

# Recursos informativos en las redes

## Nuevos productos y servicios

### Ya exportamos ingeniería lingüística a los EUA

*Bitext* (Madrid, España), empresa especializada en tecnología lingüística, y la compañía encargada de desarrollar la tecnología de búsqueda e indexación *dtSearch* (Bethesda, MD, EUA) han firmado un acuerdo de colaboración. *dtSearch* aporta su herramienta *Text retrieval engine*, y *Bitext* la inteligencia lingüística mediante *DataLexica*. *Text retrieval engine* es un motor de búsqueda e indexación, mientras que *DataLexica*, por su parte, dota de inteligencia al proceso de búsqueda e indexación, y lo convierte en algo sencillo e intuitivo, proporcionando una cobertura máxima y eliminando el ruido.

Este acuerdo abre la posibilidad de incluir inteligencia lingüística a la recuperación de información, mejorar el rendimiento de los buscadores (tanto en precisión como en exhaustividad) y potenciar su facilidad de uso. Estos productos están especialmente pensados para procesos de gestión de contenidos y conocimiento, internet, intranet, extranet y/o cd-rom.

<http://www.bitext.com/>

<http://www.dtsearch.com/>

### Tras la “usability”, la “findability”

Ante la segunda edición del famoso libro de **Louis Rosenfeld** y **Peter Morville**, *Information architecture for the world wide web*, merece la pena destacar su última invención, la idea de “findability”, casi intraducible, excepto como “¿encontrabilidad”? Este nuevo concepto a tener en cuenta, algo que llevamos reclamando desde hace largo tiempo los profesionales de la información venidos de la documentación (como él mismo), es el arte de conseguir que un usuario encuentre lo que necesita en una web —valdría lo mismo decir, claro, en cualquier aplicación informática, o en un sistema de información como una intranet/extranet, sistema de gestión del conocimiento (que es mucho más que una intranet), etc.—.



## Centro de Información y Documentación Científica

[www.cindoc.csic.es](http://www.cindoc.csic.es)

- ✓ Apoyo documental a la investigación científica.
- ✓ Recopilación y difusión de la actividad científica española.
- ✓ Desarrollo de proyectos de investigación en Documentación Científica.
- ✓ Formación de profesionales y usuarios.

Más información: [sdi@cindoc.csic.es](mailto:sdi@cindoc.csic.es) - Tel.: 915 635 482 ext.: 215



Algunas de sus ideas al respecto muestran la aplicación directa en la usabilidad y la arquitectura de la información, pilares de la moderna gestión de contenidos desde el punto de vista de la experiencia del usuario: “mi última aproximación al tema es la ‘findability’. En sí, precede a la usabilidad. No se puede usar lo que no se puede encontrar. Es extremadamente difícil medir el coste de encontrar información; pero es aún más difícil medir el coste de no localizarla. Muchos vocabularios, hablados o escritos, están fuera de control. Por ejemplo, en una web lo que uno llama ‘home’ otro puede llamarlo ‘página principal’. Ello ilustra por qué es útil estandarizar y controlar el vocabulario en determinadas situaciones”.

<http://louisrosenfeld.com/>

## Noticias de buscadores

### Premios 2002 a los mejores motores de búsqueda

Las prestigiosas webs *Pandia* y *Search Engine Watch*, especializadas en motores de búsqueda, han dado a conocer a los mejores buscadores de 2002. Reseñamos aquí las coincidencias y categorías premiadas más interesantes:

—Mejor motor de búsqueda: *Google*.

—Mejores metabuscadores: *ez2www*, *Vivisimo* y *Copernic*.

—Mejor sitio web sobre buscadores: *Search Engine Watch*.

—Mejor buscador de noticias: *Google News*.

—Mejor buscador de imágenes: *Google Images*.

—Mejores buscadores internos de e-commerce: *Yahoo Shopping* y *DealTime*.

—Mejor diseño: *Google*.

—Mejor característica de un buscador: el corrector lingüístico de *Google* (“Spell Checking”) y las capacidades de “customización” de la apariencia de la interfaz de *AllTheWeb*.

<http://www.pandia.com/post/016.html>

<http://searchenginewatch.com/awards/2002-winners.html>

### IBM promete superar a Google

IBM anuncia que en breve dará a conocer una nueva tecnología llamada *Unstructured information management architecture* (*Uima*). Asegura aunar diferentes aproximaciones de la inteligencia artificial en torno a la recuperación de información y la soñada web semántica. Se utilizará desde el cálculo estadístico de la relevancia de las respuestas de *Go-*

*gle*, xml y rdf, a técnicas de procesamiento del lenguaje natural.

<http://news.com.com/2100-1001-981256.html>

### Todos los filtros y operadores de Google

Además de los clásicos operadores booleanos, podemos utilizar una lista de comandos y filtros que nos permitirán realizar una consulta mucho más elaborada y específica. ➡

### Se lanza Documentum 5 al mercado

Se ha anunciado *Documentum 5*, última versión de la conocida plataforma de gestión de contenidos. Según su web oficial mejora las prestaciones en aquellos puntos de los que casi todos sus clientes se quejan: mayor facilidad de uso y de implementación, con lo que el tiempo de integración en los sistemas de información es menor; mayor seguridad por perfiles, e integración de gestión documental, escaneado, archivado y gestión de contenidos para el web en un solo espacio de trabajo.

[http://www.documentum.com/products/launch/documentum\\_5.htm](http://www.documentum.com/products/launch/documentum_5.htm)

### El álgebra booleana de Google

Los buscadores de internet han superado ampliamente las clásicas posibilidades de filtrado de preguntas basadas en el álgebra booleana. Por ejemplo, *Google* descarta por defecto el operador booleano “or”, asumiendo que su base de datos es tan grande que, o intenta ser muy específico, o el resultado obtenido en una consulta de este tipo tendrá demasiado ruido. Aun así, nos permite utilizarlo a voluntad en su búsqueda avanzada.

Siguiendo esta lógica, busca por defecto con el operador “and”, y es más, aplica por defecto un operador “near” creciente. Si tecleamos en la cajetilla de búsqueda dos palabras, intenta encontrar primero las adyacentes y en el orden en que se han escrito. Después aumenta de “near” 1 a “near” 2, 3, etc, independientemente del orden en que fueron escritas en la búsqueda (“query”). El orden de relevancia o ranking se obtiene a partir de ahí combinando su pertinencia a estas estrategias de búsquedas con otros parámetros, tanto externos como internos.

—Cache: la “foto” de la última vez que *googlebot* (su robot) pasó por la web y la indexó.

—Link: para buscar webs que pongan un enlace a cierta página (“sitas” o citas de sitios web).

—Related: páginas relacionadas o parecidas con la de inicio de aquella que usemos como referencia.

—Info: muestra la información que *Google* tiene de nuestra página de inicio o principal.

—Stocks: da información de bolsa sobre una empresa. Necesitamos conocer su símbolo correspondiente y para ello podemos consultar:

*http://finance.yahoo.com/l*

—Site: combinado con una palabra clave (o varias), restringe su búsqueda al sitio web indicado.

—Allintitle: limita la búsqueda únicamente al título de la página web (aquello que está entre las etiquetas `<title>` y `</title>`).

—Intitle: única forma de operador “or” que hay en *Google*. Si usamos dos palabras en nuestra ecuación de búsqueda permite buscar lo primero en el título y lo segundo en el resto del sitio web.

—Allinurl: igual que allintitle, pero buscando en la url. Un ejemplo, en:  
*http://www.webdeejemplo.com*

—Inurl: lo mismo que intitle, pero con la url.

Varios de estos comandos se pueden combinar entre sí en una misma ecuación de búsqueda. Se aprecia mejor con un ejemplo: para encontrar todo lo que tenga la palabra “documentación” en el título de la web y además sea en sitios web sólo de España, o mejor dicho, sitios web con dominio *.es* teclearemos la siguiente ecuación:  
*site:es intitle:documentacion*

No podemos combinar todas las opciones posibles, como el comando link: que parece que no permite mezclarse con ningún otro, pero sí con otros.

Otro ejemplo sería saber cuántas páginas indexa *Google* de un sitio web concreto:

*inurl:webdeejemplo  
site:webdeejemplo.com (sin http)*

Otros filtros que podemos utilizar:

—Fecha: publicado hace 3 meses, 1 año, etc.

—Por formato: pdf, Word, Excel, Power Point, etc.

—Tamaños de imagen.

—Tipo de imagen: jpeg, gif, etc.

—Color de imagen: en blanco y negro, escala de grises, colores, etc.

—Grupo de discusión (sólo en foros).

—Autor (sólo en foros).

—Identificador de mensaje (sólo en foros).

—Orden por relevancia/por fecha (sólo en noticias).

*http://www.google.com/help/operators.html*

**Jorge Serrano Cobos. Serikat Consultoría e Informática, S. A.**

*http://www.serikat.es*

*http://trucosdegoogle.blogspot.com*