Capítulo 4:

Cómo se llega al Sitio Web

Resumen

En la Guía Web Versión 1.0 se hizo una mínima referencia a la importancia de que el Sitio Web estuviera bien indexado en los sistemas de búsqueda en Internet, debido a que para la fecha de esa edición, éstos no tenían la relevancia que han adquirido con el tiempo. Esto se corrige en la presente edición, debido a que actualmente un Sitio Web corre el riesgo de "no existir" si sus contenidos no han sido indexados por los sistemas de búsqueda y recuperación de información y no tan sólo a través de la búsqueda del nombre de la institución, sino también, a través de los principales temas a los que ésta se dedica.

Este capítulo introduce el concepto de la Encontrabilidad que es una traducción libre del término inglés "findability", el cuál se puede entender como la "habilidad para ser encontrado". Para efectos de esta versión de la Guía Web, entenderemos la "Encontrabilidad" como la facilidad para que los contenidos de un Sitio Web puedan ser indexados y luego encontrados por sistemas de búsqueda externos e internos.

En el Sitio Web este capítulo es presentado en el menú con el nombre de "Encontrabilidad".

Tabla de Contenidos
Capítulo 4 - Cómo se llega al sitio web

> Introducción / Cómo se llega al sitio web

En la Guía Web Versión 1.0 se hizo una mínima referencia a la importancia de que el sitio web estuviera bien indexado en los sistemas de búsqueda en Internet, debido a que para la fecha de esa edición, éstos no tenían la gran relevancia que han adquirido con el tiempo.

Esta situación se corrige en la presente edición de la Guía, debido a que actualmente un sitio web corre el riesgo de "no existir" si sus contenidos no han sido indexados por los sistemas de búsqueda y recuperación de información y no tan sólo a través de la búsqueda del nombre de la institución, sino también, a través de los principales temas a los que ésta se dedica. Debido a lo anterior, se debe hacer un trabajo permanente tanto en la sección visible del sitio web (contenidos, imágenes y otros elementos similares) como en la invisible para los usuarios (keywords, meta tags, archivos destinados a robots de búsqueda), para asegurar que ellos siempre puedan llegar al sitio web a través de los buscadores.

Findability: facilidad para que los contenidos de un sitio web puedan ser indexados y luego encontrados por sistemas de búsqueda externos e internos.

Por lo tanto, este capítulo introduce el concepto de la Encontrabilidad que es una traducción libre del término en inglés "findability" que se puede traducir como la "habilidad para ser encontrado". Entonces, para efectos de esta versión de la Guía Web entenderemos la "Encontrabilidad" como la facilidad para que los contenidos de un sitio web puedan ser indexados y luego encontrados por sistemas de búsqueda externos e internos. Esta capacidad será el elemento clave que nos permita asegurar que los contenidos que se ofrecen a través del Sitio Web estarán adecuadamente indexados, facilitando de esa manera el acceso directo a la información desde los sistemas de búsqueda.

Para cumplir con este objetivo tan importante, los administradores de los Sitios Web de Gobierno deberán tener presente la importancia de llevar a cabo las tareas relacionadas con los siguientes aspectos:

- **1. Generación de los contenidos:** se refiere a que los contenidos del Sitio Web deben ser escritos teniendo en mente la forma en que el usuario final denomina a los temas que incluye el sitio. Se debe considerar que si se habla en el lenguaje del usuario, será más fácil que un sistema de búsqueda muestre entre

1.- Ver Capítulo II de la Guía Web 1.0 en http://www.guiaweb.gob.cl/guia/capitulos/cuatro/objetivos.htm

sus resultados los contenidos ofrecidos por el sitio ya que contendrán las mismas palabras utilizadas por el usuario que busca.

- **2. Presentación de los contenidos:** se refiere a que las páginas del sitio web deben ser preparadas para que tengan una estructura válida, en la cual haya consistencia entre los elementos de titulación y los contenidos propiamente tales, de manera que no haya alguna discordancia que las lleve a ser penalizada por los buscadores.

- **3. Apoyo a los robots de búsqueda:** se refiere al adecuado manejo de las cabeceras de las páginas (es decir, los contenidos de la etiqueta <head>); el contenido del archivo robots.txt; la generación del archivo estándar sitemaps.xml y la revisión del sitio mediante herramientas que simulan la acción de un "spider" de búsqueda.

Más información sobre Peter Morville se puede encontrar en el sitio **http://findability.org/.**

- **4. Monitoreo de sistemas de búsqueda:** se refiere que se debe prestar atención permanente a los sistemas que reflejan la forma en que los usuarios acceden al sitio web, ya que de esta manera se podrá entender qué palabras están utilizando para ese efecto y optimizar el contenido para reflejar dichos términos.

> ¿Qué es la Encontrabilidad?

Uno de los autores que ha apoyado con mayor fuerza el concepto de la Encontrabilidad es Peter Morville, a través de un texto del año 2006 titulado "Ambient Findability"[2] cuyo epígrafe indica "lo que encontramos nos transforma"[3].

Basado en la premisa "no puedes usar lo que no puedes encontrar" el autor destaca la importancia que tiene en la forma actual de usar la Internet, la capacidad de ser indexado puesto que es la forma más habitual que tendrán los usuarios de acceder al sitio web y emplear los contenidos que se ofrecen.

Adicionalmente, define el concepto de tres formas:

- 1. La capacidad de ser ubicado o de ser navegable.
- 2. El grado en el que un objeto determinado es fácil de descubrir o ubicar.

2.- Más información del libro en la Editorial O'Reilly: http://www.oreilly.com/catalog/ambient/
3.- Traducción libre de la frase "What We Find Changes Who We Become".

- **3.** El grado en el que un sistema o ambiente apoya la navegación y recuperación por sus contenidos.

Por lo mismo, podemos entender que la calidad de "encontrable" de un sitio web dependerá de dos aspectos: su capacidad para ser encontrado a través de los sistemas de búsqueda de Internet y, una vez que los usuarios decidan llegar al sitio web, de la calidad de la navegación que encuentren internamente en el propio sitio.

Esto representa desafíos interesantes para el administrador del sitio web, ya que siempre deberá estar mirando ambos aspectos para tener la seguridad de que la experiencia que se ofrece a través de sus páginas, es coherente con las expectativas de quien llega a visitarlo.

> Sitios visibles e invisibles

Uno de los desafíos más importantes de todo administrador de un sitio web consiste en permitir que sus contenidos sean indexados por los sistemas de búsqueda de Internet.

Esta característica debe tener en cuenta el hecho de que el sitio web debe estar tanto preparado para ser indexado por sistemas automatizados, los cuales están basados en programas (conocidos como robots de búsqueda o spiders) que navegan a través de los enlaces ofrecidos por el sitio web y que le permiten descubrir las páginas de contenidos disponibles. Lo anterior significa que cada página debería ofrecer enlaces en lenguaje HTML[4] hacia el resto del sitio web y, por lo mismo, que ninguna página del sitio debería estar aislada del resto.

Para apoyar esta tarea, que ya se revisó en detalle en el Capítulo 3 en el subtítulo referido a "Sistema de Navegación", es imprescindible que haya enlaces en cada una de las páginas que hagan referencia al resto del Sitio Web, en particular que lleven a la Portada y al Mapa del Sitio. Esta última página, a la que siempre se le da poca importancia, cobra a partir de esta circunstancia una relevancia mayor ya que es una colección de enlaces que debe ser visitada por el sistema de indexación de los buscadores porque constituye el punto de entrada al sitio web.

En este sentido es interesante tener en cuenta el trabajo "Características de la Web Chilena 2006"[5] llevado a cabo por el Centro de Investigación de la Web de la

4.- Enlaces del tipo `<a href.=...>` que puedan ser seguidos por los robots.

Universidad de Chile, que dirige el profesor Ricardo Baeza-Yates, a través del cual se determinó que el 21,4% de los sitios chilenos muestra una sola página.

Más información sobre Ricardo Baeza-Yates se puede encontrar en el sitio **http://www.dcc.uchile.cl/~rbaeza/spanish.html.**

En dicho estudio se indica que dentro de los motivos por los cuales se encuentra solamente una página en el sitio, destacan los siguientes:

- La página basa su navegación en la tecnología Javascript, por lo que es necesario interpretar dicho código para navegarla; como los robots de búsqueda no lo hacen, aparece como que no existen más y el contenido que exista no se incluye.

- La página necesita un plug-in de la tecnología Flash para visualizar su contenido; esto ocurre habitualmente en sitios que tienen una introducción animada que puede ser vista por humanos, pero que no ofrece puntos de entrada para el robot de búsqueda; por lo tanto, para éste el sitio sólo tiene una página.

- Lo anterior también es válido para aquellas páginas que emplean tecnología basada en Applets Java para la navegación, los cuales también impiden el acceso a los programas automáticos.

En los tres casos señalados se da el fenómeno que los robots de búsqueda no logran entender la sintaxis ofrecida en el código, ya que normalmente en los tres casos señalados de haber enlaces hacia el resto del sitio, estos se ofrecen desde el interior de programas que deben ser interpretados y no mediante enlaces basados en HTML. Debido a esto, en dichos casos los robots no logran encontrar la forma de tener acceso al interior del sitio web y sólo guardan la información de la portada del sitio web.

Cabe indicar que, tal como se explica más adelante en este capítulo, en los tres casos señalados existen formas de ofrecer acceso alternativo a los robots de búsqueda, facilitando el acceso de estos al sitio pese al uso de dichas tecnologías en la portada.

> Posicionamiento del Sitio Web

Una tarea permanente del administrador del sitio web será la de determinar la posición relativa del sitio web en los sistemas de búsqueda, respecto de las pala-

5.- Ver el estudio en http://www.ciw.cl/material/web_chilena_2006/index.html

bras más utilizadas por los usuarios. Para conocer cuáles son ellas, será muy importante que se haga un monitoreo permanente y constante de las visitas (como se indica en el artículo 6 del Decreto Supremo 100/2006) para ver cuáles son las que llegan desde los buscadores y las palabras que se usan para eso.

Posicionamiento web: se refiere a la ubicación relativa de un sitio web dentro de las páginas resultados de un buscador, para una o más palabras. Las técnicas para mejorar el posicionamiento se conocen como SEO - Search Engine Optimization (Optimización para Motores de Búsqueda).

Normalmente se deberá esperar que el sitio web esté indexado, lo que se puede comprobar escribiendo la dirección web principal del sitio en el buscador. El resultado deberá mostrar que el sitio efectivamente está indexado y aparece en la primera página de resultados.

Luego, deberá hacerse una búsqueda similar para las palabras que identifican al servicio u organismo al que pertenece al sitio web. Normalmente para las palabras más importantes, el sitio web debería aparecer entre los primeros lugares ya que de esa manera se podrá asegurar que los usuarios efectivamente verán el enlace y llegarán al sitio web por esa vía.

Figura 1. Las imagen muestra el resultado de la búsqueda "declaración de renta" a través del sistema Google: entre los primeros están los sitios de Gobierno que explican cómo hacerlo.

Para ello, el sitio web debe estar preparado para ser indexado por sistemas automatizados que llegarán como una visita más, y se deberá hacer el trabajo adicional de incluir el sitio en aquellos sistemas de directorio que sean más importantes en la web mundial.

Para atender adecuadamente estos dos aspectos, es importante entender la diferencia entre ambos ya que su comportamiento y forma de acceso varía notablemente, como también lo hace la forma en que un sitio web puede llegar a quedar incluido en ellos.

Page Rank: es el algoritmo diseñado por Google para indicar la relevancia de un sitio web respecto de la calidad de sus contenidos; entre muchas variables, se determina a partir de su actualización, cantidad de enlaces entrantes y salientes y tiempo de vida del sitio.

Inclusión en Índices automatizados

Los índices automatizados se forman gracias a la actividad realizada por los buscadores de Internet (search engines en inglés) que utilizan robots que navegan y almacenan información de páginas, que luego integran a una base de datos general, a partir de la cual los usuarios hacen las búsquedas. Al momento de la edición de esta Guía, los más conocidos son Google, Yahoo!, LiveWeb (ex MSN), Ask, Teoma y Quaero.

Para asegurar que un robot de búsqueda llegue a un sitio web se requiere de cumplir con al menos las siguientes características:

- Dar de alta el sitio web propio en algunos de los más importantes buscadores de Internet. Al menos se debe realizar esta acción en Google, Yahoo! y LiveWeb.
- Dar y recibir enlaces hacia otros sitios de Internet, porque ésta es la única manera que un robot de búsqueda pueda conocer la dirección de nuestro sitio web al detectar nuestra dirección a partir de otro sitio web; adicionalmente en el caso de Google, esto contribuirá a su mejorar su índice “Page Rank”.
- Ofrecer en la página de portada del sitio web los meta-tag que dirijan a los robots hacia los archivos robots.txt y sitemaps.xml cuyas características se analizan más adelante en este capítulo.
- Ofrecer en la página de portada del sitio web un meta-tag que indique el sitio da permiso para ser indexado, tal como se explica más adelante en este capítulo.

- Ofrecer en la página de portada del sitio web un enlace hacia la página Mapa del Sitio, en la que se entreguen enlaces en lenguaje HTML estándar hacia todas las secciones del sitio web.

Aunque hay más elementos que intervienen en la Encontrabilidad del sitio web como se analiza más adelante en este capítulo, al menos con los consejos anteriores se podrá asegurar que el sitio web está disponible para ser indexado por los robots de los sistemas de búsqueda automáticos.

Inclusión en Directorios

Los directorios son índices manuales de contenidos, en los que los propios interesados inscriben sus sitios, los cuales posteriormente son analizados y catalogados por los encargados del sistema, habitualmente humanos.

Gracias a esto, los sitios que se integran a un directorio aparecen ordenados por categorías y subcategorías, permitiendo a los usuarios encontrar listados ordenados de sitios, lo que facilita entender los sitios similares y competidores de los que se revisan. Un directorio puede tener un buscador interno, pero sólo para ubicar lo que se busca dentro de todos los registros existentes. Entre los más conocidos al momento de edición de esta Guía se encuentran Yahoo! Directory y Open Directory Project (también conocido como DMOZ).

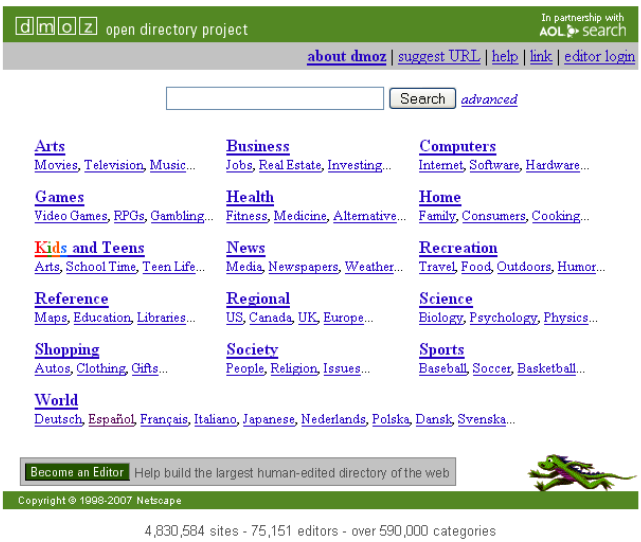

Figura 2. La imagen muestra la página de inicio de DMOZ.org, con sus contenidos separados por categorías.

Como se indicó, la única forma de integrar un directorio es por la suscripción manual del sitio, por lo que será tarea del administrador del sitio web tomar las medidas para que ello ocurra.

> Usuarios y Uso de Buscadores

En forma adicional a las tareas anteriores, el administrador del sitio web deberá tener información actualizada acerca de la forma en que sus usuarios están accediendo al sitio web desde los buscadores. Para ello será relevante revisar los informes de visita, ya que éstos cuentan con una sección en la que se analizan los referers del sitio web, que son las páginas desde las cuales llegan los visitantes gracias al uso de enlaces.

Referer: es el nombre que reciben las páginas desde las cuales un usuario accede a nuestro sitio web; su identificación se logra gracias a que quedan registradas en el log del servidor.

Gracias a esto se podrá saber cuáles son los buscadores más utilizados y cuáles son las palabras que han motivado a los usuarios a llegar al sitio web por esta vía. Este conocimiento permitirá, además, contar con una forma concreta de saber cómo se muestran los contenidos del propio sitio web en los buscadores y, a partir de ello, determinar cuáles pueden ser las reformas que se pueden aplicar para optimizar dicha experiencia.

Cabe señalar que más adelante en este capítulo se analizan algunos de los elementos principales que permiten mejorar la capacidad de los contenidos para ser encontrados, por lo que se puede seguir sus indicaciones como una guía de buenas prácticas para ayudar a la Encontrabilidad. Esto se debe a que no sólo influye el hecho de inscribir el sitio en un buscador, sino que como se verá, hay una serie de elementos que contribuyen a hacer más eficiente dicha inclusión de tal manera que cuando los usuarios busquen contenidos que están integrados al sitio web, éste aparezca siempre en la primera página. Esto último es relevante ya que abundantes estudios al respecto, indican que los usuarios siempre miran los resultados de dicha página[6], sin seguir más allá.

> ¿Cómo se mide la Encontrabilidad?

Respecto de esto último, hay que tener en cuenta que los usuarios siempre estarán intentando llegar en la menor cantidad de pasos posibles hacia los contenidos que sean de interés para resolver sus necesidades de información. Debido a esto, su

6.- Ver estudio de iProspect y Jupiter Research en http://www.iprospect.com/about/whitepaper_seuserbehavior_apr06.htm

intención siempre será que para las palabras que ingresan en los buscadores, haya algo de nuestro sitio que les permita acceder a nuestros contenidos.

La forma de asegurar esto tiene mucho que ver con los contenidos que se ofrezcan desde el sitio web, pero, principalmente, con el conocimiento de la forma en que los usuarios operan a través de estos sistemas. Esto significa un llamado a los administradores de los Sitios Web a estar permanentemente actualizados respecto de las últimas investigaciones y noticias respecto de estos temas, ya que ellas darán pistas sobre las actividades a realizar para estar más cerca de los usuarios.

En todo caso, la Encontrabilidad de un sitio web siempre estará relacionada con su habilidad para aparecer en las primeras páginas de los resultados de búsqueda de un buscador para aquellas palabras, frases y términos más relevantes relacionados con la institución, ya que será la única forma de asegurar que sea visto por quien utiliza dicho servicio. Por lo mismo, la Encontrabilidad será medida con esa característica: su habilidad para ubicarse lo más cerca posible de la parte superior de la primera página de resultados.

Relación con los motores de búsqueda

Para llegar a resultados de privilegio dentro de un sistema búsquedas, el administrador del sitio web deberá estar preocupado permanentemente de que se cumplan las buenas prácticas que se definen en las siguientes páginas, pero además deberá estar revisando frecuentemente las estadísticas de su propio sitio para averiguar los siguientes elementos:

- Determinar cuál es el buscador más utilizado en su sitio web, para aprender de sus usuarios cuáles son las palabras más usadas.

- Determinar las páginas que reciben más visitas por esta vía, de tal manera de prepararlas con más elementos que lleven a los usuarios a conocer otros contenidos relacionados a partir de ellas.

- Determinar cuál es el patrón de navegación a partir de esas páginas, para saber si ese ciclo se puede mejorar a través de contenidos más completos.

- Determinar cada cuánto tiempo están visitando el sitio web los robots de búsqueda, para tener en cuenta ese dato para efecto de las actualizaciones del sitio web.

Como se puede adivinar a partir de lo anterior, los administradores deben realizar un trabajo permanente en torno a las estadísticas e informes generados por las visitas al sitio web, ya que es la única forma de ir aprendiendo de los usuarios, los cuales siempre tendrán actividades cambiantes relacionadas con la información que existe en el ambiente y que los motiva a visitar el sitio web del servicio propietario del sitio web.

Es importante, en este sentido, que el administrador revise los contenidos de noticias generales del país referidas a los temas abordados en el sitio web, para determinar las nuevas palabras que podrían llevar a los usuarios a buscar con dichos términos. Gracias a esto, podrá modificar o mejorar sus contenidos para que los nuevos términos también permitan que más usuarios lleguen al sitio web tras una búsqueda.

Finalmente una recomendación habitual es revisar los Sitios Web que salen antes en las páginas de resultados de los buscadores para los términos en los que el sitio web tiene participación, para indagar los eventuales motivos que los llevan a tener un mejor posicionamiento que el sitio propio.

Relación con los índices

A diferencia de los motores de búsqueda, la relación con los índices es menos dinámica, ya que sólo está asociada a la acción de conseguir que el sitio web sea agregado a uno de ellos, sin que esto tenga modificaciones en el tiempo.

No obstante, hay que tener la precaución de revisar con cuidado la forma en que el sitio ha sido descrito en estos índices, ya que esto es realizado por personas que normalmente trabajan como voluntarios del sistema indexador.

¿Cómo se aumenta la Encontrabilidad?

Tras analizar las páginas anteriores queda claro que la meta de un sitio web será la de tener la mayor capacidad de ser encontrado desde los buscadores, ya que eso garantizará que los usuarios de Internet tengan acceso a la información que el sitio web puede ofrecer.

Para ello, en esta sección se aborda este desafío desde dos perspectivas: el código HTML y el contenido de las páginas, ya que desde ambos se contribuye a aumentar la capacidad de acercarse a este objetivo.

Lo primero que se analiza es el código HTML puesto que en la medida que el sitio web se desarrolle mediante el uso de código estándar habrá mayores posibilidades de que su posicionamiento sea mayor.

Robots.txt: se recomienda visitar el sitio http://www.robotstxt.org/ para obtener información acerca del uso de este protocolo.

Estándares y Códigos relacionados

Aunque la Encontrabilidad de un sitio web tiene una serie de elementos desde los cuales se puede explicar su buen resultado en los buscadores, la calidad de su código es uno de lo más relevantes.

Como se ha explicado antes, el código del sitio web debe ser estándar y por lo mismo ofrecer un cumplimiento concreto en el uso de las etiquetas HTML a lo largo de sus páginas, siendo las de la zona del `<head>` las más relevantes.

Etiquetas de `<head>`

Las páginas web bien estructuradas dividen su contenido en las zonas de `<head>` y `<body>`. La primera se ubica en la parte superior de las páginas y entrega información de referencia para el sistema computacional que utiliza y despliega la página, a fin de que pueda entender de qué manera se ha codificado el contenido y de esa manera mostrarlo adecuadamente a través del browser o programa navegador que se utilice.

Respecto de la Encontrabilidad, las etiquetas sobre las que hay que poner la mayor atención son las siguientes:

- `<title>`: permite indicar el título que aparece en el encabezado de la ventana de cada página del sitio web[7]; se recomienda que lleve el nombre del sitio web más un título que describa el contenido de la página. Por ejemplo: “Ministerio del Interior - Chile: Acerca del Ministro”. De esta manera, esta información será la que aparezca en los buscadores cuando se muestre el enlace al usuario que busca alguna palabra o frase que tenga dicha página.

- `<meta>`: una de las etiquetas “meta” de esta sección está orientada a dar una instrucción concreta a los robots de búsqueda, cual es la de indexar el contenido[8]. Para ello, su texto debe indicar lo siguiente:

Es importante considerar que los modificadores que se agregan al elemento “content” tienen efecto sobre el buscador, de la siguiente manera:

7.- Más información de este tema en la sección “Encabezado de Página” del Capítulo 2 y sección “Uso de logotipos” del Capítulo 3 de esta versión de la Guía Web.
8.- Más información en http://www.robotstxt.org/wc/meta-user.html

- **index:** indica que el contenido debe ser indexado.
- **noindex:** indica que el contenido no debe ser indexado.
- **follow:** indica que los enlaces existentes en la página deben ser seguidos.
- **nofollow:** indica que los enlaces existentes en la página no deben ser seguidos.

Uso de robots.txt

En forma paralela a lo que se indique en cada página, para el sitio se debe generar un archivo que cumple una función similar a la señalada para la etiqueta `<meta>` anterior, cual es la de indicar a los robots de los buscadores cuál es la acción global que debe desarrollar en el sitio web.

Para ello, en la raíz del servidor se debe incluir un archivo de texto que lleve el nombre robots.txt y en el que se indique la información acerca de la acción a desarrollar[9]. El contenido estándar[10] está dado por dos líneas, que son las siguientes:

```
User-agent: *
Disallow:
```

Se debe considerar que la línea “User-agent” puede incluir el nombre de cualquier robot y que si tiene un asterisco, indica que la directiva se aplica a todos; en tanto que la línea “Disallow” permite indicar los directorios del sitio web que no se desee incluir en la indexación. Si está en blanco, indica que permite indexar todo el contenido del sitio web.

Es importante considerar que este archivo es revisado por todos los robots de búsqueda que acceden al sitio web por lo que es muy importante su presencia, ya que constituye una de las buenas prácticas en torno a los buscadores, debido a que forman parte de una suerte de bienvenida formal a todos los programas enviados por los sistemas de búsqueda de Internet.

Cómo mostrar contenidos

De acuerdo a lo indicado en los párrafos precedentes, el sitio web deberá cumplir con tener los siguientes elementos para asegurar que los buscadores de Internet los indexen:

- `<meta>`: en esta sección la línea debe indicar lo siguiente:
  `<META NAME="robots" content="index,follow">`

9.- Ver Capítulo IV Guía Web, http://www.guiaweb.gob.cl/guia/capitulos/cuatro/queprobar.htm#t03robots
10.- Más información en http://www.robotstxt.org/wc/exclusion.html

- robots.txt: en este archivo el contenido debe indicar lo siguiente:

```
User-agent: *
Disallow:
```

Sitemaps.xml: se recomienda visitar el sitio **http://www.sitemaps.org/es/protocol.php para obtener** información acerca del uso de este protocolo.

Cómo esconder contenidos

Para evitar que el contenido del sitio web sea indexado, se debe tener el siguiente contenido en las páginas que no se desee incluir en los sistemas de búsquda:

- <meta>: en esta sección la línea debe indicar una de los siguientes contenidos:

```
<META NAME="robots" content="noindex,follow">
<META NAME="robots" content="noindex,nofollow">
```

- Con el primero se consigue no indexar el contenido, pero que el robot siga los enlaces ofrecidos; con el segundo se consigue que no haya indexación ni que se sigan los enlaces existentes.

- **robots.txt:** en este archivo el contenido debe indicar lo siguiente, dependiendo del caso:

```
User-agent: *
Disallow: /
```

- Con la primera línea se indica que la instrucción es para todos los robots y con la segunda, se señala que desde la raíz en adelante, no se debe indexar nada.

```
User-agent: *
Disallow: /fotos/
```

- Con la primera línea se indica que la instrucción es para todos los robots y con la segunda, se señala que el directorio llamado fotos no debe ser indexado.

Uso de sitemaps.xml

Como se revisó en las páginas anteriores, una de las dificultades más importantes referidas a la indexación en buscadores dice relación con la manera de indicar a

estos sistemas cuáles son las direcciones de las páginas web que se desea incluir en ellos.

Para enfrentar este tema, desde los sistemas de búsquedas se planteó el uso de un protocolo denominado Sitemaps que consiste en un archivo XML en el que se enumeran todas las URL de un sitio junto, a las que se agregan metadatos adicionales acerca de cada una de ellas. Por ejemplo, se indica la fecha de la última actualización, la frecuencia de modificación de sus contenidos y la importancia relativa de la página en el sitio.

- Un archivo estándar de este tipo tiene el siguiente contenido:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

Los elementos que se definen por cada línea son los siguientes[11]:

- `<urlset>`: su contenido es obligatorio y es el que permite encapsular el archivo, haciendo referencia al protocolo sitemaps vigente.

- `<url>`: también es obligatorio y es la etiqueta que permite definir cada una de las páginas web diferentes que se desea incluir en el archivo.

- `<loc>` también es obligatorio y permite indicar la dirección o URL de la página que se incluye. Debe comenzar con el protocolo correspondiente (http en el caso del web) y termina con un slash (barra diagonal).

- `<lastmod>`: es un valor opcional que permite indicar la fecha de la última modificación del archivo que se está incluyendo; para la fecha se usa el formato AAAA-MM-DD.

- `<changefreq>`: es un valor opcional que hace referencia a la frecuencia con la

11.- Se puede ver más información en http://www.sitemaps.org/es/protocol.php

que cambia la página a la que se hace referencia; sus valores son en idioma inglés y corresponden a siempre (always), cada hora (hourly), diariamente (daily), semanalmente (weekly), mensualmente (monthly), anualmente (yearly) y nunca (never). Es importante considerar que el valor "always" se utiliza para describir documentos que cambian cada vez que se accede a ellos, mientras que "never" se utiliza para describir URL archivadas.

- `<priority>`: es un valor opcional que permite informar a los motores de búsqueda las páginas que se consideran más importantes dentro del sitio web. Los valores aceptados abarcan desde 0,0 a 1,0. La prioridad predeterminada de una página es 0,5. De acuerdo a la información del protocolo, los motores de búsqueda pueden utilizar esta información para elegir entre varias URL del mismo sitio.

Es importante considerar que el protocolo Sitemaps es un estándar que ya fue aceptado por Google, Yahoo! y LiveWeb, lo que garantiza que su uso permite atender a los principales buscadores actuales de la Internet.

Una vez que el archivo ha sido creado y contiene todas las direcciones de páginas web que se desea indexar, hay que hacer referencia de él al sitio web mediante una de las siguientes actividades:

- **Mediante la interfaz de envío del motor de búsqueda:** se debe consultar la documentación ofrecida por los propios motores de búsqueda, los que indican la forma de hacerlo.

- **Mediante el archivo robots.txt:** se debe añadir una línea al final del archivo que consigne la ubicación del archivo con el sitemap. Dicha línea deben indicar:

- **Sitemap:** `<ubicación_sitemap>`

De acuerdo a lo que se indica en el sitio web en que se difunde este protocolo, la forma de ingresar la información correspondiente a la <ubicación del sitemap> debe ser la URL completa del Sitemap, como por ejemplo: http://www.sitioweb.gob.cl/sitemap.xml.

Mediante una solicitud http desde el browser: para hacerlo se debe utilizar el propio browser y en la línea de la dirección escribir lo siguiente:

- <searchengine_URL>/ping?sitemap=sitemap_url

Administración de contenidos

Tal como se indicó antes, la calidad del código HTML de un sitio web es uno de lo aspectos más relevantes para conseguir que sus contenidos sean incorporados adecuadamente en los índices y aparezcan entre los primeros lugares de las búsquedas realizadas por los usuarios.

No obstante, tal como se indica en el párrafo anterior, quienes visitan los buscadores siempre solicitarán contenidos y por tanto, será la calidad de éstos apoyada por la forma en que se ha creado el código HTML de la página, lo que determine lo forma de aparecer en las páginas de resultados.

A continuación se revisan los tres grandes elementos que se deben vigilar en cada página, para apoyar una adecuada Encontrabilidad:

- Estructura de Contenidos
- Características de los Contenidos
- Calidad de los Contenidos

Estructura de contenidos

Para que los contenidos de una página web reciban un trato adecuado en los motores de búsqueda debe existir una concordancia entre las diferentes artes del código. En este sentido, es importante velar porque el texto que aparezca en la etiqueta `<title>` de la sección `<head>` sea el mismo que aparece en la etiqueta `<h1>` de la sección `<body>`. También es adecuado que las palabras que aparezcan en la etiqueta `<meta ... content="keywords">` de la sección `<head>` incluya palabras que también aparezcan en la etiqueta `<h1>` de la sección `<body>`

Gracias a esta relación, se estará dando una prueba de que la página se refiere a los contenidos que se exponen en estas etiquetas, generando una demostración de credibilidad que es valorada dentro de los parámetros de los buscadores.

Otro elemento de interés es que el contenido esté estructurado utilizando etiquetas del tipo `<h>` para los subtítulos, ya que gracias a ello se demostrará que se ha utilizado el estándar. Adicionalmente, es interesante utilizar la etiqueta `<strong>` o `<em>` para indicar contenidos destacados, dejando de lado la etiqueta `<b>` que sólo denota negritas. Además, si dichas etiquetas se asignan a palabras que están en la lista de palabras claves (“keywords”) antes señaladas,

se aumentará la correspondencia interna de la página lo cual, a su vez, ayudará en la calidad de su indexación.

Características de los contenidos
Un tema central de la página tiene que ver con la titulación de la página, vale decir, con la frase que aparece repetida tanto en la etiqueta `<title>` como en la etiqueta `<h1>`. Se debería intentar que dicha frase incorporara la forma en que el contenido es llamado por los usuarios a través de los buscadores.

Por ejemplo, si el contenido se refiere a la "Cédula de Identificación", será interesante utilizar la denominación "Carné o Carnet de Identidad" en lugar de su nombre oficial. De esta manera, habrá más posibilidades que al ser indexada, la página tenga las palabras que sean más cercanas a lo que las personas utilizarán para hacer la búsqueda respectiva.

Otra de las prácticas habituales para apoyar la Encontrabilidad de los Sitios Web y fomentar su posicionamiento en las páginas de resultados, tiene que ver con el hecho de que se debe "convencer" a los robots de búsqueda de que la página se refiere a los temas que aparecen tanto en `<title>` como en `<h1>`. Para ello es vital que dentro del texto aparezcan varias veces los términos utilizados en dichas secciones. La lógica detrás de esta situación es que si una página se refiere a un tema determinado, es natural que en su contenido, las palabras utilizadas en los títulos (de la página y del texto), aparezcan nombradas con cierta frecuencia. Gracias a esto, se busca reprimir una mala práctica realizada por algunos sitios que para subir en su posicionamiento, ponen ciertas palabras en la lista de palabras claves ("keywords") pero luego no las usan en los contenidos.

Esto puede ser apoyado por los enlaces que ofrezca la página, que deberían ir naturalmente hacia otros sitios donde también se encuentren las mismas palabras, con lo que se reforzará el contenido de la propia página. Adicionalmente dichos enlaces deberían usar el elemento "title" en su sintaxis, de tal manera de poner allí alguna frase que refuerce la idea de que se accederá a contenidos relacionados con el tema de la propia página.

¿Cuántas veces se deben repetir los contenidos? La respuesta tiene que ver con la redacción: se debe repetir tantas veces como sea necesario para la comprensión del texto por parte de un "humano" que esté leyendo y menos de las que se puedan

interpretar como que se está haciendo dicha repetición sólo para el robot de búsqueda.

logs: son archivos de texto en los cuales se va registrando cada uno de los archivos que son mostrados por un servidor web, a raíz de las acciones que realiza un usuario que visita un sitio web mediante un browser. Su análisis permite entender lo más visitado, entre otros aspectos.

Otro elemento de interés en este sentido, es que los buscadores valoran el hecho de que haya enlaces que apunten hacia el contenido que se ofrece. En este sentido, aparece como una herramienta importante, la capacidad que tenga el sitio web de ofrecer elementos que puedan ser enlazados desde diferentes sitios. En la medida que se haga dicha acción, aumentará la posibilidad de que los contenidos del sitio aparezcan en mejores lugares en las páginas de resultados de los buscadores.

Calidad de los contenidos

Por último y aunque esto se planteó previamente, es importante reconocer que por muy importante que sean los buscadores, los contenidos que se ofrecen serán leído por personas y por lo tanto deberán ser creados para fomentar su comprensión por parte de ellas.

En este sentido, hay que convenir que en la medida que los contenidos que se ofrezcan sean de calidad y provengan de una fuente importante como es el servicio público propietario del sitio web, será bien recibido, creído y, eventualmente, enlazado desde otros Sitios Web creando de esta manera el círculo virtuoso que permite mejorar la presencia y posición en las páginas de resultados de los buscadores.

> Minería Web y Encontrabilidad

La Minería de la Web[12] es una disciplina que permite generar información acerca del comportamiento de los usuarios en un sitio web, mediante el análisis de los datos que ellos mismos van dejando a medida que visitan los Sitios Web. Utilizando técnicas provenientes de las ciencias sociales, entre las que se cuenta la clasificación, asociación y agrupación mediante patrones, es posible caracterizar a posvisitantes con el objetivo de ofrecerles productos o servicios que vayan de acuerdo a las necesidades que se asignen a los tipos de usuario que se hayan definido.

12.- Ver más información en http://www.infovis.net/printMag.php?num=172&lang=1

Para mejorar su efectividad, se define que la minería del web se puede hacer en tres áreas que se refieren al contenido del sitio y la estructura de navegación, más el comportamiento de los usuarios respecto de los dos primeros[13].

El objetivo de utilizarla en el contexto de esta versión de la Guía Web, es ofrecer una alternativa que permita la generación de mayor información acerca de las actividades que llevan a cabo los usuarios que visitan el Sitio Web.

Cabe señalar además que la minería web debe ir de la mano del monitoreo del sitio web que es apoyado desde el Decreto Supremo 100/2006 del Ministerio Secretaría General de la Presidencia (en su articulo 6)[14], en que se plantea esta tarea como una de las prioritarias para que los administradores de los Sitios Web puedan conocer las necesidades de los usuarios y la forma en que están utilizando el sitio web.

Respecto de esto es importante tener en cuenta que las acciones que realizan los usuarios son registradas anónimamente en archivos de texto también conocidos como logs (o bitácoras, en español), en los cuales se va registrando cada uno de los archivos que son mostrados por un servidor tras la petición de un cliente que visita un sitio web. Por lo mismo, cada visita genera decenas o centenares de líneas de información que al ser procesadas con software especializado de análisis[15], permiten tener información agregada acerca de visitas, zonas del sitio que son más visitadas y otros elementos básicos de información similares. Asimismo es posible tener información acerca de las palabras ingresadas en los buscadores externos e internos del sitio web, lo que ayuda a entender cuáles son los términos más buscados y para los cuales el sitio web constituye una fuente de información.

En este sentido, la posibilidad de analizar por ejemplo, los logs de la navegación del sitio web o bien los logs de las palabras ingresadas en un buscador, permitirán al administrador del sitio web, tener información de primer orden para tomar decisiones acerca de contenidos, de la forma que tiene el sitio web e incluso, para tomar decisiones de reorganización de las secciones existentes en el mismo.

Quién busca y qué busca

Respecto de los usuarios del sitio web se debe intentar conseguir información acerca de qué están buscando en el sitio web y cuáles son los objetivos que persiguen al visitarlo.

13.- Más información en http://www.webtaller.com/maletin/articulos/web-mining-diseno-sitios-web.ph
14.- Se puede obtener copia del artículo en www.modernizacion.cl y también en el Capítulo 1 de esta Guía.
15.- Ver software de este tipo en http://www.download.com/Site-Management/3150-2181_4-0.html?tag=catat

Aunque la información que se obtiene mediante minería de web será anónima, ya que está basada en elementos de este tipo, será posible activar otros tipos de recursos para conocer al usuario y de esta manera saber más acerca de sus necesidades y las razones que lo llevan a visitar al sitio web.

Por lo mismo, se apoya como una buena práctica que el administrador del sitio web pueda dedicar algún tiempo de su jornada semanal a las siguientes actividades:

- Responder correos electrónicos de los usuarios que tengan relación con la operación del sitio web, ya que en el intercambio con ellos será posible entender su percepción del sitio web y las necesidades de información que lo llevan a visitarlo. Desde allí, será más fácil conocer si hay satisfacción de las necesidades y qué pasos se deben dar para conseguirla.

- Responder llamados telefónicos de usuarios que no consigan terminar sus operaciones y para quienes los sistemas de ayuda tradicionales ofrecidos por el servicio u organización dueña del sitio web tampoco aporten información adecuada. El contacto directo con los usuarios será una herramienta valiosa para perfeccionar los contenidos y funcionalidades del sitio web.

- Desarrollar tests de usuario tendientes a entender de qué manera se relacionan los usuarios con los contenidos y a partir de esto, establecer las mejoras necesarias en los ámbitos que se requieran (este tema será tratado con más profundidad en el Capítulo 5 de esta Guía).

Los seis tipos de contenidos según R. Baeza

Cuando se esté trabajando con usuarios, será importante avanzar en el conocimiento de los seis tipos de contenidos que el usuario viene a buscar en el sitio web, los cuales quedan normalmente reflejados a través de las palabras que usa en el buscador interno del sitio para encontrar aquellos términos que el sistema de navegación no le puede aportar o no le muestra dónde pueden estar ubicado dentro de la organización actual de contenidos.

Basado en un diagrama de árbol[16] que se aprecia en la figura anterior, el académico Ricardo Baeza-Yates plantea que los seis tipos de contenidos que los usuarios buscan a través del buscador interno del sitio son:

16.- Este diagrama fue publicado por el autor en el artículo "Excavando la Web" que apareció en "El Profesional de la Información" (http://www.dcc.uchile.cl/~rbaeza/inf/EPIexcavando.pdf)

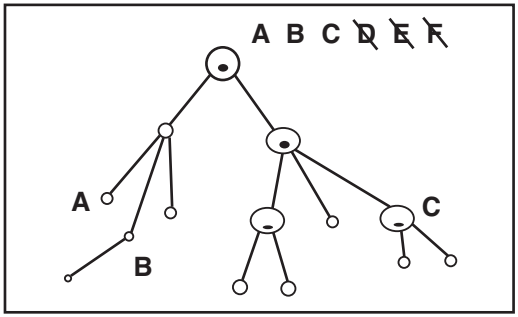

Figura 3. La imagen muestra el árbol de organización de contenidos donde aparecen los seis tipos definidos por el académico.

- A) Contenidos principales y que están en el sitio web, pero que no son destacados adecuadamente por el sistema de navegación o de contenidos.
- B) Contenidos que constituyen un tema secundario del sitio web y que no son destacados como elemento de importancia.
- C) Contenidos que no son destacados en el sitio web y que los usuarios asumen que son parte de los contenidos, por lo que los buscan en el sitio web.
- D) Contenidos que existen en el sitio web pero que están registrados con otro nombre.
- E) Contenidos que no existen en el sitio web pero que deberían estar, ya que forman parte de los contenidos que deberían utilizarse.
- F) Contenidos que no existen en el sitio web y para cuya inexistencia se cuenta con definiciones editoriales o políticas de la organización o servicio.

El académico recalca que los últimos tres son muy importantes porque revelan que los usuarios pueden dar pistas de mucho interés para la creación de contenidos, para los cuales el sitio web es considerado una fuente principal o relevante.

Influencia de la Minería en los contenidos

Basado en la información anterior, es evidente que el desarrollo de una política de minería de web sobre los contenidos permite tener un aporte contundente para la generación de contenidos, ya que se trata de un mecanismo de feedback efectivo para entender las necesidades de información de los usuarios.

Por lo anterior, deberá constituir una buena práctica la revisión permanente de los informes de actividad del sitio web más la información que aporten los informes de minería web, ya que basados en ellos se podrán tomar decisiones editoriales que permitan responder a las necesidades que los usuarios manifiesten a través de su navegación por el sitio web.