IX CONGRESS CONGRESO
ISKO-SPAIN ISKO-ESPAÑA
Valencia 11th, 12th, 13th March 2009 11, 12 y 13 de Marzo de 2009
New Perspectives for the organisation and dissemination of knowledge
Nuevas perspectivas para la difusión y organización del conocimiento

UN NUEVO DIRECTORIO TEMÁTICO PARA LA ACCESSIBILIDAD AL CONOCIMIENTO DE LA WEB INVISIBLE

Sra Eniko Pajor

SZTE Juhász Gyula Főiskolai Kar Felnőttképzési Intézet Könyvtártudományi és Múzeumpedagógiai Szakcsoport (Universidad de Szeged, Instituto Andragogía, Dép. de la Biblioteconomía y Museopedagogía – Hungría)
Correo electrónico: pajor@jgypk.u-szeged.hu , pajorm1em@t-online.hu

RESUMEN

El tema de mi comunicación es la presentación de uno de mis trabajos cuyo finalidad es ayudar, facilitar el acceso al conocimiento de la red invisible. Éste es un directorio temático que ya he construido en húngaro, pero estoy reorganizándolo para (y en el futuro con) mis alumnos de ERASMUS españoles. En mi presentación los temas se siguen según sus relaciones, asociaciones ligadas con Internet invisible. Algunas, a modo de ejemplo: explicación del concepto/noción (hidden, dark, invisible etc.), primeros sondeos, resúmenes de la web invisible, obras, fuentes fundamentales, fuentes rusas, fuentes alemanas, fuentes francesas, fuentes españolas, técnicas de recuperación de información, ensayos de soluciones para hacer visible el contenido invisible, buscadores especializados y desarrollados exactamente para Internet invisible, agentes inteligentes, portales, plataformas, tutoriales de web invisible, tecnología topic map, buscadores que utilizan la tecnología topic map, buscadores visuales, percepción visual, blogs, videos de Internet invisible, etc.

ABSTRACT

The subject of my communication is the presentation of one of my works whose purpose is to help, to facilitate the access to the knowledge of the invisible network. This one is a thematic directory that already I have constructed in Hungarian, but I am reorganizing it for (and in the future with) my Spanish students of ERASMUS. In my presentation the subjects are followed according to their relations, associations ligatures with invisible Internet. Some, as a example: explanation of the concept/notion (they hidden, dark, hair net etc.), first soundings, fundamental, Russian sources, German sources, French sources, Spanish, technical sources of test, information retrieval of solutions to make the invisible content visible, seeking summaries of the invisible Web, works, sources specialized and developed exactly for invisible Internet, intelligent agents, vestibules, platforms, visual tutorials of invisible Web, technología topic map,

finders that use the technology topic map, finders, visual perception, blogs, videos of invisible Internet, etc.

PALABRAS CALVES

Organización del conocimiento, Internet invisible, Web invisible, red profunda, aplicaciones de topic map, buscadores visuales, percepción visual.

KEY WORDS

Knowledge organization, application of topic map, invisible web, hidden web, visual search engines, visual perception.

INTRODUCCIÓN

„Estamos perdidos en las informaciones,
a pesar de tenemos sed de la sabiduría."

Para seguir los cambios generados por la sociedad de información-orientada de hoy, no es suficiente utilizar el Internet basado en experiencia incluso sin embargo que el uso es extremadamente intensivo. Buscando las informaciones el sentido, el comportamiento consciente son dos factores esenciales. Pues la vida cotidiana ha cambiado considerablemente, una nueva clase de conocimiento y las nuevas habilidades son necesarias, que se basan sobre todo en la tecnología de comunicación de la información (ICT.) Aunque sea inverosímil decir que el Internet puede ofrecer una respuesta o una solución a todos y a cualquier problema, „la Red de las Redes" se ha convertido el lugar más importante para la adquisición de la información profesional. A pesar de esto, está consiguiendo que más difícil encontrar la información necesaria. Este fenómeno se puede detectar en la educación superior también. Debido a la cantidad de información de forma aplastante - varios millones por más que necesitado -, ofrecida en el Internet, los estudiantes se deben conocer, aprender la estructura y las técnicas del Internet. Estes conocimientos les permitirían a encontrar la información de calidad.

En la docencia dedicada al estudio de las fuentes de la web invisible y a la recuperación de información, me enfrento habitualmente a un problema repetido: en los portales y directorios temáticos que incluyen un gran número de bases de datos, informaciones de contenidos especializados, etc., el usuario se encuentra abrumado y sobrecargado a causa de la abundancia de información. Sin embargo, en el acceso temático, acceso formal o por materias, la búsqueda es compleja y el usuario se encuentra, igualmente, perdido. Las direcciones URL no son significativas, está desorientado y desconoce cuáles son las fuentes verdaderamente relevantes para su trabajo entre millares de las localizadas en la web. Atendiendo a este punto de partida he construido un directorio temático de una parte de la Red, llamada „Red invisible/Red profunda" -, primero en húngaro, pero he reorganizándolo en español también para mis alumnos de ERASMUS españoles.

EL DIRECTORIO

„La red invisible es el término utilizado para describir la información, generalmente almacenada y accesible mediante bases de datos, que no es recuperada interrogando a los buscadores convencionales."
(http://www.internetinvisible.com)

Este directorio tiene trece objetivos: 1) reunir las mejores fuentes de la red invisible de cualquier lugar en siete idiomas (en inglés, en ruso, en húngaro, en español, en francés, en italiano, en alemán), 2) hacer perceptible la multiplicidad de interpretación de las soluciones para acceder a éstas informaciones y 3) dar a conocer a los usuarios esta parte de la red que es cinco cientos veces mayor (Bergman, 2001) de lo que los motores

de investigaciones son capaces utilizar actualmente. El directorio cambia con regularidad para dar las novedades de este tema y reúne solamente las mejores fuentes de la red profunda/invisible de alta cualidad. La cantidad de los vínculos no sobrepasa nunca un límite determinado (aproximadamente de 200 vínculos), puesto que cada fuente misma es un tesoro de millares y millares de informaciones también, por eso querría eliminar el sobrecargo intelectual.

La página húngara que contiene el directorio en español: http://melyweb.lap.hu

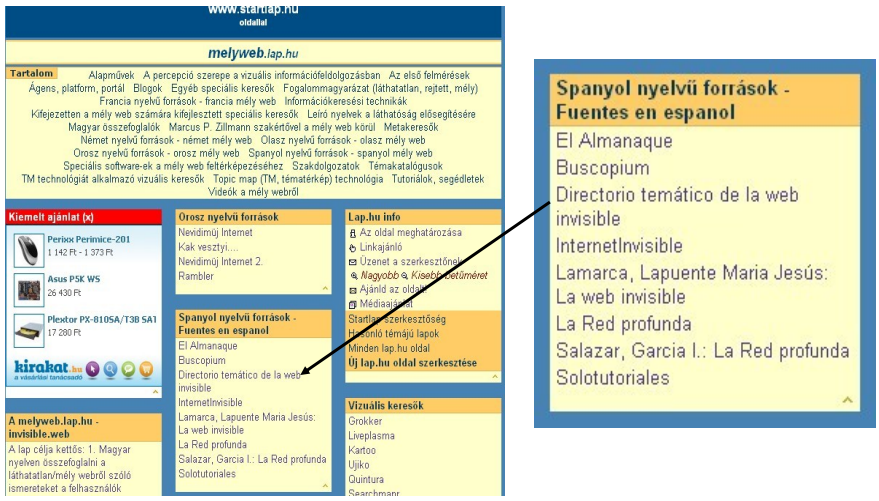

Figura 1. – El directorio temático sobre la página húngara encajado en las fuentes españolas

EL CONTENIDO TEMÁTICO Y LA ESTRUCTURA DEL DIRECTORIO

Para que los estudiantes entiendan lo más fácilmente posible las relaciones y la parte fundamental del tema, las palabras clave que caracterizan la red invisible se siguen enciclopédico y no en orden alfabético. Para facilitar su utilización y la selección de una fuente o herramienta electrónica, cuando se sitúa el cursor sobre la URL, se activa un mensaje en español que resume el contenido o las características importantes de la fuente de información. De este modo el usuario puede decidir inmediatamente, si abrirlo o no, si este sitio de web es interesante o no para él.
Pues, veamos unos subtemas paso a paso!

Los primeros sondeos y conceptos

„… It would be a site that's possibly reasonably designed,
but they didn't bother to register it with any of the search engines.
So, no one can find them! You're hidden.
I call that the invisible Web."
(Jill H. Ellsworth, 1996.)

En la literatura especializada anglófona uno más grande de los acontecimientos de estos último años es la aparición de una monografía y de los resultados de un sondeo. Michael K Bergman - fundador de la sociedad Bright Planet - hacía el sondeo (Bergman, 2001), y los dos autores del libro eran Chris Sherman y Gary Price (Sherman-Price,). El tema que trastornó la vida de los expertos informática, de los informáticos y de todo el mundo que se ocupa de la búsqueda de la información de calidad era el fenómeno llamado „la red invisible", había mencionado el primera vez en 1996 por Jill Ellsworth.

A nuestros días es ya muy conocido que al principio la red era enriquecida por documentos textos que tenían la estructura hypertext. Para facilitar su investigación los motores de investigación se desarrollaban de la esta óptica que fueran capaces de recuperar los documentos en formato HTML. Esta etapa no duró mucho tiempo, porque el internet se invadía de los documentos en distinto formato: fotografías, vídeos, partes de música, otros objetos audios, ficheros post escritura? (post script), animaciones flash etc., en particular, todas las cosas que podía nacer con el desarrollo de la informática e Internet. A partir de final de los años 1990 los usuarios se dieron cuenta que los motores de investigaciones dan mucho o demasiado o los pocos resultados, pero los resultados de calidad y de no calidades son mezclados. Otro tipo era cuando se conocía la existencia de los documentos pero los buscadores no los encontraron o los usuarios no podían tener acceso a las bases de datos. Al resumen: los buscadores titulados al HTML no podían encontrar uno grandes partes de información que tenían otro formato que HTML (java, perl, php, flash, formatos compresivos, páginas dinámicos etc.). La parte que hoy la literatura especializada ya se llama y tiene en cuenta como red invisible.

En la estructura de red, la visibilidad, la accesibilidad y la indexabilidad son los tres pilares importantes. Si entre estos tres criterios hay exactamente uno que no se realiza, la página se sitúa en la parte invisible de la red.

Para la visibilidad la estructura de red tiene una gran importancia. Los dos investigadores (San José, San Mateo, IBM, Compaque y Alta Vista) en 2000 dibujaron el primer „mapa integral" de red donde se puede bien proseguir los vínculos y los fronteros que separaban los distintos, diferentes lugares en la Red. Estas eran las fronteras que pueden dificultar la navegación entre sitios o hacerla imposible. Sobre en mapa se ve la totalidad de las páginas que se conectan y son visibles, que se conectan y

no son ya visibles y que no eran nunca visibles. Hasta a los años de 2000 esta estructura fue llamada como el modelo „Bow Tie Theory" (BTT).

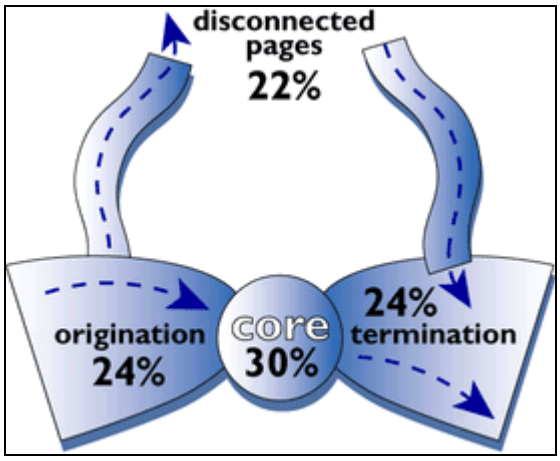

Figura 2. – La teoría de pajaríta elaborada hace 8 años (The „Bow Tie Theory")
Fuente: http://www.almaden.ibm.com/software/images/bowtie.gif> [consulta: 30 de noviembre de 2008]

La teoría de pajarita estableció que la Red dejaba de ser un fenómeno de desarrollo espontáneo aparentemente sin reglas que explicasen su crecimiento y difusión, para convertirse en una materia mesurable y previsible. Con ésta teoría podían desarrollar un modelo como medio para explicar, a la vez que ilustrar, el comportamiento dinámico de la Web, ayudando a comprender la Red como una organización. (<http://www.rizomatica.net/la-teoria-de-la-pajarita/>). Desde este tiempo el BTT ya no existe. La Red en totalidad actualmente no tenga ese carácter espontáneo y este formato que se le atribuye. El desarrollo de red se pasa tanto rápidamente y ella es tanta inmensa que la red se volvió autoorganizada con una estructura graf y unas efectos centrípetos de algunos sitios web. (Carr, 2008). Este cambio influye sobre el indexabilité lo que es la condición previa del accesibilidad. Y los dos juntos son la condición previa de la visibilidad también. Hay muchas razones de la visibilidad o no visibilidad.

Las fuentes que se ve en esta partida del directorio están ligadas a este problema y dan acceso al libros, tesis, fuentes electrónicas que trataban el primero el fenómeno „web invisible/web profunda".

Sistemas de tratamiento y recuperación de información

„El saber total de la Antigüedad –al menos en su forma final, custodiado en la Biblioteca de Alejandría – se ha calculado en 0,8 terabytes …. Los veinte millones de libros de la Biblioteca de Congreso de Estados Unidos ocuparían (sin contar sus ilustraciones) 20 terabytes. La totalidad de un corte actual de la Web visible

daría 7,5 terabytes de texto, y la Web invisible (para Bright Planet), 7. 5000 terabytes de información."
(José Antonio Millán)

El tamaño de la proporción de los artículos en la literatura especializada prueba la importancia del proceso de recuperación de información. (González, 2000; Pálvölgyi, 2001; Gonzálo, 2004; Danneberg, 2005; Lamarca Lapuente, 2006; LIK, 2008.) Sin duda, los métodos más rápidos y eficaces de recuperación de documentos/informaciones de la Red son los que proporciona la propia Web a través de los buscadores. Tiene que acentuar, que la recuperación de información no es igual a la recuperación de datos, por que ella consiste el tratamiento y procesamiento de documentos, no sólamente datos y registros. La recuperación de información de web invisible o profunda es más difícil porque:

- la mayor parte de información se encuentra en bases de datos que tienen páginas dinámicas y por eso no están indexadas. Para que una página sea indexada, esta debe ser estática no dinámica y contener enlaces hacia otra páginas. Sin indexación la página no se vea? visto?, y las „arañas" son impotentes frente a esta.
- Otro problema es que no todo es texto en Internet, en Web. Las herramientas interactivas, bibliotecas digitales, los OPACS etc. tienen páginas dinámicas también.
- Un otro motivo es la cantidad gigantesca de la web invisible. Los tamaños de la web superficie y de la web profunda no pueden ser indicados claramente. No tenemos que números aproximados. Dependiendo de las técnicas aplicadas, puede ser dicho que el invisible/profunda es cerca de 500 veces más grande que la web superficial, o que la web tiene 7,5 terabytes enfrente de la web invisible que contiene 7. 500 [!] de información!

Una otra opinión: Maria Jesús Lamarca Lapuente en su tesis cita las palabras de Ricardo Baeza Yates:"la web tiene actualmente al menos unas cuatro mil millones de páginas estáticas y un número cientos de veces mayor de dinámicas … Además, tenemos que agregar toda la web invisible, en intranets o páginas con acceso restringido. La web oculta es seguramente miles de veces más grande que la pública". (Lamarca Lapuente, 2006)
Buscadores especiales son necesarios para la recuperación de información de las páginas dinámicas. Los motores de investigaciones especiales no pueden trabajar correctamente solamente. Para tener un buen resultado, es necesario que las palabras clave del documento y las palabras clave de la investigación, de la cuestión sean las mismas o sus corte. Para resolver el problema de las ambigüedades en la recuperación, tiene que utilizar lenguajes documentales de representación de conocimiento (lenguajes controlados, tesauros, actualmente metadatos y ontologías.) Estos requisitos desempeñarán un papel aún más importante en el trabajo de la generación siguiente de web la que está conocida como „web semántica.". Sin embargo todo esto no resolvía el problema de la abrumadora cantidad de documentos e informaciones de referencias de

Red, pero con sus utilizaciones el usuario puede obtener respuestas más eficaces y pertinentes acordes con sus necesidades.

Por eso en esta partida del directorio se hallan las portadas para la aprendizaje de las etapas de recuperación de la información, las herramientas, buscadores y bases de datos especiales desarrolladas exactamente para encontrar la información de la Red profunda. Aquí se presentan unos de nuevos experimentos, de proyectos y de progresos técnicos. Entre éstos, los más prominentes son los que han sido creados por los investigadores que habían descubierto la „web invisible/web profunda". Esto es seguido por una selección metabuscadores que se han desarrollado para encontrar diversos tipos de documentos o de formatos de archivo.

Herramientas especializadas utilizando las nuevas tecnologías

"Making your page visible on the Web!"
(Jill H. Ellsworth)

En los últimos anos, el desarrollo de técnicas de Internet y, en concreto, las aplicaciones de XML han dado como resultado una aplicación espectacular: el sistema llamado „topic map." Éste visualiza las informaciones y su implementación aparece en varios sitios, desde los tesauros hasta los buscadores (Pajor, 2006). Como es conocido la visualización de información ayudar a las usuarios mejorar su experiencia de navegación entre los resultados del contenido del sitio o de la página. Con la ayuda de la percepción visual y la hipótesis de interpretación el visitante debería poder hacerse una idea instantánea de "de qué va el sitio", de qué temas trata predominantemente. (Dürsteler, 2008.) La busca, la comprensión del contenido, las relaciones como la elección de las fuentes se pasa en uno minuto, por que el imagen lo que el cerebro memorizaba una vez se puede revelar/reestrenar? en cualquier tiempo con sus informaciones.

Esta parte del directorio abraca los mejores buscadores que visualizan las informaciones (Grokker, Kartoo, Ujiko, Quintura etc.), da tutoriales para utilizar, entender la web semántica, las nubes de etiquetas, la psicología de la visualización y aplicar las tecnologías como XML, SGML etc. Cada recurso al que se accede mantiene un lenguaje de consulta distinto, adaptado al tipo del información que contiene o administra, para aprovechar al máximo de sus posibilidades.

CONCLUSIÓN

Este directorio en español está creado exactamente para ser el primera puerta de acceso como introducción por mis estudiantes ERASMUS espanoles, para los navegantes quién no conocen bastante bien la metodología de la recuperación de información. Las fuentes elegidas, controladas y anotadas abarcan de todos aspetos y relaciones del tema y de este modo expanden el horizonte de la red para el usuario pasivo (debutante) como para el usuario activo o dinámico. Los estudiantes para la obtención del conocimiento

científico se deben utilizar métodos y técnicas rigurosos que permitan tener la confiabilidad y validez. Utilisando esta puerta de acceso al informaciones de web invisible los estudiantes aprenden más fácil y serán capaz elaborar sus trabajos, también les brindará las bases para empezar a elaborar el protocolo de su tesis BA o MA.

BIBLIOGRAFÍA CITADA

BERGMAN, M. K. *The Deep Web: Surfacing Hidden Value*. [en línea] [consulta: 30 de noviembre de 2008] Disponible en Web: <http://www.brightplanet.com>

The *Bow Tie Theory (BTT)* [en línea] [consulta: 18 de noviembre de 2008] Disponible en Web: <http://www.almaden.ibm.com/almaden/webmap_press.html>

CARR, N. *The centripetal web*. En Nicholas Carr's blog. 19. de octobre de 2008. [en línea] [consulta: 30 de noviembre de 2008] Disponible en Web: <http://www.roughtype.com/archives/2008/10/the_centripetal.php>

DANNENBERG, D.; HERZIG, B.; RENGER, H.; *Leitfaden zur Entwicklung von Unterrichtseinheiten zur Förderung von Informationskompetenz*. En: Bertelsmann Stiftung (Hrsg.); Ministerium für Städtebau und Wohnen, Kultur und Sport des Landes NRW (Hrsg.) : Kooperation macht stärker : Medienpartner Bibliothek & Schule. CD-ROM. Gütersloh: Bertelsmann Stiftung, 2005.

DÜRSTELER, J. C. *Revista Inf@Vis : Vizualización de Información*. No. 196. [en línea] [consulta: 30 de noviembre de 2008] Disponible en Web: <http://www.infovis.net/>

ELLSWORTH, J. H.; ELLSWORTH, M. V. *Marketing on the Internet.* 2nd ed. New York; Chichester; Brisbane; Toronto; Singapore; Weinheim : Wiley Computer Publishing; John Wiley & Sons, Inc., 1997. 428 p

GONZÁLEZ, A. *Resumen de Nuevas Técnicas de búsqueda en Internet*. Madrid, Facultad de Informática, Universidad Complutense, 2000. [en línea] [consulta: 30 de noviembre de 2008] Disponible en Web: <http://www.fdi.ucm.es>

GONZÁLO, C. *La selectión de palabras clave para el positionamiento en buscadores : conceptos y herramientas de estudio*. En Anuario Hipertext.net mayo de 2004. [en línea] [consulta: 30 de noviembre de 2008] Disponible en Web: <http://www.hipertext.net>

LAPUENTE, M: J: L. *Hipertexto : El nuevo concepto de documento en la cultura de la imagen.* Thesis doctoral Madrid : Universidad Complutense, 2006. [en línea] [consulta: 30 de noviembre de 2008] Disponible en Web: <http://www.hipertexto.info>

LERNSYSTEM *Infromations Kompetenz* (LIK) [en línea] [consulta: 15 de noviembre de 2008] Disponible en Web: <http://www.lik-online.de/html/>

MILLÁN, J. A. *El libro de medio billión páginas : La ecología lingüistíca de la web.* [en línea] [consulta: 16 de noviembre de 2008] Disponible en Web: <http://www.jamillan.com/ecoling.htm>

PAJOR, E. *Una aplicación de topic map que puede ser un modelo posible* En: La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico : Interdisciplinarity and transdisciplinarity in the organization of science knowledge. VIII. Congreso ISKO – Espana. León: Universidad de León, 2007. 245–252. [en línea] [consulta: 15 de noviembre de 2008] Disponible en Web: <http://isko2007.unileon.es/presentaciones/pajor.ppt>

PÁLVÖLGYI, M. *Keresőnyelvek és fogalomtárak általános, ismeretreprezentációs és technológiai tendenciái* 2001. [en línea] [consulta: 29 de noviembre de 2008] Disponible en Web: <http://www.medinfo.hu/forum/palvolgyi-trend.htm>

SHERMAN,C.; PRICE, G.; The web invisible : Uncovering Information Sources Searche Engine Can't See. Medford; New Yersey : Information Today, 2001. 439 p. (CyberAge Books)

La *teoría de la pajaríta*. [consulta: 17 de noviembre de 2008] Disponible en Web: <http://www.rizomatica.net/la-teoria-de-la-pajarita/>