabética y por fuente, además permite desplegar una ventana adicional donde se resume y extrae el contenido de un documento en tiempo real. También dispone de una sección de búsqueda de noticias (figura 16).

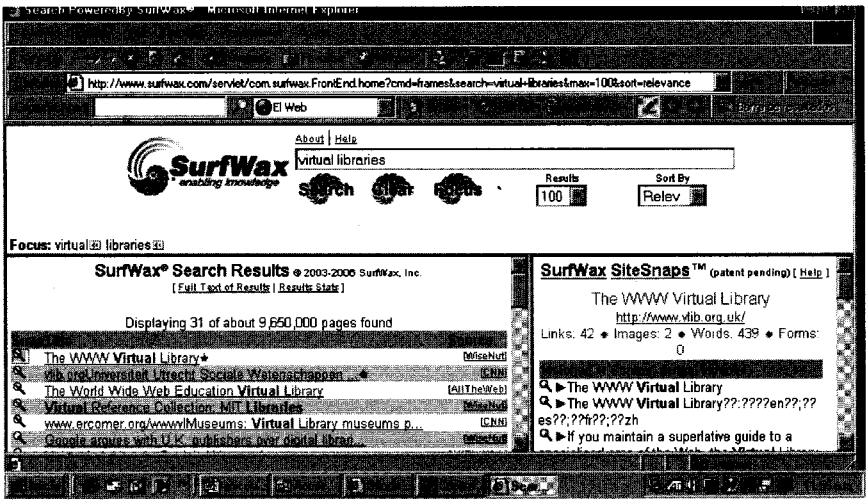

Figura 16. Página *web* de Surfwax.

Directorios o índices

Otros recursos de gran importancia para recuperar información de la *web* invisible son los directorios o índices. A continuación se presentan los directorios más potentes que nos permiten la navegación, búsqueda y recuperación de información en la *web* invisible:

Internet invisible

Internet Invisible es un índice que recopila, describe y ofrece el enlace a unas 2 500 bases de datos existentes en internet, se encuen-

tran organizadas en un directorio por grupos temáticos y materias específicas. Cuenta con un equipo de profesionales que se encargan de la selección, evaluación, recopilación y descripción de las mismas (figura 17).

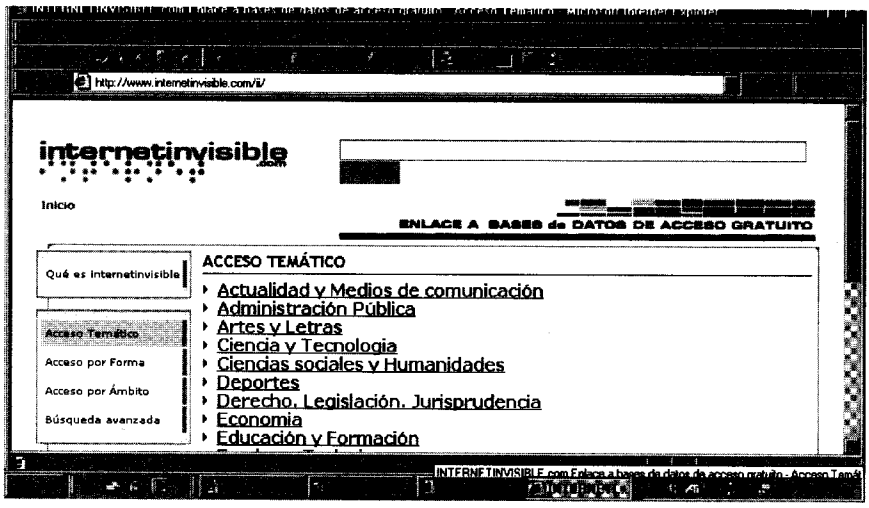

Figura 17. Página *web* de internet invisible.

CompletePlanet

CompletePlanet es un índice que recoge más de 70 000 bases de datos y buscadores especializados, organizados en un directorio temático. Contiene un listado completo de bases de datos dinámicas, las cuales contienen información relevante que no puede ser indizada por los buscadores de la *web visible*. Dispone de un formulario de búsqueda básico y otro avanzado. Los resultados incluyen el título del recurso, la dirección *web*, el resumen, la relevancia y el tamaño de la página (figura 18).

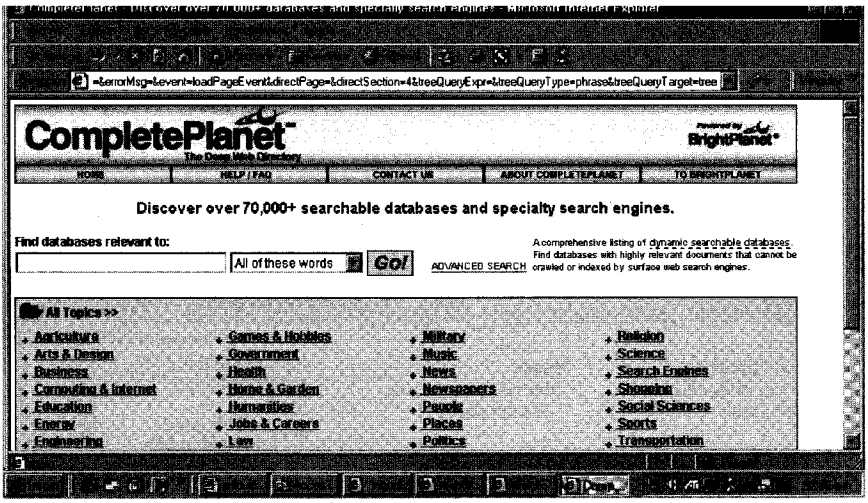

Figura 18. Página *web* de Complete Planet.

Librarians' Internet Index

Librarians' Internet Index es un directorio temático con más de 12 000 sitios de la *web*, seleccionados y evaluados por bibliotecarios, con base en su utilidad, calidad y relevancia. Tiene una categoría específica de salud y medicina donde los temas aparecen listados alfabéticamente. Ofrece el título, el resumen, materias con hipervínculos, especificación de la persona que ha creado el registro y fecha de la última actualización. Dispone de un formulario básico y otro avanzado (figura 19).

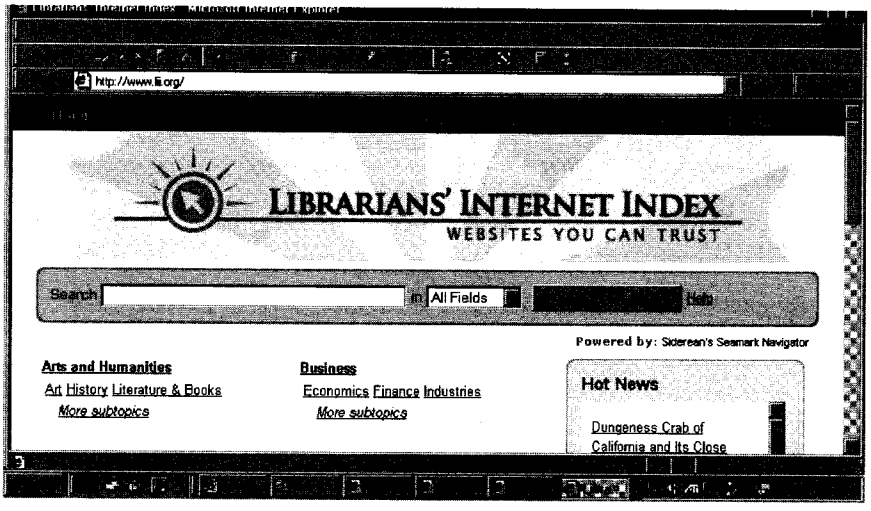

Figura 19. Página *web* de Librarians' Internet Index.

Infomine

Infomine es una biblioteca virtual con alrededor de 140 000 recursos de internet, provenientes del ámbito académico y universitario, entre los que se encuentran: bases de datos, revistas y libros electrónicos, artículos, etcétera. Permite buscar por categorías temáticas, por determinados campos e índices, por determinadas fuentes de información, por fuentes de acceso libre o de pago, por fuentes seleccionadas por bibliotecarios o por robots de búsqueda, etcétera. No contiene directorio temático. La información de cada recurso es muy completa (figura 20).

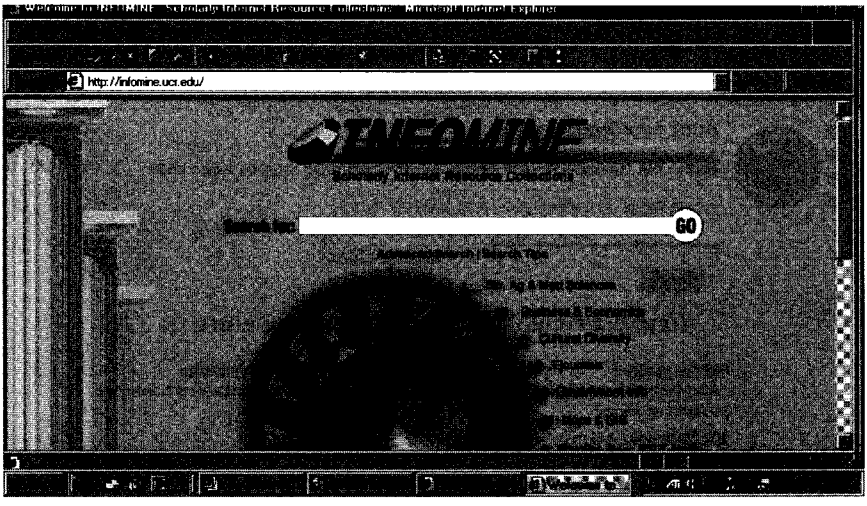

Figura 20. Página *web* de Infomine.

Bubl Link

Bubl Link es un directorio de recursos de internet con alrededor de 11 000 enlaces, que cubre todas las áreas del conocimiento. Las fuentes son seleccionadas, evaluadas, catalogadas y descritas. Las categorías temáticas principales se dividen en subcategorías. Proporciona el título, dirección *web*, resumen, autor, materia, clasificación, tipo de fuentes que contiene, localización y última verificación del enlace (figura 21).

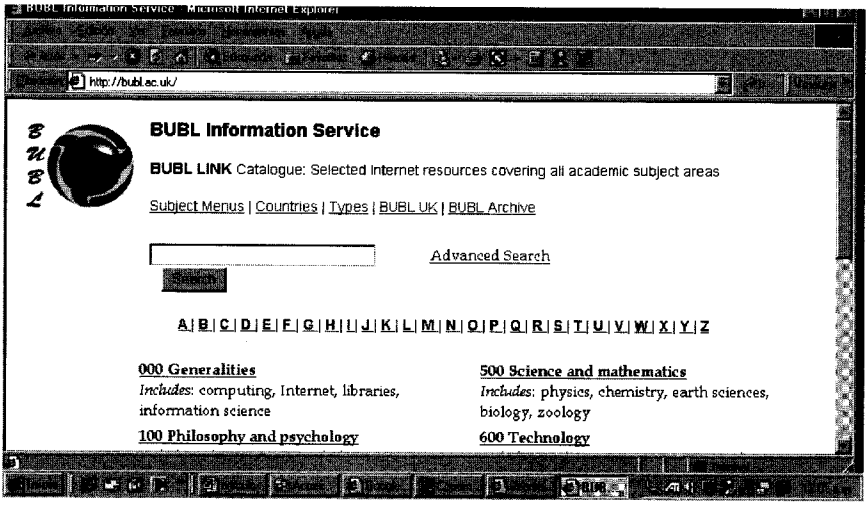

Figura 21. Página *web* de Bubl Link.

Web Brain

Web Brain es una interesante propuesta que rastrea bases de datos de la *web invisible*. Una vez que se realiza una búsqueda, se despliegan en la parte superior a manera de mapa conceptual, todos aquellos términos asociados a esa búsqueda de información (figura 22).

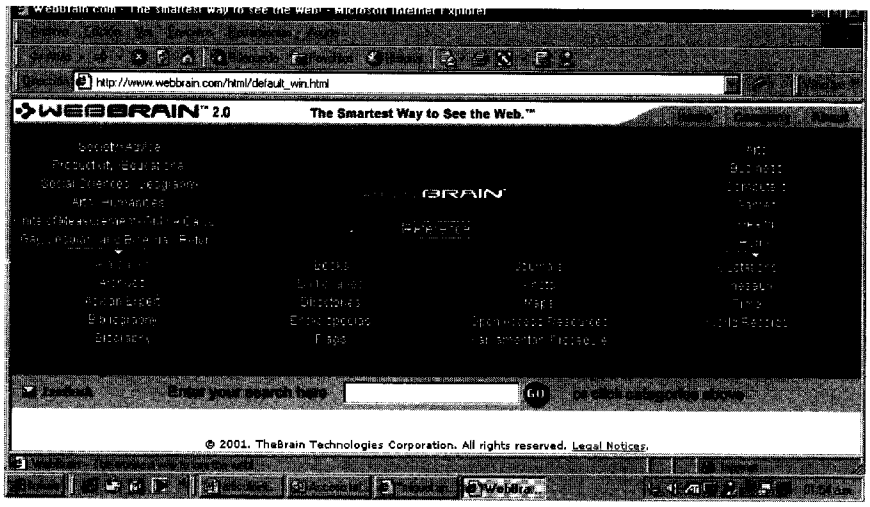

Figura 22. Página *web* de Web Brain.

Beaucoup

Beaucoup es un directorio y buscador de motores de búsqueda especializados, que además funciona como un metabuscador, pues cuenta con un motor de búsqueda que le permite realizar las búsquedas en los diferentes buscadores que cubre. Está organizado a manera de directorio, con 15 categorías principales y 42 subcategorias (figura 23).

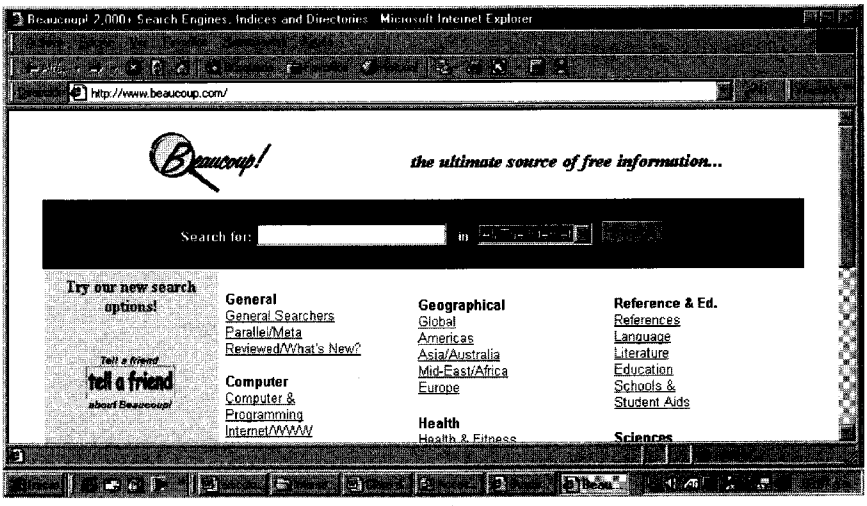

Figura 23. Página *web* de Beaucoup.

Conclusiones

Es importante que los bibliotecólogos dedicados a la búsqueda y recuperación de información en internet conozcan aquello que existe como parte no visible en la *web* convencional, la denominada *web* invisible.

Asimismo, al conocer cuáles son los tipos de documentos que integran la *web* invisible, se adopta un enfoque vital para la navegación, búsqueda y recuperación de este tipo de recursos de información. Esto permitirá a los bibliotecólogos ofrecer una mayor cantidad de información relevante para satisfacer las necesidades de información de sus usuarios y porque no, las de ellos mismos.

Es necesario que los bibliotecólogos conozcan cómo funcionan los recursos que permiten el acceso a la información de la *web* invisible y en dónde pueden localizarlos, para así poder explotar la *web* en toda la extensión de la palabra y para destruir ese mito de que internet y principalmente la *web* está llena de información chatarra o basura.

Si aprendemos a trabajar con las herramientas que nos permiten recuperar información de sitios y documentos que conforman la web invisible, nuestro panorama de acción en la recuperación de información se ampliará no diez ni veinte, sino una cantidad de veces sensiblemente superior a la que nos proporcionan los métodos tradicionales de búsqueda en la *web*.

Lista alfabética de los sitios *web*, buscadores y servicios mencionados en el capítulo

- Agricola: *http://agricola.nal.usda.gov/*.
- Alphadictionary: *http://www.alphadictionary.com/*.
- Altavista: *http://www.altavista.com*.
- Alltheweb: *http://www.alltheweb.com*.
- Beaucoup: *http://www.beaucoup.com/*.
- Bubl Link: *http://bubl.ac.uk/*.
- CompletePlanet: *www.completeplanet.com/*.
- Copernic Agent Basic: *http://www.copernic.com/en/index. html*.
- DOAJ, Directory of open access journals: *http://www.doaj.org/*.
- E-journals.org: *http://www.e-journals.org*.
- Ejournal SiteGuide: *http://www.library.ubc.ca/ejour/*.
- ERIC, Education Resources Information Center: *http://www.eric.ed.gov/*.
- Google: *http://www.google.com/*.
- Infomine: *http://infomine.ucr.edu/*.
- Ingenta: *http://www.ingenta.com/*.
- Internet invisible: *http://www.internetinvisible.com/ii/*.
- Librarians' Index to the Internet: *http://lii.org/*.
- Librunam: *http://www.dgbiblio.unam.mx/*.
- Mamma: *http://www.mamma.com*.
- Medline PubMed: *http://www.pubmed.gov*.
- Scholarly Journals: *http://info.lib.uh.edu/wj/webjour.html*.
- ScienceDirect: *http://www.sciencedirect.com/*.

- Scirus: *http://www.scirus.com/.*
- SurfWax: *http://www.surfwax.com.*
- Turbo 10: *http://turbo10.com.*
- Vivísimo: *http://vivisimo.com.*
- Yahoo: *http://www.yahoo.com/.*
- Yale Medical Library: *www.med.yale.edu/library/.*
- Web Brain: *http://www.webbrain.com/.*

Referencias

AGUILLO, I. (2000). Internet invisible o infranet: definición, clasificación y evaluación. En *VII Jornadas Españolas de Documentación.* Bilbao: Universidad del País Vasco.

AMAYA RAMÍREZ, M. A. (2006). Estrategias de búsqueda para la recuperación de información en la *web.* En H. A. Figueroa Alcántara, C. A. Ramírez Velázquez (Coord.). *Servicios bibliotecarios.* México: UNAM, Facultad de Filosofía y Letras: Dirección General de Asuntos del Personal Académico.

BERGMAN, M. K. (2000). *The deep Web: surfacing hidden value.* Documento en línea. Recuperado el 20 de julio, 2006 de: *http://www.brightplanet.com/images/stories/pdf/deepwebwhitepaper.pdf.*

BERNERS-LEE, Tim. *Tejiendo la red.* Madrid: Siglo XXI, 2000.

CODINA, L. (2003). *Internet invisible y web semántica: ¿el futuro de los sistemas de información en línea?* Documento en línea. Recuperado el 8 de julio, 2006 de: *http://www.lluiscodina.com/articulos/web semantica.pdf.*

ELLSWORTH, J. (1995). *Marketing on the Internet: multimedia strategies for the World Wide Web.* New York: John Wiley & Sons.

Evolución de internet (2001). Documento en línea. Recuperado el 20 de junio, 2006 de: *http://www.alu.ua.es/r/rac6/HInternet/origenes.html.*

GLOBAL REACH (2004). *Global internet statistics.* Documento en línea. Recuperado el 3 de abril, 2007 de: *http://global-reach.biz/globstats/index.php3.*

MARTÍNEZ, M. y OÑA, A. (1997). Aplicación de las comunicaciones y nuevas tecnologías al campo del aprendizaje motor. *Motricidad,* 3, 89-108.

SHERMAN, C. y PRICE, G. (2001). *The invisible web: uncovering information sources search engines can't see.* Medford, New Jersey: Cyber Age Books.

TURNER, L. (2005). *Deep Web Search Tools*. Documento en línea. Recuperado el 20 de junio, 2006 de: *http://www.bhsu.edu/education/edfaculty/lturner/deep_web_search_tools.htm*.