Xarxa Punt TIC

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

La web privada / la web propietaria / la web realmente invisible

plan avanza2
GOBIERNO DE ESPAÑA
PlanE
MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES Y PARA LA SOCIEDAD DE LA INFORMACIÓN
Generalitat de Catalunya

➔ ÍNDICE

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

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

Los directorios anotados o guías pueden incluir, además, algún mecanismo de búsqueda en sus páginas o en la Web en general (Moreno Jiménez, 2004). Comúnmente no basta con conocer la variedad de herramientas de búsqueda disponibles en la Web, sino que se requiere una orientación sobre su funcionamiento, sobre qué estrategias seguir para trazar una adecuada ruta de búsqueda y sobre cómo elegir los mejores instrumentos para cada necesidad. De ello se ocupan los tutoriales. *How to Choose a Search Engine or Directory,* de la Universidad de Albany, en Estados Unidos, y las guías de *SearchAbility* y de la Universidad de Leiden en Holanda *A Collection of Special Search Engines* orientan al usuario en el amplio mundo tanto de los recursos especializados en la Web como de las maquinarias que permiten su localización.

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

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

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

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

- ➔ Extensión de la indización: por economía, no todas las páginas de un sitio son indizadas en los buscadores.

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

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

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

La Web propietaria incluye aquellas páginas en las que es necesario registrarse para tener acceso al contenido, ya sea de forma gratuita o pagada. Se dice que al menos el 95% de la Web profunda contiene información de acceso público y gratuito (Turner, 2003).

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

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

Herramientas de búsqueda en la Web profunda

Los motores de búsqueda han mejorado su desempeño en los últimos años, permitiendo un mayor nivel de precisión en las búsquedas y ofreciendo los resultados en formas cada vez más convenientes para el usuario, pero aún son muchos los buscadores que sólo pueden recuperar directamente la información

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

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

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

Estrategias de búsqueda en la Web profunda

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

Para la búsqueda de información especializada:

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

información académica de calidad.

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

Para realizar búsquedas avanzadas:

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

Para evaluar la información disponible en la Web:

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

Para buscar información en bases de datos:

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

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

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

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

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

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

El segmento de la Web privada no representa una pérdida de gran valor, en términos de la información que contiene, ya que en general se trata de documentos excluidos deliberadamente del circuito informacional por su escasa utilidad. En cualquier caso, son los dueños de la información los que deciden no hacerla disponible, por lo que difícilmente se podrán encontrar mecanismos legítimos para franquear esa barrera. Además, los archivos robots.txt sirven para evitar que los robots caigan en “agujeros negros”, que les hagan entrar en procesos circulares interminables, mermando así la eficiencia en su funcionamiento.

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

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

La Web pública está dominada por contenidos originados en los Estados Unidos, escritos en inglés. Esto nos lleva a pensar que probablemente haya

más recursos invisibles en páginas originadas en otros países (distintos a los Estados Unidos) y en otros idiomas.

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

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

Como hemos visto, la definición más genérica de lo que constituye la Web invisible o profunda apunta a los recursos que no pueden ser recuperados mediante las herramientas comunes de búsqueda. Para verificar la visibilidad de la Web profunda, que ha sido identificada por los autores de *The Invisible Web*, Moreno Jiménez (2003) ha seleccionado al azar diez recursos de su *The Invisible Web Directory* y realizó la búsqueda en un buscador, un directorio, un metabuscador y un agente metabuscador avanzado en su versión gratuita. Los resultados de esta sencilla prueba aparecen reflejados en el cuadro siguiente:

Recurso	**MSN**	**Yahoo**	**MetaCrawler**	**Copernic**
Artcyclopedia	SI	SI	SI (6 buscadores)	SI (8 buscadores)
CRA Forsythe List	SI	SI	SI (3 buscadores)	SI (5 buscadores)
Current Films in the Work (BoxofficeHollywood Hot Set)	SI	SI	SI (3 buscadores)	SI (4buscadores)
Employee Benefits	SI	SI	SI (2	SI (3

INFOSOURCE			buscadores)	buscadores)
Hamnet	SI	SI	SI (4 buscadores)	SI (6 buscadores)
Infonation	SI	SI	SI (5 buscadores)	SI (7 buscadores)
Jourlit	SI	SI	SI (3 buscadores)	SI (7 buscadores)
Scholarly Societies Project	SI	SI	SI (4 buscadores)	SI (6 buscadores)
Vessel Registration Query System	SI	SI	SI (2 buscadores)	SI (6buscadores)
Who's who in American Art (AskArt)	SI	SI	SI (6 buscadores)	SI (8 buscadores)

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

Todos los recursos seleccionados de *The Invisible Web Directory* son localizables con las actuales herramientas de búsqueda. Además, en los resultados se observa que existen múltiples referencias en otras páginas, es decir, que se trata de páginas "conectadas". La única dificultad para encontrarlas consiste, en algunos casos, en las palabras con las cuales se denomina el sitio o el recurso. Por ejemplo, en el *The Invisible WebDirectory* aparece "Vessel Query Registration System", en lugar de "Vessel Registration Query System", lo cual hace que la búsqueda por todas las palabras sea exitosa, pero la búsqueda por frase no. Igualmente, la denominación de "Who's who in American Art" para el sitio de "AskArt" dificulta la búsqueda, mientras que si se busca directamente por su nombre aparece en numerosos buscadores. La tabla refleja además cómo el solapamiento entre buscadores es variable.

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

- ➔ Las páginas desconectadas.
- ➔ Las páginas no clasificadas que contienen principalmente imágenes,

- audio o vídeo.
- El contenido de las bases de datos relacionales.
- El contenido que se genera en tiempo real.

Pero:

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