Capítulo X

Internet, fuente documental para el periodismo

David García Martul

Introducción

Nuestro objetivo del capítulo es mostrar los fundamentos de Internet que pueden resultar útiles en la labor del periodista. Expondremos las soluciones que Internet aporta en la resolución de los problemas con los que se encuentra el periodista en el curso de rutinas de elaboración del mensaje periodístico. Es nuestro propósito, demostrar que la red va mucho más allá del empleo del motor de búsqueda Google, si bien explicaremos posibilidades de la herramienta que pueden resultar útiles en la redacción del mensaje periodístico y que sin embargo, son poco conocidas. Asimismo, mostraremos las diferencias y momentos más adecuados para utilizar las búsquedas en motores y en directorios. Qué nos pueden aportar los lenguajes controlados y los metadatos. Qué es un portal horizontal y uno vertical, así como aquellos más apropiados para la labor del periodista tanto en su faceta profesional como académica.

1. La Documentación Informativa o Periodística

Según Galdón (2002), la documentación informativa es un saber práctico que tiene por objeto la valoración, selección, clasificación y archivo de los textos y referencias que, una vez recuperados, sirven para elaborar una información periodística verdadera, inteligible y adecuada y difundir información de base periodística. En efecto, es ante todo un conjunto de técnicas orientadas a que el periodista sea capaz de buscar documentos, con independencia de su formato, procedencia o finalidad. Pero, ¿qué tipo de documentos? Pues aquellos que sean relevantes para la elaboración de una información periodística contrastada. En este sentido, la documentación no es un saber científico, ni un conjunto de textos narrativos acerca de la experiencia social y humana, sino un saber práctico con el cual se satisface la inquietud del periodista por localizar información contrastable y veraz, al mismo tiempo que aprende a resolver el problema que Wurrnan (2001) denominó la angustia informativa.

Más sintética y concreta es la definición que Codina y Fuentes (2000) nos proporcionan acerca de lo que se concibe por Documentación informativa o periodística. Nos dicen que es el conjunto de ciencias y técnicas documentales al servicio de: a) la producción de informaciones de actualidad, b) el incremento de su calidad, c) su almacenamiento y conservación y d) su difusión y reutilización. La ventaja de esta definición es que entra de lleno en el papel de la documentación en el contexto de la convergencia tecnológica en las empresas informativas y más concretamente en la unificación tanto de las redacciones como de todos los servicios, incluido el servicio de documentación, en una única redacción digital. García Avilés (2009) nos señala la preocupante carencia de los periodistas en cuanto a formación adecuada para la recuperación de datos. Esta sería una de las causas de que la calidad de sus producciones se encuentre en descenso. ¿Por qué?

El sistema gestor de contenidos cada vez se concibe más como una factoría unificada cuyas noticias se distribuyen a través de los diversos soportes de la empresa informativa. Esto supone que se necesite un único centro documental que permita y facilite la realización de economías de escala entre distintas plataformas. De hecho, en los medios, con independencia de su canal de difusión preferente, ofrecen incluso a sus clientes la posibilidad de consultar la información retrospectiva almacenada en sus bases de datos documentales, bien sean textuales como de imagen fija o en movimiento.

El papel de los servicios de documentación ha ganado tanto peso con Internet que Codina (2000) ya nos avisaba de la aparición de una nueva rutina en la labor del periodista: la investigación en línea. Esta tarea, o más bien habilidad del saber profesional del periodista,

estaría constituida por cuatro componentes: 1) La búsqueda y obtención de información tanto en Internet como en repositorios y bases de datos especializadas. 2) La evaluación, selección y descripción de recursos informativos por medio del empleo de metadatos o descripciones de las propiedades de los recursos seleccionados. 3) El conocimiento de técnicas de posicionamiento web y marketing; es decir, que el periodista sea capaz de proporcionar al producto de su trabajo la mayor difusión posible en el ámbito adecuado al contenido de su noticia. 4) La especialización en contenidos que por sus características intrínsecas obligarán al conocimiento de fuentes de difusión propias como pueden ser bases de datos de información científica, el conocimiento de revistas de difusión científica. Esto se debe a que el abanico de recursos se ha ampliado tanto que se ha generado una necesidad por profundizar en el conocimiento que el periodista tiene de sus fuentes.

En este capítulo, no podemos tratar con todos las posibles fuentes con las que cuenta el periodista, pero vamos a tratar de proporcionar de forma sintética algunas habilidades relacionadas con los cuatro componentes que constituyen el campo de la investigación en línea. Si bien, nos centraremos más en técnicas y recursos de recuperación de información en Internet que puedan resultarle útiles para la edición de noticias con independencia de su formato y medio de comunicación.

2. La información en Internet

Actualmente, se cuenta con un elevado número de recursos de información en Internet, en general de calidad informativa variable. La primera pregunta que se hace el periodista es acerca de cómo puede buscar información. La segunda pregunta se refiere a la posibilidad de localizar información pertinente, no sólo a sus demandas informativas, sino sobre todo a las posibilidades de redacción de un mensaje suficientemente preciso para la comunicación del mensaje a sus clientes.

Han sido muchos los trabajos acerca del potencial del hipertexto como modo alternativo de transmisión de mensajes y sobre todo de lectura, entre los cuales vamos a destacar la obra de Landow (1997) sobre teoría del hipertexto y el trabajo de Díaz Noci (2001) acerca del papel del hipertexto en la construcción del discurso informativo en periodismo. Nosotros, vamos a quedarnos con la concepción de que el hipertexto es una tecnología que organiza la información en bloques distintos de contenidos, conectados a través de una serie de enlaces cuya activación o selección, provoca la recuperación de información realizando saltos entre un documento y otro. Esto supuso una revolución en la transmisión del mensaje al introducir un tipo de lectura asociativa y romper con el marco estructural tradicional de lectura lineal de los mensajes.

Este avance permitió la introducción de un conjunto de tecnologías en la web que facilitaran el tratamiento hipertextual de la información. Así, se introdujeron los lenguajes de marcas como SGML, luego simplificado en HTML con sus sucesivas versiones hasta llegar a XHTML. El paradigma explicativo del desarrollo de estos lenguajes fue la tendencia hacia la separación entre el contenido y la estructura.

En un periódico en formato papel conocemos la importancia que tiene la maquetación de los mensajes informativos. La inserción de los mismos en una rejilla proporciona un producto con identidad propia. Sin embargo, ¿qué ocurre si deja de ser importante la posición del texto en el medio?

Cuando se comenzó a editar páginas HTML los autores no se preocupaban más que de escribir contenidos para la web sin reparar en cuestiones de posicionamiento web, de organización de la información o de usabilidad y accesibilidad. Sin embargo, a medida que el número de páginas web se iba incrementando se fue haciendo más difícil su recuperación, a pesar de la existencia de numerosas obras con consejos de diseño web. De esta manera comenzaron a aparecer los primeros trabajos de arquitectura de información. Entre ellos, destacamos la obra de Morville y Rosenfeld (2002). Estos autores, documentalistas de formación, fueron los primeros en definir el campo de la arquitectura de información, así como los ámbitos a los que afectaba. Entre ellos se encuentra el campo de los lenguajes controlados y los metadatos. Ellos aportan varias posibles definiciones del campo tales como:

1. Combinación de organización, etiquetado, y esquemas de navegación dentro de un sistema de información.

2. El diseño estructural de un espacio de información que facilita la tarea de dotar al usuario para acceder de forma intuitiva al contenido.

3. El arte y la ciencia de estructurar y clasificar sitios web para ayudar a los usuarios a localizar y gestionar los recursos de información.

3. ¿Directorio o motor de búsqueda?

Un directorio o buscador basado en índices como los denomina Bradley (2004) se basa en la clasificación del conocimiento bajo una lista de encabezamientos y subencabezamientos por medio de una estructura en árbol, de lo más genérico a lo más específico. La ventaja que ofrecen reside en que los encabezamientos permiten al usuario guiarse durante el proceso de búsqueda hasta la localización por si mismo de los recursos de información que resuelven sus demandas informativas. Otra ventaja, es que no exigen al usuario un conocimiento profundo de la materia que buscan ni conocer técnicas de búsqueda o de selección de la información. El usuario es quien decide el nivel de profundidad de la búsqueda. Entre los directorios más conocidos contamos con el directorio de Google y el de Yahoo. Sin embargo, quizás para el ámbito del periodismo el más interesante sea BUBL. Se trata de un directorio organizado con el sistema de clasificación Dewey. Esto permite al periodista la localización de un repositorio de recursos, seleccionados, comentados y almacenados por especialistas, y al que se le ha asignado una materia del sistema Dewey para que pueda ser localizable de forma lógica por el periodista, tal y como vemos en la figura siguiente.

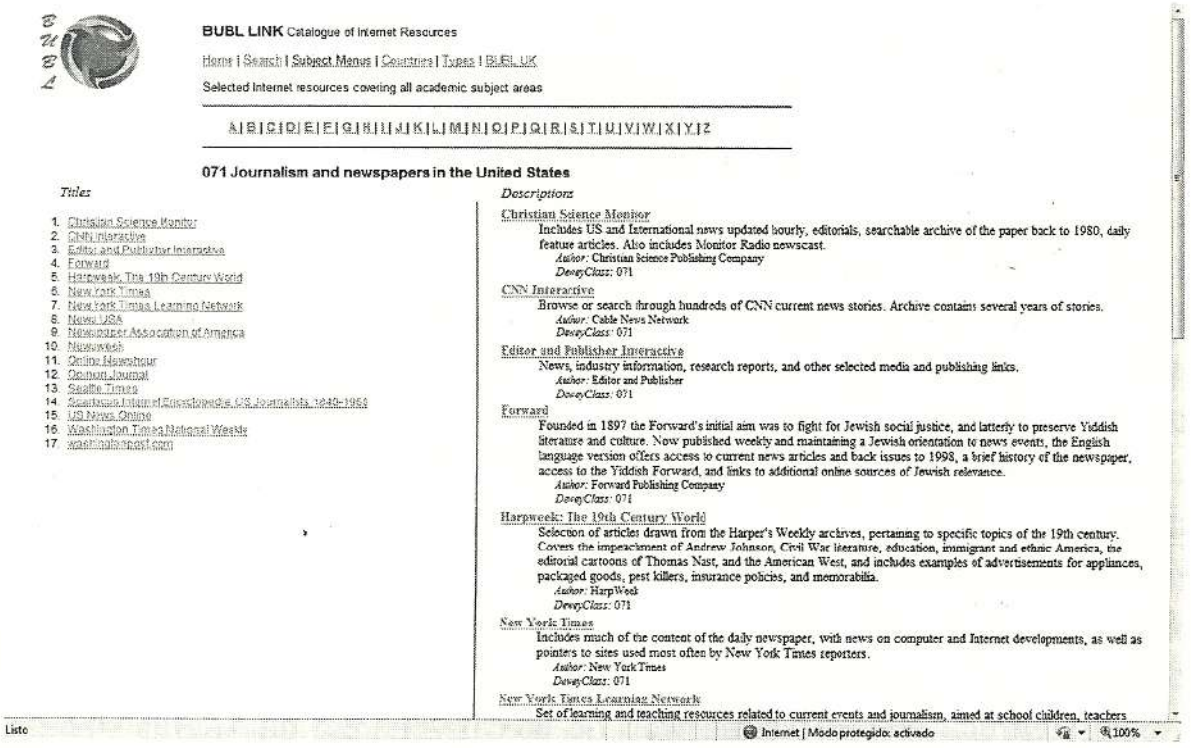

Figura 1. Directorio BUBL. En: www.bubl.ac.uk. Consultado el 10/02/2010.

Como vemos, el directorio nos permite mostrar una visión global de todos los sitios pertenecientes a una categoría particular. Por ejemplo, si quisiéramos conocer qué diarios norteamericanos cuentan con sitios web resultaría difícil hacerlo con un motor de búsqueda, o al menos nos demoraríamos bastante. En cambio, navegando por los subencabezamientos de un directorio como BUBL o Yahoo resulta bastante intuitivo acceder a una lista con una relación de periódicos con sus propios sitios recopilada por algún documentalista.

Entonces, ¿para qué nos pueden servir los motores de búsqueda? Los directorios son muy útiles cuando se busca información de calidad y se sabe lo que se está buscando, pero tienen el peligro de encontrar silencios o muy escasos resultados, especialmente cuando las búsquedas piden información muy específica. En cambio, los motores de búsqueda nos van a proporcionar mucho ruido, especialmente si no sabemos muy bien lo que buscamos, pero por lo menos nos ofrece pistas por medio de los documentos relacionados, o con la opción de la búsqueda sencilla "Voy a tener suerte", por la cual es el propio motor quien selecciona el resultado que más se aproxima a nuestra demanda y nos lo abre en el navegador ¿Cómo lo hace? Para ello, vamos a explicar cómo funciona un motor de búsqueda convencional como Google.

Cada motor de búsqueda tiene su propio algoritmo que indica cómo un programa, denominado crawler, debe recoger la información contenida en las páginas web. En el caso de Google, este algoritmo se denomina "*PageRank*". Fundamentalmente, consiste en que asigna de manera automática una nota de 0 a 10 cada página web indizada. Así, si una página obtiene un *PageRank* de 0 a 3, su nivel no es muy alto en Google. Por el contrario, si el *PageRank* obtenido es de 5 y 6, la página es mejor considerada y por lo general suele ocupar los primeros puestos de la lista de resultados recuperados. Es decir, el *PageRank* es lo que determina el éxito en la recuperación de una página en Internet.

El *PageRank* de la mayoría de páginas se puede ver en la barra de herramientas de Google. Si queremos usar seriamente Google se recomienda su utilización. Esta barra permite buscar directamente en Google sin necesidad de acceder a los sitios web, pero sobre todo integra la opción *PageRank*. Este indicador opera para calcular la visibilidad de las páginas web y no de los sitios web. Esto significa que la página web de la Universidad de Sevilla tiene una nota de 8, resultado muy elevado, lo que indica que la página es importante y está bien posicionada en los resultados de búsqueda, en tanto que la página web del departamento de periodismo 11 tiene un resultado de 3, lo cual refleja cierta dificultad para la recuperación de la página en este motor de búsqueda. Sin embargo, es posible verificar periódicamente el éxito en los esfuerzos por mejorar la visibilidad de la página. Siempre debe tenerse en cuenta que los cambios en los resultados tardan al menos un mes, tiempo medio de espera para la actualización de las bases de datos de Google.

Supongamos que deseamos trabajar como periodistas autónomos en un blog creado por nosotros mismos. La pregunta es, ¿cómo hacer que nuestro blog sea fácilmente recuperable en la red? La primera respuesta es obvia, pagando a Google por medio del servicio Adwords. Esto nos permite tener el blog bien posicionado para Google pero no para otros motores de búsqueda, además hace que nuestro medio pase a estar con otros medios publicitados. Y, finalmente, si somos periodistas profesionales nos interesa que nuestro medio tenga una visibilidad suficientemente importante como para poder constituir una plataforma publicitaria lo bastante rentable para poder autosostenerse financieramente. De hecho, existen numerosas experiencias de este tipo, pero las que sobreviven son aquellas sostenidas por empresas multinacionales. Entre las plataformas de comunicación más conocidas están Google News o Terra del grupo Telefónica, las cuales se pueden permitir tener pérdidas económicas por estar integradas como servicio de valor añadido en las plataformas de su casa matriz.

La idea esencial de la visibilidad con Google, y en general con cualquier motor de búsqueda, es que la cantidad de enlaces que apunten a la página web de nuestro blog será muy importante para el *PageRank*. Este principio por el cual la cantidad de enlaces que apuntan a una página es esencial para la determinación del rango de una página a partir de sus resultados de búsqueda. Esto permite la búsqueda eficaz de páginas con contenidos de calidad en Internet.

Ya sabemos cuál es el criterio por el cual se presentan los resultados. Pero, ¿de dónde se obtienen? En primer lugar debemos tener en cuenta que la información en la web está escrita en un lenguaje de marcas denominado HTML, siglas de Hipertext Markup Language ¿Qué significa esta expresión? Pues que con un editor de texto plano, sin ningún tipo de formato se han escrito los contenidos de la página web con etiquetas tal que <HEAD> </HEAD> para que pueda ser interpretable por un programa denominado navegador. Ya hemos visto cómo en la red existen unos programas, crawlers, que están constantemente recogiendo los datos contenidos entre las etiquetas HTML de las páginas web. Estos datos se pueden consultar desde un navegador solicitando su código fuente. Concretamente, los crawlers registran la información contenida entre las etiquetas <HEAD> o cabecera como el título, el autor, etc. Así que si una página web tiene poca información en su cabecera, entonces poco podrá el crawler registrar de esa página para llevarla a la base de datos de Google. Sin embargo, no por incluir mucha en las cabeceras va a hacer más visible la página. La clave está en la inclusión de etiquetas meta o metadatos. El metadato se define como la descripción que se hace de las características de un recurso digital, tales como el autor, título, materias, lengua, etc. Pero, al igual que la catalogación de la colección de una biblioteca, se hace con una norma de catalogación. En metadatos existen muchas normas pero la más conocida es Dublin Core, creada por un consorcio de bibliotecas norteamericanas para tratar de ordenar la información en Internet. Un ejemplo de ello lo mostramos a continuación para el código fuente de la página web de la Universidad Carlos III:

```
<head>
<title>Universidad Carlos III de Madrid</title>
<meta HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">
<meta name="DC.title" content="Universidad Carlos III de Madrid - Universidad
Carlos III de Madrid">
<meta name="DC.description" content="Web Oficial de la Universidad Carlos III
de Madrid">
<meta name="DC.date" content="2007-10-08">
<meta name="DC.format" content="text/html">
<meta name="DC.language" content="es">
<meta name="DC.publisher" content="Universidad Carlos III de Madrid">
```

Como podemos ver incluye un título como cualquier página web, pero incluye unas etiquetas meta que los crawlers registran de manera prioritaria. Como vemos las etiquetas incluyen la expresión “DC” que se refiere al esquema “Dublin Core” y a continuación el tipo de propiedad descrita como la fecha de edición, formato, lengua o editor de la página.

Una vez el crawler ha registrado la información de millones de páginas web, la envía a la base de datos hospedada en los servidores de Google. Si hacemos la búsqueda de un medio de comunicación con Google como “DiariodeSevilla.es” en los resultados nos muestra el tiempo que ha transcurrido desde la última actualización, y en el enlace de “caché” se nos muestra la página web guardada por última vez en la base de datos de Google. Esto significa que la página puede haber cambiado y Google conserva durante un periodo de tiempo la página guardada desactualizada.

Qué tiene esto de importante para un periodista digital? Pues, que cada motor de búsqueda tiene sus propias bases de datos y sus propios programas de indización con sus

correspondientes algoritmos. Esto significa que si construimos un medio de comunicación digital se debe tener en cuenta que los resultados en la recuperación de cada motor de búsqueda son diferentes.

El programa Thumshots permite comparar los resultados de búsqueda en dos motores de búsqueda. Como vemos en la siguiente figura el resultado de buscar "diariodesevilla.es" con Google y con Yahoo es diferente. Google nos lo muestra en el segundo resultado y Yahoo en el primero tal como vemos marcado por los puntos azules.

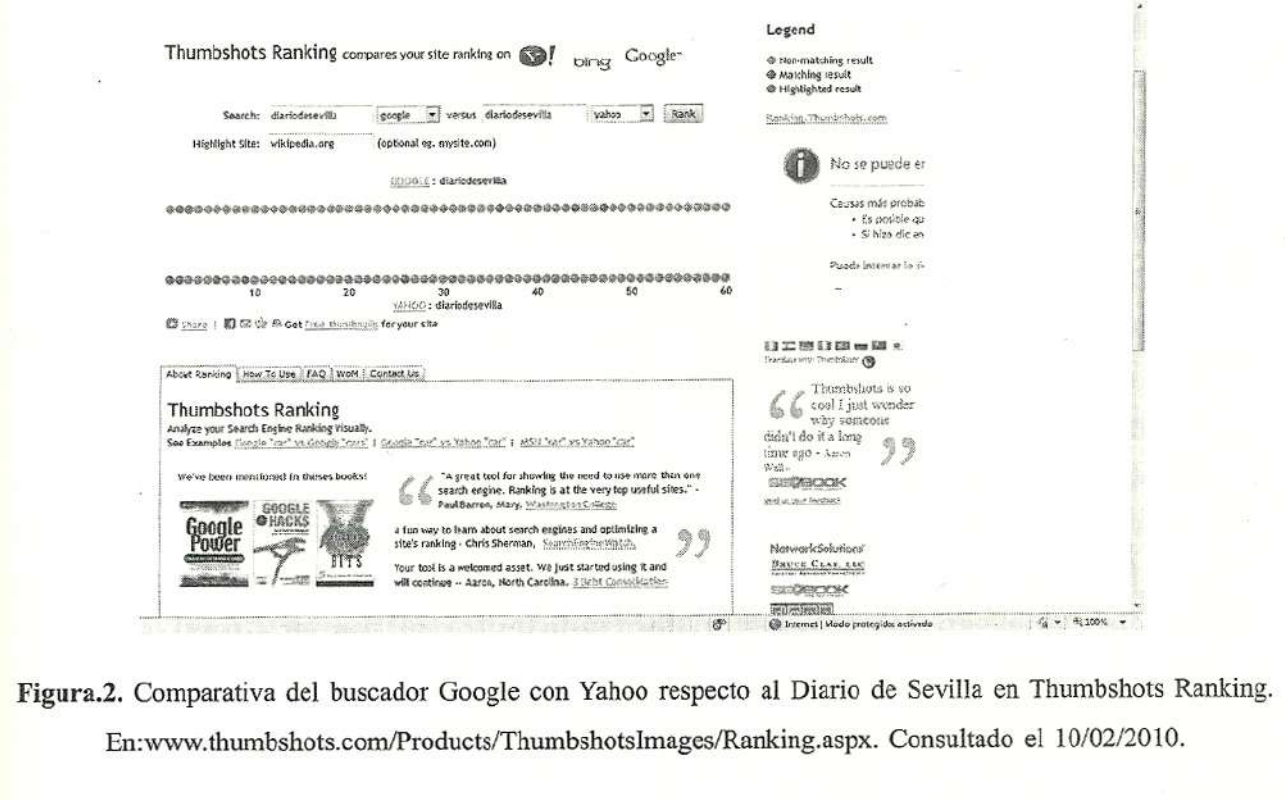

Figura.2. Comparativa del buscador Google con Yahoo respecto al Diario de Sevilla en Thumbshots Ranking.

En:www.thumbshots.com/Products/ThumbshotsImages/Ranking.aspx. Consultado el 10/02/2010.

No obstante, existen programas que permiten búsquedas simultáneas en varios motores. Es decir, en vez de comparar los resultados que cada buscador recupera se ofrece un formulario de búsqueda común a varios de ellos para que los términos a localizar puedan ser comparados con aquellos guardados en las bases de datos de los buscadores. A estos programas se les llaman multibuscadores, el más conocido es *Metacrawler*, que lanza simultáneamente las búsquedas en Google, Yahoo, Bing y Ask.

4. Técnicas de recuperación

4.1. Técnicas básicas de recuperación.

Lo primero que debemos conocer al plantearnos una búsqueda en la web son los conceptos de ruido y silencio. El ruido se concibe como los resultados de una búsqueda que en nada satisfacen la consulta planteada por el usuario; es decir, aquellos documentos recuperados por el sistema pero que no son relevantes. Suele ocurrir cuando la estrategia de búsqueda se ha definido de manera demasiado genérica. El silencio, en cambio, se plantea como la imposibilidad por recuperar información suficiente para satisfacer las demandas informativas que el usuario plantea con su consulta; se concibe como aquellos documentos relevantes para nuestra

búsqueda que no han sido recuperados. Aunque tratan el tema de consulta, no contienen los términos exactos que hemos utilizado.

¿Por qué es importante valorar el ruido y el silencio en la recuperación? Para que el periodista sea capaz de seleccionar las herramientas de búsqueda y tratamiento de fuentes informativas más adecuadas a su labor de redacción periodística. Veremos como en el caso de tener que redactar una noticia con mucha información digital, lo más adecuado es el empleo de herramientas que traten con información filtrada como los directorios; y, en el caso de necesitar una información que sea escasa será más apropiado contar con los servicios de noticias de los motores de búsqueda. Sin embargo, según la calidad que desee imprimir a su noticia deberá contar con dos conceptos básicos cuando se plantea la búsqueda de información que corrobore su artículo. Son los relativos a la precisión y la exhaustividad. Más adelante veremos el papel que juegan, pero vamos a definir por lo pronto la precisión como el número de documentos que satisfacen nuestra demanda entre el número total de documentos recuperados.

Esto se puede cuantificar a través de la siguiente expresión matemática:

$$Exhaustividad = \frac{\text{Nº de docs. relevantes recuperados}}{\text{Total de docs. relevantes}}$$

La exhaustividad, se define con la fórmula del número de documentos relevantes recuperados entre el total de documentos relevantes que debieron ser recuperados. Su expresión matemática es:

$$Precisión = \frac{\text{Nº de docs. relevantes recuperados}}{\text{Total de docs. recuperados}}$$

4.2. Truncamientos y Operadores

Cuando el periodista se plantea una búsqueda en Internet, comienza por introducir lo que considera las palabras clave correspondientes a su demanda en el formulario, el recuadro, del motor de búsqueda. Sin embargo, debe ser consciente de que los motores de búsqueda son herramientas que no piensan. Si introducimos la palabra “de”, preposición y por tanto es una palabra vacía como los artículos y los determinantes, el buscador lo interpreta como una cadena de caracteres constituida por el carácter “d”, que en código máquina se interpreta como (0100) y el carácter “e” con código máquina (1100). Así, buscar documentos que integren la palabra “de” significará para el buscador (01000,1100). Es decir, los buscadores sólo comparan las cadenas de caracteres de su base de datos con la cadena de caracteres introducida en su formulario de búsqueda. Esto supone que son incapaces de recuperar por la semántica de las palabras, lo cual supone la necesidad de que el periodista sea capaz no sólo de seleccionar las herramientas de búsqueda más adecuadas, sino de emplear filtros que depuren las palabras clave escogidas para la búsqueda.

Los filtros pueden ser automáticos o bien manuales. Los automáticos se encuentran fundamentalmente en la búsqueda avanzada de Google, si bien es posible emplearlos como comandos en la búsqueda sencilla como ya hemos visto.

Los filtros manuales permiten a los usuarios introducirlos en la declaración de su expresión de búsqueda Fundamentalmente contamos con dos tipos de filtros: i) Truncamientos y ii) Operadores. Vamos a ver algunos de los más útiles para la labor periodística.

i) Truncamientos: El truncamiento nos permite sustituir caracteres de los términos por los que deseemos realizar la búsqueda. Fundamentalmente contamos con dos: "?" para sustituir un único carácter de la cadena de caracteres y "*" para la sustitución seguida de varios caracteres dentro de una misma cadena. Los utilizamos cuando no estamos seguros del género o el número de los términos de búsqueda. Por ejemplo: Si un periodista está buscando información acerca de Bush y no recordamos la manera de designar al padre, al hijo o a uno de los padres de internet, Vannevar Bush, podemos utilizar el truncamiento "*" como antecedente de Bush para no tener que distinguir a cada uno de los tres por el nombre.

ii) Operadores Booleanos: Constan de tres tipos: 1) El operador "and" que permite la recuperación de documentos que simultáneamente contengan los dos términos que relaciona. Por ejemplo, si deseamos localizar documentos acerca de Universidad y Periodismo. Con este operador se localizan todas las páginas web que en su texto tengan ambos términos. 2) El operador "or" permite localizar todos aquellos documentos que contengan alguno de los dos términos que relaciona. Por ejemplo, si no encontramos información acerca del proceso de desamortización en la Sevilla, podemos buscar por "or" con desamortización o Mendizábal en Sevilla. 3) El operador "not" nos permite localizar todos aquellos documentos que contengan un término pero indicando que en ningún caso tengan otro término. Así podemos indicar a nuestro buscador que encuentre todas las universidades de Sevilla excepto las privadas. La sintaxis de búsqueda es: "Universidad* Sevilla" not privad* . La cadena de caracteres entre comillas se busca primero junta en un mismo documento. Los truncamientos permiten buscar documentos con expresiones como Universidad de Sevilla, Universidades Sevillanas, Universidades de Sevilla. A continuación, se filtran los resultados eliminando aquellos resultados que contengan las palabras: privado/os/a/as.

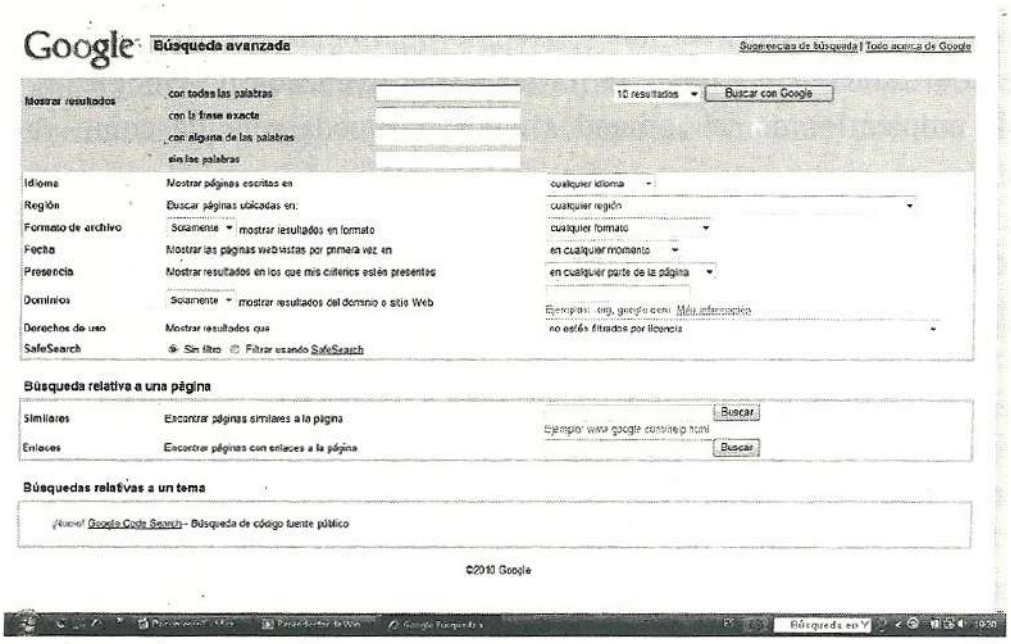

Figura 3. Formulario de búsqueda avanzada en Google. En:www.google.es/advanced_search?hl=es. Consultado el 10/02/2010.

iii) Operadores de Posición. Los operadores de posición (SAME, WITH, NEAR, ADJ) localizan registros en los que los términos son próximos dentro de un mismo documento. Si bien muchos buscadores no los emplean, son especialmente útiles en bases de datos de comunicación. Esencialmente, tienen la función de localizar frases hechas o expresiones que contengan palabras necesariamente próximas. Resultan muy útiles para las búsquedas en lengua inglesa dada la elevada“ cantidad de términos compuestos que utiliza y para relacionar palabras dentro de un campo de búsqueda pero no entre campos de búsqueda diferentes a diferencia de los operadores booleanos.

4.3. Comandos

La búsqueda con comandos permite al periodista contar con herramientas de filtrado que pueden estar recogidas en la búsqueda avanzada o no, pero que en cualquier caso permiten una adaptación de las consultas a las demandas concretas de cada usuario. Vamos a ver algunas de las más importantes.

Filetype: Permite la recuperación con las palabras introducidas en documentos con el tipo de formato exigido. Por ejemplo: “Estructura información” filetype:pdf. Nos permite la recuperación de documentos en pdf sobre estructura de la información.

Link: Muestra los vínculos que apuntan a la página especificada. De esta manera podemos saber los sitios que hacen referencia a nuestro blog o medio de comunicación digital. Se expresa así: link:diariodesevilla.es Si bien en la búsqueda avanzada de Google lo tenemos en el apartado de búsqueda relativa a una página, en Enlaces.

Cache: Con este comando podemos acceder a la última copia de una página web que se encuentra guardada en la base de datos de Google, procedente del momento en que su crawler registró la página. El comando se expresa: cache:www.publico.es Si incluimos palabras adicionales, estas se muestran subrayadas en la página de resultados. Por ejemplo: cachez:www.publico.es Sevilla.

Site: La inclusión de “site:” en la consulta limita los resultados a los sitios indicados. Esto significa que podemos buscar palabras en cualquier medio de internet. Un ejemplo sería que quisiéramos buscar noticias culturales sobre Bergamín en el diario digital elPais.es. Escribiríamos: Bergamín site:www.elpais.com.

Intitle: Restringe la búsqueda a los títulos de los sitios web, en lugar de tener en cuenta toda la página. Por ejemplo, supongamos que deseamos buscar páginas en cuya cabecera <HEAD> se encuentre por título la palabra Ámbitos. Veremos cómo el código fuente de todas las páginas que nos muestra tiene esta palabra en el título. Este comando limita de forma sensible el número de resultados que se muestran respecto a realizar la búsqueda únicamente con el término.

Inurl: Es similar al comando Intitle en cuanto a su funcionalidad. Sin embargo, en este caso la búsqueda se realiza dentro de la URL de los sitios. Vamos a tomar el mismo ejemplo que antes y buscamos: Inurl:ámbitos. Vemos que obtenemos mucho ruido en la web. Entonces, ¿para qué nos puede servir este comando? Pues para la localización de archivos concretos en sitios web. Así, podemos buscar el catálogo de la Universidad de Sevilla combinando los dos comandos, “insite:www.us.es inurl:fama”, para acceder al catálogo general de la Universidad de Sevilla.

Define: Comando que le permite al periodista obtener definiciones sobre algún término. Es especialmente útil para términos técnicos, médicos y jurídicos. Por ejemplo, si estamos cubriendo una noticia acerca de una medicación para combatir una “encefalopatía espongiforme”. Debemos señalar que el mayor número de definiciones las encontramos en lengua inglesa, con lo que se hace necesario el empleo de un traductor.

4.4. Recuperación de imagen estática

La recuperación de imagen puede realizarse por contexto y por contenido. La recuperación por contexto es aquella que realizamos empleando términos para que el buscador al compararlos con las palabras clave o descriptores de su base de datos pueda recuperar las imágenes asignadas a éstas. Este tipo de recuperación se produce fundamentalmente en tres ámbitos: I) En los motores de búsqueda comerciales, como Googleimages o Yahooimages. En ellos, se emplea el mismo formulario de búsqueda que el empleado para la búsqueda web, tanto en la búsqueda sencilla como avanzada. Esto simplifica notablemente su utilización, dado que las técnicas de búsqueda son las mismas que para las búsquedas de texto. Con la única salvedad del comando filetype que nos permite realizar las búsquedas por distintos formatos de imagen, según la resolución y calidad de imagen que se desee obtener de las mismas. Un ejemplo de ello es la recuperación de imágenes sobre la Universidad de Sevilla en formato jpg. Acudimos a Googleimage y escribimos: "Universidad de Sevilla filetype:jpg. La gran ventaja que tienen estos motores frente a los que expondremos a continuación es que permiten también la obtención de infografías y anagramas. Para ello, sólo tenemos que pedir un formato "gif" en filetype. La búsqueda avanzada con la que cuentan evita que el usuario tenga que conocer los comandos y los formatos de recuperación. Asimismo, añade filtros para la recuperación como la posibilidad de recuperar por derechos de uso, la coloración o el tipo de contenido. II) en los repositorios comerciales de fotografías como Gettyimages o Jupiter. Aquí, los resultados de las búsquedas permiten no sólo obtener más información acerca de las imágenes recuperadas y una garantía de su calidad, sino que ofrece una seguridad jurídica al periodista sobre su empleo para su propio medio de comunicación. Las imágenes, como cualquier obra original, cuentan con una protección de los derechos de autoría. Cuando utilizamos una imagen obtenida en Internet, no solemos averiguar el autor de la misma, y menos todavía nos ponemos en contacto con él para solicitar su utilización. Sin embargo, el periodista debe ser muy cauto con el empleo de la información obtenida en Internet, ya que tiene la obligación de contar con la autorización de los autores de aquellos documentos de los que no es responsable, para no incurrir en la violación de los derechos de autor. En estos repositorios, el motor de búsqueda permite la recuperación de imágenes libres de derecho de autor. Si desea obtener imágenes con royalties tiene la posibilidad de comprarlas a Gettyimages, para lo cual se envía la imagen sin marca de agua de la empresa. El pago evita cualquier problema legal en el futuro con los autores. Asimismo, cuentan con unos motores de búsqueda más adecuados para la recuperación de fotografías, con filtros que permiten considerar la información técnica acerca de las cámaras, los planos o los filtros empleados para la recuperación de fotografías. Esto, junto con los filtros empleados para la recuperación por los contenidos de las fotografías resulta de gran valor para el fotoperiodista. III) En la web social, a través de herramientas específicas para compartir imágenes, como Flickr o Picasa. Sin embargo, los descriptores empleados para la indización de fotografías en la web social no son parte de un lenguaje controlado sino que son los propios usuarios que cuelgan las fotografías quienes deciden los descriptores a asignar a sus imágenes. Dando lugar así a un lenguaje específico denominado Folksonomy. Si bien se trata de un repositorio muy importante de fotografías, debemos decir que su calidad es variada y no cuentan con motores de búsqueda sistemáticos, capaces de recuperar fotografías de acuerdo con las necesidades prefijadas por el periodista. Eso sí, resulta fácil pedir autorización a los autores para su empleo, ya que las cuentas en la web social van asociadas a un correo electrónico, quienes suelen autorizarlo gratuitamente si no se trata de contenidos del ámbito personal.

La recuperación por contenido se produce con herramientas creadas ex profeso para ello. Se caracteriza por la recuperación de imágenes únicamente por sus características intrínsecas y no por términos que le son asignados. La recuperación no se mediante la formulación de una consulta textual sino seleccionando las propiedades de búsqueda como los códigos de color empleados o las formas representadas en la imagen. La herramienta más conocida al respecto es QBIC de la empresa IBM para la recuperación de imágenes sobre las obras artísticas expuestas en el museo del Hermitage en San Petersburgo. De momento, no se están empleando este tipo de herramientas en los centros de documentación periodísticos dado lo oneroso que resulta y los resultados cuestionables que ofrece. Sin embargo, pensamos que los futuros sistemas gestores de contenidos de las redacciones digitales las irán integrando de manera conjunta con las tradicionales herramientas de recuperación de imagen por contexto.

4.5. Recuperación de imagen en movimiento.

Dada su complejidad, los motores de búsqueda comerciales cuentan con la posibilidad de recuperar videos, pero todavía son muy escasos los filtros que pone a disposición del periodista para poder seleccionarlos. Concretamente, Google videos, a las opciones de la búsqueda avanzada con las que cuenta para la búsqueda web le añade los filtros específicos de vídeo para la búsqueda por duración (corto, medio y largo), si cuentan con subtítulos, si son reproducibles en Google y la posibilidad de seleccionar el tipo de formato de vídeo por el cual recuperar. Para Yahoo videos, no existe una opción de búsqueda avanzada, sino que únicamente añade una pestaña para la recuperación con el criterio de su duración. Respecto a Youtube, debemos aclarar que es parte de la iniciativa de web social, con lo cual la calidad de los vídeos es muy variable. En cuanto a las opciones de recuperación resultan muy escasas, ya que únicamente permite la recuperación por tres tipos de materias (deportes, música y ocio), por dos clases de duración (corto, largo), por la fecha de subida del vídeo y por características como si tiene subtítulos o si se trata de vídeos en alta definición. Estas carencias de los motores de búsqueda de videos se deben a las dificultades intrínsecas de sus características y a que los centros de documentación de los medios audiovisuales cuentan con sistemas gestores muy costosos por la dificultad de la descripción de los contenidos y la constante evolución de los formatos de vídeo.

5. Noticias en Internet

Hoy día, contamos con numerosas plataformas para la difusión de noticias. Consciente de ello, las empresas dedicadas a ofrecer servicios de búsqueda, como Google o Yahoo, ofrecen su propio servicio de noticias multilingüe, con noticias específicas de cada unos de los países a los que se ofrece el servicio. La gran ventaja que ofrece respecto a los medios de comunicación digitales es que aglutinan muy variados canales de información, desde la ofrecida por los diarios digitales hasta los weblogs o las listas de distribución. El servicio de noticias de los buscadores captura todas las informaciones provenientes de todo este tipo de canales, las une, las ordena un poco y las refleja hacia los usuarios de Internet en forma de Noticias o News y grupos de noticias del buscador. Según Calishain, la página de noticias de Google consulta 4.500 fuentes de información diferentes. Su página principal se actualiza diariamente a través de unos algoritmos propios sin necesidad de la intervención de un periodista, si bien las noticias que integran provienen de medios donde las han escrito periodistas. Las noticias más relevantes aparecen en la parte superior de la página. Estas se organizan en grupos, en el margen izquierdo, uniendo información sobre la noticia y fotografías procedentes tanto de agencias de noticias como de medios digitales. Debajo de cada una de las noticias se muestran sus fuentes proveedoras, así como el tratamiento que cada medio le ha dado a esa misma noticia. Normalmente, en el caso de Google Noticias si lo tiene, cuenta con una opción de búsqueda avanzada que permite a los usuarios contar, no sólo con las posibilidades de búsqueda propias de la web, sino con filtros específicos para noticias. Uno de ellos es el correspondiente a “Fecha” que nos permite realizar el seguimiento de una noticia con los antecedentes de un mes. Asimismo se ofrece la posibilidad de seleccionar la aparición de una noticia en un periodo de tiempo definido. En caso de necesitar el acceso retrospectivo a noticias con mayor antigüedad a un mes, se permite el acceso a una hemeroteca. Finalmente, se permite la elección de los medios sobre los cuales realizar las búsquedas, lo cual permite la realización de análisis comparativos acerca del tratamiento de las noticias en los distintos medios.

Debemos destacar que todas estas opciones, no sólo son aplicables en cada una de las secciones, sino que también se permite la selección de la estructura de la noticia sobre la cual realizar las búsquedas, en los titulares, en el cuerpo de texto, etc. Asimismo, al igual que teníamos para las búsquedas en web, contamos con un conjunto de comandos específicos para la localización de noticias que van más allá de la interfaz de búsqueda avanzada. Vamos a ver cuáles nos pueden ser de más utilidad, teniendo en cuenta que sólo podemos emplearlos desde la interfaz de Google News:

Intitle. Para la búsqueda de palabras clave en los titulares de los artículos. Su variación **allintitle** permite localizar noticias en las que varias palabras clave aparecen en el titular.

Intext. Busca los términos en el cuerpo de la noticia. También cuenta con la variante allintext para la recuperación de noticias donde una serie de términos aparecen en el cuerpo del artículo.

Inurl. Busca las palabras clave en el URL de una noticia.

Source. Permite la búsqueda de artículos procedentes de fuentes seleccionadas con anterioridad a la búsqueda. Si bien, no están todos los diarios digitales sino fundamentalmente los anglosajones. Por tanto, es muy útil para conocer el tratamiento de un tipo de información particular entre los medios anglosajones. Por ejemplo, podemos buscar noticias sobre el tratamiento de la cumbre de Copenague en el New York Times así: copenhagen summit source: new_york_times.

Location. Lo podemos emplear para filtrar fuentes pertenecientes a un determinado espacio geopolítico. Por ejemplo, podemos buscar artículos publicados en la prensa nipona sobre Haiti así: Haiti location:japan.

No obstante, los servicios de noticias de los motores de búsqueda comerciales no son los únicos en ofrecer noticias. Últimamente, con la expansión de la comunicación por Internet, han aparecido motores de búsqueda monográficos de noticias. Dos de los más conocidos son: i) Rocketinfo. Se trata de un buscador que emplea fuentes de noticias muy técnicas. Asimismo, proporciona acceso directo a noticias de agencias internacionales, así como permitir la sindicación directa del usuario a cada una de las noticias de agencia; ii) Yahoo Daily News. Muestra su lista de fuentes en la página de búsqueda avanzada. Al contar con un índice de 30 días, es posible localizar noticias ya eliminadas de otros buscadores. También cuenta con la posibilidad de recuperar noticias escritas en un idioma determinado, o perteneciente a una sección. Incorpora formularios que evitan el empleo de los comandos que acabamos de ver para la recuperación de artículos.

Finalmente, vamos a comentar un servicio de reciente auge entre los medios de comunicación digital, la sindicación de contenidos o RSS. Este servicio de valor añadido ha tenido especial incidencia en los portales de comunicación porque permite a sus clientes estar actualizados acerca de cualquiera de las noticias difundidas. Hasta hace poco existían programas específicos para la sindicación, en la prensa digital española el más conocido fue el “mini20”, un agregador de noticias creado por el diario gratuito digital 20minutos para que sus usuarios pudieran seguir las novedades de un acontecimiento. La sindicación, según Guillermina Franco, no es un concepto nuevo en el periodismo. Surgió a finales del siglo XIX entre la prensa estadounidense aprovechando la extensión de las líneas de telégrafo por todo su territorio. Los diarios mantenían corresponsales en las principales ciudades del país quienes, a través del telégrafo, podían mantener informado, con noticias redactadas telegráficamente, a su periódico de los eventos que sucedían en su región. En un país con tantas migraciones como Estados Unidos, esto permitía a los grandes periódicos mantener una sección de noticias breves y locales para unos clientes con interés por los acontecimientos de un territorio muy amplio. Con el tiempo, el concepto fue cayendo en desuso hasta que aparecieron los blogs, especialmente durante la guerra del Golfo Pérsico de 1991. Los soldados, empezaron a tener un papel protagonista en la generación de noticias sobre la guerra al emplear sus blogs como canales de información de la situación que estaban viviendo. Sin embargo, los medios de comunicación y los ciudadanos no podían estar constantemente revisando los blogs de centenas de soldados que no se sabía cuando actualizaban sus blogs. Por ello, se puso en marcha una herramienta denominada sindicador de contenidos. Esta permitía a los usuarios decidir los blogs o canales a los que deseaban suscribirse. Con ello, los usuarios empleando un agregador de noticias recibían un mensaje telegráfico con las últimas modificaciones de los canales a los que se suscribían. Hoy día, los navegadores han incorporado la posibilidad de suscribirse a canales de información, con lo que han desaparecido los agregadores de noticias.

LAS CLAVES

La documentación informativa es un saber práctico destinado a la recuperación de información, mejora de su calidad, almacenamiento, difusión y reutilización.

El periodista debe considerar el ruido y el silencio en la recuperación antes de seleccionar la herramienta de recuperación.

Los portales de comunicación proporcionan una plataforma muy útil para que el periodista comience a seleccionar sus fuentes por la calidad de sus recursos.

El periodista cuenta con directorios y mapas de sitio que los sitios web utilizan para la organización de la información. Es muy útil su empleo cuando se conoce el leguaje y los recursos informativos de la materia sobre la que se redacta.

Internet no es un medio periodístico pero tiene un gran valor para el periodista tanto por los nuevos formatos narrativos que ofrece como por dotar a sus clientes de un papel protagonista de la noticia.

Contamos con el servicio de noticias de los buscadores comerciales como fuente de información de noticias y fuentes de información.

La recuperación de imagen estática se puede realizar con buscadores comerciales, repositorios de imágenes y repositorios de web social. Pero debemos consultar las licencias de autor.

La recuperación de imágenes en movimiento es la más compleja y la que cuenta con menos herramientas a nivel de usuario para su recuperación.

El crecimiento de los medios en Internet hará que aparezcan nuevas herramientas de búsqueda y difusión de información que el periodista no podrá desconocer si quiere sobrevivir en un mundo de la información cada vez más competitivo.

La sindicación de contenidos proporciona al periodista un servicio rápido y escueto para el seguimiento de noticias, pero no sirve de nada si no las "contrasta con sus fuentes de información y las inserta en una noticia periodística.

CONSEJOS PRÁCTICOS

Antes de realizar una búsqueda recoger en papel la estrategia de búsqueda que nos planteamos.

Emplear los buscadores de los medios especializados que utiliza antes que los comerciales.

Utilizar las búsquedas avanzadas preferentemente para poder combinar varios criterios de búsqueda y obtener información más contrastada y veraz.

El periodista debe editar su propio directorio de fuentes de información en Internet por medio de una herramienta destinada a ello como “Favoritos” o “Bookmarks”.

Se aconseja la participación del periodista en listas de distribución y foros acerca de las noticias sobre las que redacta, dado que constituyen una fuente rica de información.

Es importante que el periodista aprenda a utilizar herramientas de sindicación de contenidos para poder mantenerse actualizado acerca de las fuentes de información a las que está suscrito en Internet.

Las tecnologías relacionadas con Internet tienen una renovación muy rápida y una creciente presencia en la comunicación. Es importante mantenerse actualizado acerca de la aparición de nuevas tecnologías para la comunicación en las secciones especializadas en nuevas tecnologías de los medios.

Emplear los archivos multimedia y hemerotecas que tanto los diarios digitales, como las agencias de noticias o los gabinetes de prensa ponen al servicio del periodista gratuitamente o con una suscripción.

Contar con un gestor de contenidos capaz de trabajas con múltiples formatos, con el cual el periodista pueda almacenar la información seleccionada para la redacción de una noticia.

Emplear los traductores web para acceder a información publicada en entornos digitales que utilizan un alfabeto distinto al occidental. Esto permite al periodista contar con un rango de fuentes que van más allá de las occidentales.

Fuentes documentales

a) Bibliografía esencial

Fuentes i Pujol, M. E. *Manual de documentación periodística*. Madrid: Síntesis, 1995.

Galdón, G. (coord.). *Teoría y práctica de la documentación informativa*. Barcelona: Ariel, 2002.

García Gutiérrez, A. (ed.). *Introducción a la Documentación Informativa y Periodística*. Sevilla : MAD, 1999.

Moreiro González, J .A.(coord.). *Manual de documentación informativa*. Madrid : Cátedra, 2000.

Rubio Lacoba, M. *Documentación informativa en el periodismo digital*. Madrid: Síntesis, 2007.

b) Bibliografia complementaria.

Bradley, P. *The advanced internet searcher is handbook*. London : Facet Publishing, 2004.

Calishain, T. ; Domfest, R. *Google. Los mejores trucos*. Madrid : Anaya Multimedia, 2005.

Maldonado Martínez, A. ; Rodríguez Yunta, L. *La información especializada en Internet. Directorio de recursos de interés académico y profesional.* Madrid: Consejo Superior de Investigaciones Científicas, 2006.

Pareja, V.(coord.). *Guía de Internet para periodistas*. Madrid: Consejo Superior de Investigaciones Científicas, 2002.

Salazar, I. *Las profundidades de Internet. Accede a la información que los buscadores no encuentran y descubre el futuro inteligente de la Red*. Gijón : Trea, 2005.

b) Otras Fuentes:

Alemany Martínez, L. "La disciplina Documentación informativa en los planes de estudio de las licenciaturas de Publicidad y Relaciones Públicas". *Cuadernos de Documentación Multimedia*, nº10, 2000. Consultado el 10/02/2010 en: http://www.ucm.es/info/multidoc/multidoc/revista/num10/ paginas/pdfs/Lalemanx.pdf..

Codina, Ll. "La Documentación en los medios de comunicación: situación actual y perspectivas de futuro". *Actas del I Congreso Universitario de Ciencias de la Documentación*. 2000. En:www.ucm.es/info/multidoc/multidoc/revista/num10/paginas/pdfs/Codina.pdf. Consultado el 10/02/2010.

Codina, Ll. "Documentación Periodística". En: http://www.mindomo.com/view.htm?m=06523ef513ff404a9048b0d95911fad3. Consultado el 10/02/2010.

e-periodistas. En: hgg://www.uuav.es/fcom/guia/. Consultado el 10/02/2010.

InCOM-UAB. Portal de la Comunicación. El portal de los estudios de comunicación. En: http://www.portalcomunicacion.com/esp/home.asp. Consultado el 10/02/2010.

Infoamérica.org. El portal de la Comunicación. En: http://www.infoamerica.org/. Consultado el 10/02/2010.

IPTC. International Press Communications Council. En: http//www.iptc.org/cms/site/index.html/channel=CH0086. Consultado el 10/02/2010.

Micó Sanz, J. Ll. ; Masip Masip, P. ; García Avilés, J.A. "Periodistas que ejercen de documentalistas (¿y viceversa?): nuevas relaciones entre la redacción y el archivo tras la digitalización de los medios". *El Profesional de la Información*.vol.l8, nº3, 2009, p. 284-290.

Newslink. En: http://www.newslink.org/. Consultado el 10/02/2010.

Tejedor, C. "Documentación en medios de comunicación. Agencia EFE". En: www.ucm.es/info/multidoc/multidoc/cursos/verano/material/EFE_CONCHA%. Consultado el 10/02/2010.