Internet Invisible: un recurso terciario en la Red

Por **M. Fernanda Peset**, **Ricardo Albiñana** y **Silvia Morales**

Introducción

1. El tamaño de la información en la Red

S. Lawrence y **L. Giles**, en un estudio realizado en 1998[1], consideraron que el porcentaje de la web visible indizada por los motores de búsqueda era del 60%. Un año más tarde, estos mismos autores estimaron que en su conjunto no eran capaces de indizar más del 42%. La razón de esta pérdida de cobertura se debe al tiempo que invierten en realizar su trabajo pues no son capaces de actualizarse al mismo ritmo que crece la Red y pueden tardar meses en indizar una nueva página o una que haya sido modificada[2].

Por otra parte, en un estudio de *Datasearch* de 1997, **S. Feldman**[3] afirma que, aproximadamente, el 50% de la web no era indizable: se trataba de información almacenada en bases de datos. El tamaño estimado por **S. Lawrence** y **L. Giles** de la web pública e indizable en 1999 era de unos 800 millones de páginas. Pero ¿qué tamaño tiene aquella que no está contenida en estas paginas visibles?

2. El control de la información invisible en internet

Desde el punto de vista de los buscadores, internet se divide en webs visibles, constituidas por páginas que pueden ser indizadas en su totalidad, y otras invisibles, las bases de datos, cuyo contenido es infranqueable para estos sistemas de consulta.

Los motores de búsqueda disponen de programas, conocidos comúnmente como arañas (*spiders*) que exploran la Red, la rastrean continuamente en busca de nuevas páginas, desmenuzando sus contenidos y depositándolos en las grandes bases de datos de dichos buscadores. Pero cuando llegan a un sitio que tiene la información almacenada en bases de datos lo más que pueden hacer es *darse una vuelta por el jardín y decirnos dónde está la casa* porque sus contenidos no pueden ser indizados. Para estas arañas existe una información que es invisible, la que contienen las bases de datos residentes en internet.

> **«El tamaño estimado de la web pública e indizable en 1999 era de unos 800 millones de páginas. Pero ¿qué tamaño tiene la que no está contenida en estas paginas visibles?»**

Todo ello evidencia que la localización de información no sólo es cada día más complicada, sino que los buscadores no son el único instrumento válido o útil para encontrarla. Puesto que estos motores no pueden informar de todos los recursos accesibles en la Red, recientemente han surgido alternativas para ayudar a descubrir lo que se está buscando. Una de ellas es la recopilación de las bases de datos accesibles en internet[4].

Internet Invisible

1. Origen y desarrollo

En este contexto se origina *Internet Invisible*, una recopilación de bases de datos de acceso gratuito, que nace de la decisión de poner a disposición de los profesionales de la información y público en general una fuente de información gratuita accesible en la Red.

http://www.internetinvisible.com

Por otra parte, contribuye a paliar un fenómeno extendido en el ámbito de la información: el acceso desigual a la misma. El mundo desarrollado invierte cada vez más recursos para hacerla accesible, de manera que produce un desequilibrio en su difusión. Este hecho destaca notablemente si se compara la situación de Estados Unidos con el resto de países ya que la mayor parte de información estructurada, es decir controlada, tiene su origen en este país. Lo mismo ocurre con los buscadores, que indizan especialmente páginas ubicadas en lugares estadounidenses relegando el resto de sitios web del mundo.

2. Ámbito temático

Internet Invisible es un directorio de bases de datos principalmente de contexto español, entendiendo como tal tanto lo producido en el estado español, en cualquiera de sus lenguas oficiales, como lo procedente de otros lugares geográficos de temática o habla española o autonómica. No obstante, también incluye recursos ajenos a este criterio cuando constituyen un punto de referencia en su campo de aplicación.

3. ¿Cómo se encuentra estructurada?

Cada recurso recopilado ofrece una ficha técnica (figura 1), sin descuidar que la ayuda propia de la base de datos quede a la vista. En ella se recoge:

—enlace, procurando que apunte directamente a la pantalla de búsqueda,

—entidad o persona responsable de su creación, y

—breve descripción de su contenido, incluyendo naturaleza de los datos, opciones de búsqueda, idioma, etc.

4. Organización de la información

Se ha elegido un modelo habitual en muchos buscadores, al que

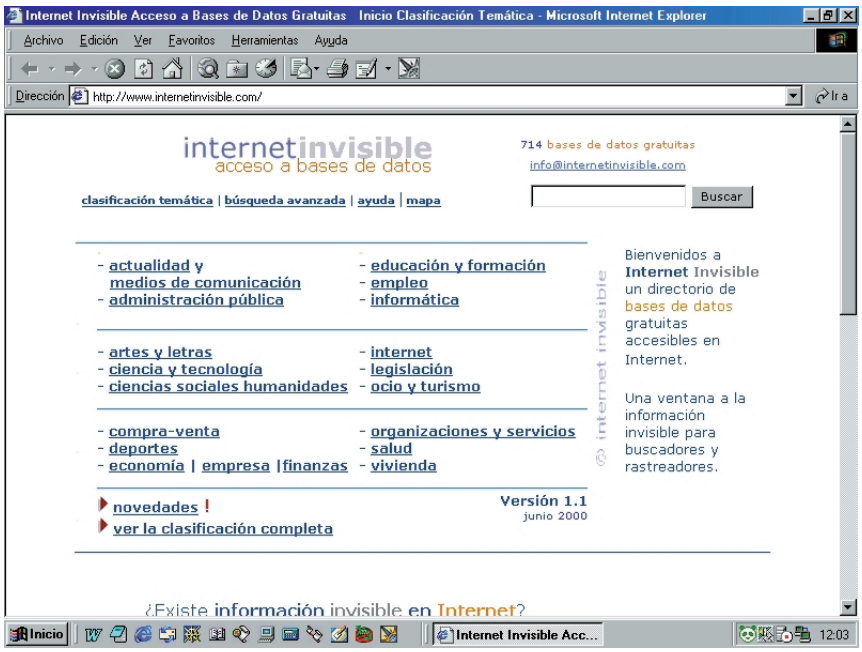

Figura 1

los usuarios ya se han adaptado. Permite dos modos de acceso: un directorio temático, un formulario simple o caja de búsqueda presente en todo el web y una opción de búsqueda avanzada.

Como se aprecia en la imagen (figura 2), las bases de datos están organizadas en grupos temáticos, los cuales se subdividen a su vez en materias más específicas ampliables a medida que son descubiertas nuevas fuentes en la Red. Las subdivisiones de esta clasificación son**:**

—actualidad y medios de comunicación: agendas culturales, cine, moda, premios y certámenes, prensa, radio y televisión,

—artes y letras: archivos, arte y diseño, bibliotecas, editoriales, lengua y literatura, librerías, materiales de consulta (diccionarios, enciclopedias), música, revistas, teatro,

—ciencia y tecnología: ciencias naturales, ciencias exactas, ingeniería y tecnología,

—ciencias sociales: antropología, arqueología, filosofía, geografía, historia, política, sicología, sociología,

—empleo: público, privado, e

—informática: hardware, software, etc.

5. Opciones de búsqueda

Además del directorio temático, existe la posibilidad de localizar los recursos recopilados mediante búsqueda por palabras. Aparecen dos modalidades, la simple (presente en todas las pantallas) y la avanzada, siendo esta última la que permite la utilización de operadores booleanos: *and*, *or* y *not*.

6. Sugerencias

— Hay que introducir las palabras acentuadas, pues el sistema es sensible a los acentos.

— Se puede escribir en mayúsculas o minúsculas y omitir los artículos, preposiciones y conjunciones (palabras vacías).

— No es necesario utilizar el operador de truncamiento para obtener todos los términos con la misma raíz.

— Si no se obtiene una respuesta satisfactoria, es necesario revisar la escritura de las palabras introducidas o probar con términos relacionados o sinónimos.

No hay que olvidar que *Internet Invisible* es una puerta de acceso, cada base de datos a la que se accede mantiene un lenguaje de consulta distinto, adaptado al tipo de información que contiene. Por lo tanto, para aprovechar al máximo sus posibilidades es conveniente familiarizarse con cada una de ellas.

7. Recursos técnicos de almacenamiento y gestión de la información

El servidor web está ubicado en Florida (Estados Unidos) y trabaja mediante la configuración mostrada en la tabla 1. En su actual versión (1.0) los recursos están almacenados en una base de datos relacional (*Microsoft SQL Server 7.0*, sistema de gestión de bases de datos relacionales para *Windows*). Su funcionamiento se asienta sobre un sistema que compara las palabras introducidas por el usuario con las existentes en la base de datos buscando las ocurrencias de esas palabras en los registros de la misma.

«Para los buscadores existe una información que es invisible, la que contienen las bases de datos residentes en internet»

Los resultados de la búsqueda se traducen en páginas *ASP* (*active server pages*) que muestran en un

Componentes	Versión
Sistema operativo	
Microsoft Windows NT Server	4.0 SP5
Software de la base de datos	
Microsoft SQL Server *Active Server Pages (ASP)*	7.0 SP1
Software de correo electrónico	
Microsoft Exchange Server	5.5 SP2
Software de estadísticas	
MediaHouse Statistics Server	5.01
Software del servidor web	
Microsoft Internet Information Server	5.0

Tabla 1

listado la descripción de los recursos junto con su enlace externo. *ASP* es parte del *Internet Information Server* (*IIS*) de *Microsoft*, una tecnología de páginas activas que permite el uso de diferentes scripts y componentes en conjunto con html para mostrar páginas generadas dinámicamente. *Microsoft* introdujo esta tecnología en 1996 y la define como "un ambiente de aplicación abierto y gratuito en el que se puede combinar código html, scripts y componentes *ActiveX* del servidor para crear soluciones dinámicas para el web".

La normalización de la base de datos (diseño y análisis del modelo de datos) está codificada en un esquema relacional en *tercera forma normal*, utilizándose el lenguaje de scripts de *Microsoft* para enlazar las páginas *ASP* con la base de datos.

8. Referencias

Internet Invisible está referenciado y es utilizado por los siguientes servicios de información:

— **Codina, Lluís**. "Internet Invisible: acceso a bases de datos". En: *Biblioteca digital web de la semana. Curso de postgrado en documentación digital* [en línea]. Barcelona: *Universitat Pompeu Fabra*, 2000.

http://docdigital.upf.es/digital/

— **Fornas Carrasco, Ricardo**. "Internet Invisible". En: *Internet todas las claves para navegar*. Madrid: El país Aguilar, 2000.

— **Fornas Carrasco, Ricardo**. *Buscopio, buscador de buscadores* [en línea]. Madrid: Proel. Consultado en: 20-05-2000.

http://www.buscopio.com/

— **Peset Mancebo, Fernanda.** "Internet Invisible: un recurso terciario en la Red" [en línea]. En: *Iwetel*, 2000, 19 de mayo. Consultado en: 30-05-2000.

http://listserv.rediris.es/cgi-bin/wa?A1=ind0005d&L=iwetel

— *Argos*: servidor de información. Enlaces de interés [en línea]: València: *Generalitat Valenciana, Subsecretaría de planificación y estudios*. Consultado en: 20-05-2000.

http://inf1.pre.gva.es/argos/index.html

— *Bases de dades a internet* [en línea]. Barcelona: *Universitat*, Biblioteca, Secció de referència electrònica. Consultado en: 30-05-2000.

http://www.bib.ub.es/www5/5bd8.htm

— *Bases de dades d'accés públic* [en línea]. València: *Universitat,* Servei d'informació bibliogràfica. Consultado en: 30-05-2000.

http://www.uv.es/~infobib/enlaces/enlaces_bases_c.html

— *Bases de datos* [en línea]. Madrid: *Universidad Complutense, Facultad de Ciencias Políticas y Sociología*, Biblioteca. Consultado en: 30-05-2000.

http://www.ucm.es/BUCM/cps/home.htm

— *Bases de datos en línea* [en línea]. *Universidad de Cantabria,* Biblioteca. Consultado en: 30-05-2000.

http://www.buc.unican.es/Recursos/basesdatos.htm

— *Bases de datos* [en línea]. León: *Universidad,* Biblioteca. Consultado en: 20-05-2000.

http://www3.unileon.es/ser/bu/bd-cdrom.html

— *Bases de dades i bibliografies* [en línea]. Barcelona: *Universitat,* Biblioteca. Biblioteconomia i Documentació.

http://www.bib.ub.es/www5/5bd74.htm

— *Bases de dades i bibliografies* [en línea]. Barcelona: *Universitat.*

http://www.bib.ub.es/www5/5bd74.htm

— *Herramientas de búsqueda en internet* [en línea]. [Zaragoza]: *Gobierno de Aragón, Departamen-*

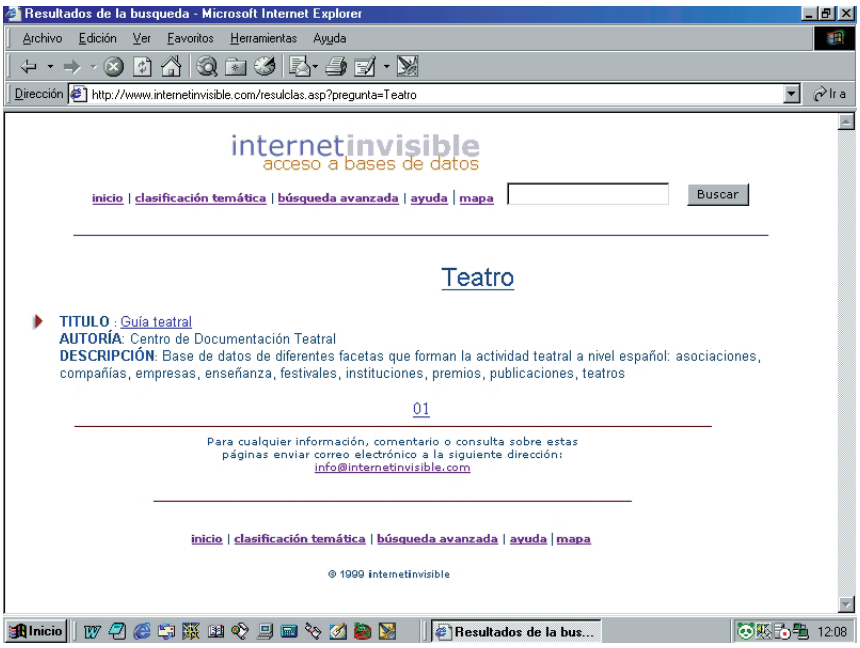

Figura 2

> **«*Internet Invisible* nace de la decisión de poner a disposición de los profesionales de la información una fuente de información gratuita accesible en la Red»**

to de educación y ciencia. Consultado en: 30-05-2000.

http://www.aragob.es/educa/

— *ICTnet. Semanario digital de la comunidad de los profesionales*. Barcelona: *Institut Català de Tecnologia*, *Newsletter,* n. 153 (24-05-2000). Consultado en: 2-05-2000.

http://www.ictnet.es/esp/servicios/noticias/novedades/hemeroteca/

— *Recursos de información* [en línea]. Salamanca: *Universidad Pontificia*, Biblioteca. Consultado en: 30-05-2000.

http://www.upsa.es/%7E/servicios/biblioteca/recursos.html

— *Recursos en internet* [en línea]. Madrid: *Universidad Complutense*, *Facultad de Farmacia*, Biblioteca. Consultado en: 30-05-2000.

http://www.ucm.es/BUCM/far/busca.htm

— *Sistemas de búsqueda* [en línea]. [Madrid]: *Universidad Nacional de Educación a Distancia*, Biblioteca. Consultado en: 30-05-2000.

http://info.uned.es/biblioteca/sistemasdebusqueda.htm

9. Recursos humanos

La selección, recopilación y descripción de las bases de datos se lleva a cabo por profesionales de la información, por lo que el factor humano asegura una selección de alta calidad. Emulando a las arañas de los motores de búsqueda, personal especializado se sumerge en la Red buscando los recursos que, por la calidad o por el interés de sus contenidos, son susceptibles de ser introducidos en *Internet Invisible*.

Notas

1. **Lawrence**, **S**.; **Giles**, **L**. "Searching the world wide web". En: *Science*, 1998, v. 280, n. 5360, pp. 98-100. Consultado en: 30-05-2000.
http://www.neci.nj.nec.com/homepages/lawrence/papers.html

2. **Lawrence**, **S**.; **Giles**, **L**. "Accessibility of information on the web". En: *Nature*, 1999, v. 400, pp. 107-109. Consultado en: 30–05-2000.
http://www.neci.nj.nec.com/homepages/lawrence/papers.html

3. **Feldman**, **S**. *New study of www search engine coverage published*. Consultado en: 30-05-2000.
http://www.infotoday.com/newsbreaks/nb0712-1.htm

4. Recopilaciones de bases de datos:

— AlphaSearch.
http://www.calvin.edu/library/searreso/internet/as/

— BigHub.
http://www.thebighub.com

— Direct Search.
http://gwis2.circ.gwu.edu/~gprice/direct.htm

— Invisible Web.
http://www.invisibleweb.com

— Lycos Invisible Web Catalog.
http://dir.lycos.com/Reference/Searchable_Databases/

— Search.com.
http://www.search.com

— WebData.
http://www.webdata.com

— Search Engines databases and Newswires.
http://www.internets.com

***Mª Fernanda Peset**. Universidad Politécnica de Valencia, Departamento Comunicación Audiovisual, Documentación e Historia del Arte.*
mpesetm@upvnet.upv.es

***Ricardo Albiñana**. Fundación Pascual Tomás, responsable del área de cultura.*
ricardo@internetinvisible.com

***Silvia Morales**. Generalitat Valenciana, documentalista.*
silvia@internetinvisible.com