# Métodos de búsqueda bibliográfica

C. Espelt

Escola Universitària J. Rubió i Balaguer de Biblioteconomia i Documentació. Barcelona.

## Introducción

A modo de introducción se presenta la influencia de los avances de las comunicaciones informáticas y la aparición de los soportes ópticos en las búsquedas realizadas por usuarios finales. Se describen a continuación las etapas que comprende la búsqueda desde que surge una necesidad de información hasta la transformación en una estrategia, para más adelante pasar al análisis de los distintos elementos que intervienen en los procesos de recuperación de la información: la estructura de la base de datos, el lenguaje usado en la búsqueda y las posibilidades del *logical* de recuperación.

## La información electrónica. Avances en la última década

Antes de presentar las etapas que intervienen en la búsqueda de información, creemos de utilidad ofrecer una breve panorámica de la situación actual, con el objetivo de apuntar los factores que condicionan al intermediario y al usuario final ante una búsqueda.

Dejando a un lado el crecimiento exponencial de la información científica, en los últimos 10 años los avances en el área de las telecomunicaciones y de los nuevos soportes electrónicos han modificado sustancialmente los procesos de búsqueda bibliográfica.

El uso generalizado del soporte óptico, y especialmente del CD-ROM, presentado comercialmente en la Feria del Libro de Frankfurt de 1985, es uno de los factores esenciales del cambio. Actualmente el mercado de la información científica electrónica se reparte entre los productos en CD-ROM, 30 novedades en el último semestre de 1994, y las bases de datos en línea, 44 en el mismo período[1]. En el contexto temático que nos afecta citaremos tres tipologías: monografías de referencia (*The Particle Atlas Electronic Edition* publicado por MicroDataware), publicaciones periódicas en texto completo (*FDA on CD-ROM* de la Food and Drug Administration, *Drugs of the Future on CD-ROM* de Prous) y las clásicas bases de datos bibliográficas (*BiblioMed Professional Series* de National Library of Medicine o *Excerpta Medica CD-Drugs and Pharmacology* de Elsevier Science Publishers).

El CD-ROM ha supuesto la familiarización del usuario final con el soporte electrónico en la visualización de la información y de técnicas de búsqueda. Una vez realizada la suscripción a un producto determinado en CD-ROM, el factor tiempo y el número de búsquedas realizadas no implican coste alguno; de esta forma, el usuario final dispone de un eficaz instrumento de aprendizaje de búsqueda.

El avance de las telecomunicaciones y, más concretamente, el acceso a Internet, ha tenido también importantes consecuencias que intentamos resumir a continuación:

### *Cantidad y variedad de información accesible*

Las fuentes de información disponibles por vía telemática incluyen actualmente una gran variedad de recursos de información como catálogos de bibliotecas, publicaciones electrónicas en texto completo, información informal disponible en listas de distribución, etc., al lado de las bases de datos bibliográficas, numéricas y directorios distribuidas por *hosts* comerciales.

### *Ampliación de los recursos de información desde la propia estación de trabajo*

Las redes locales de ordenadores personales permiten instalar el programa de comunicaciones en el servidor, acercando y ampliando las posibilidades de uso de recursos externos al usuario, sin necesidad de desplazarse.

*Incorporación de otras funcionalidades*

Un claro ejemplo de ello es la ampliación de las posibilidades de petición y recepción de documentos incluyendo los gráficos correspondientes, usando *file transfer protocol* para importarlos a través de Internet.

*Ahorro de costes de telecomunicaciones*

Desde el enfoque de esta comunicación, la mayor ventaja del acceso por vía Internet es la posibilidad de crear sistemas que permitan el acceso a través de una única *interfaz* a todo tipo de recursos informativos de interés para profesionales de un área específica. Los *logicales* de acceso temático a *World Wide Web* son instrumentos válidos, aunque su rendimiento es muy bajo en búsquedas concretas[2].

Pero se advierten también consecuencias negativas, que habrán de conllevar modificaciones del sistema actual. Martha Williams[3] ofrece una lista de 20 limitaciones, que agruparemos en: falta de control (probablemente de seguridad, propiedad intelectual, privacidad); falta de coherencia (cambios frecuentes, deficiencias de gestión, coexistencia de recursos de diferente nivel de calidad), y desconocimiento de los recursos disponibles y del tipo de indización aplicada.

Si consideramos la información como un recurso estratégico coincidiremos en afirmar que más información no equivale a mejor información, y los cambios que estamos viviendo nos hacen avanzar rápidamente por la vía de la cantidad pero, siendo optimistas, sólo muy lentamente en la de la calidad.

## Etapas previas a la búsqueda

El elemento que origina una búsqueda es siempre una necesidad de información; en el momento en que se decide satisfacerla se expresa y se transforma así en una demanda.

Antes de llegar al planteamiento de los argumentos de búsqueda, se requiere un análisis de la demanda y la selección de los recursos que utilizaremos para satisfacerla. La estrategia deberá tener en cuenta el análisis previo y las características específicas de los recursos seleccionados.

Aunque cuando al referirse a técnicas de búsqueda se suelen obviar estas fases previas, de ellas dependen en gran medida los resultados que obtendremos. La coincidencia entre necesidad de información y demanda no es nunca exacta; en el caso de actuar un intermediario, para el usuario final es difícil expresar todos los matices de la necesidad; cuando es el mismo usuario quien realiza la búsqueda da por sentado que sabe muy bien lo que busca y, al no tener necesidad de expresarlo, confía en su capacidad de valorar los resultados sin tener bien presentes los objetivos de la búsqueda.

El análisis de la demanda comprende la delimitación del contexto y la delimitación conceptual. La delimitación del contexto ayuda a detectar el nivel de dificultad de la búsqueda y a seleccionar los recursos a utilizar. Se deberá tener en cuenta el volumen de información existente sobre el tema, el número aproximado de registros que el usuario considera adecuado para el tiempo que dispone, y la especificidad de los conceptos que incluye la demanda, todo ello estrechamente vinculado al nivel de exhaustividad requerido. Los aspectos formales, como la lengua, el período cronológico que interesa abarcar y el tipo de documento completan la delimitación del contexto.

La delimitación conceptual consiste en la identificación de los elementos conceptuales independientes que constituyen la demanda y el tipo de relación que se establece entre ellos.

Partiendo de un buen análisis conceptual y formal se dispone de la información necesaria para la selección de la fuente de información adecuada a las necesidades. En la selección, aparte de los factores de contenido (área de especialización, cobertura geográfica, tipos de documentos o actualización), intervienen también factores externos como el coste, la accesibilidad y el grado de familiarización. Algunas búsquedas de patentes pueden realizarse en *CA Search*, cuyo precio de conexión por minuto y por registro visualizado es menos elevado que las tarifas de *Derwent World Patents Index*.

En el caso de las bases de datos en línea, algunos distribuidores ofrecen instrumentos que facilitan la selección. DIALOG permite a través de *Dialindex* probar una determinada estrategia en un grupo muy amplio de bases de datos, para identificar las que contienen más información.

También existe la posibilidad de realizar búsquedas en un conjunto de bases de datos e identificar los registros duplicados antes de visualizar los resultados.

## La estrategia de búsqueda

Los tres elementos a tener en cuenta en la preparación de una estrategia son la estructu-

ra de la base de datos, el lenguaje que utiliza y las posibilidades del *logical* de recuperación.

*Estructura de la información*

Las bases de datos están formadas por registros que, a su vez, están divididos en campos. Cada uno contiene un elemento de información en el registro. La fragmentación en campos permite la creación de índices y la posterior recuperación de los registros a partir de los elementos de la estrategia de búsqueda. Algunos campos no se indizan debido a que la información que contienen no es previsible que sea útil como elemento de recuperación (p. ej., la paginación). La indización de los campos puede ser por bloques de campo completo, palabra a palabra (excluyendo las palabras vacías) y por frase o secuencia de palabras.

Cuando introducimos un argumento de búsqueda sin ningún tipo de prefijo ni sufijo el ordenador lo busca en una serie de campos, unos indizados por palabras y otros por frases, que expresan contenido (título, descriptores, identificadores y resumen). Los campos indizados por bloques no presentan tal flexibilidad para la recuperación ya que sólo son recuperables por la primera parte del contenido del campo.

La preparación de una estrategia debe tener en cuenta qué campos son los idóneos para cada concepto y las características de indización de los campos elegidos.

*Lenguajes*

El uso del lenguaje natural (o libre) para la recuperación en bases de datos bibliográficas y de texto completo es prácticamente obligado en todo tipo de búsquedas. La principal ventaja que ofrece es la comunicación directa entre el autor del documento y el usuario, y su uso es esencial para conceptos muy específicos o novedosos. Las búsquedas en lenguaje natural requieren la inclusión de un número elevado de términos para cada concepto, debido a la abundancia de sinónimos y términos de diferente nivel jerárquico en la comunicación científica.

Algunas bases de datos disponen de ayudas que permiten recuperar singular y plural, nombres completos y abreviaturas o siglas, independientemente de la forma empleada por el usuario.

La ambigüedad del lenguaje natural provoca ruido (recuperación de documentos que no son adecuados) y silencio (no se recuperan documentos que interesarían). Para compensar fallos de ruido, especialmente en bases de datos de texto completo, se han introducido menús diseñados para facilitar la búsqueda como el sistema Target en Dialog, que selecciona los 50 registros con un porcentaje más alto de ocurrencia de los términos de la búsqueda[4].

Pero la forma tradicional de atenuar los fallos de recuperación consiste en utilizar lenguajes artificiales que pretenden ser unívocos, asignando una única forma a cada concepto. Se pueden diferenciar dos tipos básicos: los sistemas de codificación sin estructura y los lenguajes controlados.

Los primeros se usan para representar conceptos geográficos, nombres de empresas, tipo de documento, etc. (un claro referente sería el número de registro de *Chemical Abstracts*). La mayoría de bases de datos incluyen más de un campo destinado a este tipo de códigos.

Los lenguajes controlados representan conceptos temáticos estructurados de forma jerárquica o facetada. El uso de descriptores y códigos de clasificación permite aumentar la recuperación utilizando un único término y, por otra parte, seleccionar únicamente el contexto temático que es pertinente. Su estructuración ayuda al usuario a identificar los términos que debe incluir la estrategia para recuperar los conceptos en el grado de especificidad y enfoque requeridos. En este sentido, la utilidad de los lenguajes controlados depende de su calidad. Las bases de datos científicas y técnicas tienen una gran tradición en la elaboración de sistemas de indización y clasificación de calidad adaptados a sus necesidades.

La estrategia recomendada consiste en determinar el nivel de generalidad de cada concepto y aplicar el tipo de lenguaje más apropiado en cada caso: códigos de clasificación para los más generales (disciplina, enfoque o tratamiento), descriptores para la mayoría de conceptos, y lenguaje libre para los que no están representados en lenguaje controlado por su gran especificidad o reciente introducción.

Logical *de recuperación*

Los dos elementos anteriores son propios de cada base de datos, mientras que el *logical* de recuperación es independiente de éstas. A una misma base de datos se le aplican distintos *logicales* según el distribuidor o su presentación en CD-ROM. El instrumento de diálogo del usuario con el almacén de datos es el *logical* de recuperación, que ofrece un conjunto de instrucciones para la interrogación.

La dificultad principal de una búsqueda no radica en la correcta utilización de las instrucciones del lenguaje de recuperación, sino en seleccionar la fuente de información adecuada según el tipo de información que contiene y las posibilidades de búsqueda que ofrece su estructura, y las características de los lenguajes controlados que aplica.

Las principales funciones de búsqueda del *logical* son la truncación, los operadores booleanos y los operadores de proximidad. La truncación nos permite recuperar palabras que comparten una misma secuencia de caracteres, variaciones singular/plural, formas verbales o adjetivas, una misma raíz sin tener en cuenta prefijos o sufijos, etc. Amplía la recuperación y su aplicación resulta imprescindible en el caso de lenguaje libre.

Los operadores booleanos continúan siendo el método de recuperación más utilizado en la recuperación de la información. Los resultados de una búsqueda dependen en gran medida de la correcta definición de las relaciones lógicas entre los conceptos que intervienen en la demanda, y estas relaciones se representan mediante las operaciones de adición, intersección y negación.

El principal inconveniente de este tipo de operaciones es la falta de flexibilidad. La lógica booleana no acepta la aproximación y, por ello, es difícil definir estrategias booleanas para recuperar un determinado número de documentos. A menudo con la adición de varios términos para cada concepto se consiguen resultados parciales prometedores, pero al realizar la intersección entre los conjuntos iniciales el resultado final puede ser de 0 documentos. En este sentido, es especialmente peligroso utilizar el operador de negación, ya que se pueden eliminar registros pertinentes que incluyan el término en cuestión.

Los operadores de proximidad son conectores que especifican la distancia entre las palabras del argumento de búsqueda, por tanto, implican un nivel superior de restricción respecto al operador de intersección. Los términos deben figurar en el mismo registro o campo pero, además, se añade la condición de que aparezcan a una determinada distancia. Los operadores de proximidad se aplican en la recuperación a partir de campos de texto indizados por palabras, como título, resumen o texto completo.

Dentro de los procedimientos de selección los *logicales* de recuperación permiten otras funciones de gran utilidad como los límites por año de publicación, lengua original del documento, etc., o la posibilidad de guardar códigos o términos obtenidos mediante la consulta de diccionarios para utilizarlos como elementos de búsqueda en otras bases de datos.

## Visualización de índices y resultados

Por el momento, la información electrónica resulta menos transparente para el usuario que la información impresa. Es más fácil hacerse una idea de una biblioteca dando un paseo entre las estanterías que sentándose frente a un terminal, y todos habremos encontrado alguna vez una referencia interesante simplemente hojeando una revista de *abstracts*, o incluso mientras buscábamos otra cosa. Con las aplicaciones de hipertexto se están ampliando las posibilidades de navegar sin las limitaciones de la ordenación física, pero los resultados no permiten todavía obtener visiones de conjunto.

En esta línea, los *logicales* ofrecen la posibilidad de visualizar los diversos índices de la base de datos. Esta función es muy útil en el caso de los campos indizados por unidad para comprobar antes de la selección de las formas que debemos introducir. Un ejemplo de ello es la búsqueda por autores, ya que debido a la falta de normalización pueden figurar en el índice bajo más de una forma. Las bases de datos que aplican lenguajes controlados permiten generalmente la consulta del tesauro en soporte electrónico, con la ventaja adicional de una permanente actualización y de poder valorar la inclusión de un descriptor según el número de documentos que recuperará.

La visualización del conjunto final tiene como función principal la evaluación de los resultados obtenidos, para determinar si es necesario modificar la estrategia y volver a ejecutar la búsqueda. Existen una gran variedad de formatos disponibles y, en general, el usuario puede definir su propio formato. El formato de visualización del contexto en el que aparecen los términos del argumento de búsqueda es muy útil para localizar truncaciones demasiado amplias, operadores de proximidad poco restrictivos, etc.

Sin embargo, los usuarios finales, influenciados por sus necesidades, tienden a leer detenidamente los textos recuperados y valorar directamente el contenido de los registros en sí mismos, sin plantearse si responden a la ecuación de búsqueda planteada. El usuario final no está preparado para detectar los fallos de la estrategia o juzgar si el sistema de recuperación actúa de forma correcta. Los estudios sobre búsquedas realizadas por usuarios finales coin-

ciden en señalar la aceptación de resultados muy mediocres como buenos. Cuanto mayor es el desconocimiento de un producto, las expectativas son menores y la satisfacción más alta[5].

## Tendencias de futuro

Coincidiendo con el incremento de productos dirigidos al usuario final se ha facilitado la búsqueda en el lenguaje de interrogación con el uso de menús para usuarios ocasionales, o con las pasarelas *(gateways)* que permiten usar un único lenguaje de interrogación para acceder a varios distribuidores, ofreciendo distintos niveles según la experiencia del usuario[6].

Siguiendo esta línea sólo se trabaja en uno de los tres elementos básicos de la recuperación de la información, no se tienen en cuenta ni la estructura ni el lenguaje propio de cada base de datos.

La tendencia lógica y necesaria es que las *interfaces* consideren también estos elementos y un único *logical* reconozca los distintos formatos y haga de enlace con los lenguajes controlados que los productores utilizan en sus bases de datos. Un ejemplo de este planteamiento limitado a la base de datos *Medline* es la *interfaz* CANSEARCH desarrollada como ayuda en las búsquedas de terapia del cáncer. Está basada en una pantalla táctil que proporciona un menú para expresar el tema de la búsqueda con una combinación de cuatro facetas: *a)* localización del cáncer (órgano, tipo de tejido, etc.); *b)* tipo de terapia; *c)* detalles del paciente, y *d)* conceptos variados.

Las dos primeras se subdividen en varios niveles jerárquicos a través de menús sucesivos. Cuando se llega al nivel específico deseado, el sistema proporciona los términos del *Medical Subject Headings* que corresponden a los términos introducidos por el usuario. Si el usuario entra más de un término por faceta, el sistema usa el operador OR, y entre facetas distintas aplica el operador AND. Este mismo instrumento se ha aplicado también a la base de datos completa, partiendo de las 15 categorías de la clasificación de *Medline* y desarrollando niveles jerárquicos a partir de este punto.

La situación actual está caracterizada por la coexistencia de diferentes ciclos desde que se origina la información hasta su uso final, las clásicas figuras del autor, editor, productor, distribuidor, intermediario y usuario final ya no mantienen su independencia. Alan Gilchrist en su intervención en las *V Jornades Catalanes de Documentació* y sobre la calidad en la recuperación de la información, contrapuso esta situación a la rigidez de las normas que controlan los procesos que intervienen en la industria farmacéutica[7]. El objetivo en este camino hacia la calidad es la simplificación sin olvidar los recursos que ayudan a mejorar el rendimiento de las fuentes de información.

## BIBLIOGRAFÍA


1. Williams ME, Smith LC. New database products. Science, technology, and medicine (Issue 6). Online & CD-ROM Review 1995; 19: 211-218.
2. Winship IR. World Wide Web searching tools: an evaluation. Vine 1995; 99: 49-54.
3. Williams ME. The Internet: implications for the information industry and database providers. Online & CD-ROM Review 1994; 18: 149-156.
4. Dialog Information Services. Target on Dialog. «How-to» guide. Palo Alto, CA: Dialog, 1993.
5. Meadow CT, Marchionini G. Speculations on the measurement and use of user characteristics in information retrieval experimentation. Can J Inf Libr Sci 1994; 19: 1-22.
6. Hartley RJ, Keen EM, Large JA, Tedd LA. Online searching: principles and practice. Londres: Bowker-Saur, 1990.
7. Gilchrist A. Who is responsible for information quality in the information society? En: V Jornades Catalanes de Documentació, 1995 Oct 25-27. Barcelona: Les Jornades, 1995; 19-32.


## DISCUSIÓN

M. PORTA: Anteriormente se ha comentado —a mi juicio con acierto— que en la actualidad accedemos mejor a la información, pero que quizá no se ha producido un aumento proporcional en la calidad de la misma. Los investigadores nos quejamos de la falta de tiempo pues, a pesar de la aparición de estrategias e instrumentos mejores de búsqueda bibliográfica, al final nada puede sustituir al estudio y al análisis «artesanal». Uno de los instrumentos más importantes para cualquier artesano de la investigación es leer, y precisamente de lo que no hay mucho tiempo es de leer. Esto plantea entonces la cuestión de los *usos* que realmente hacemos de los nuevos instrumentos bibliográficos y bibliométricos. Como no se lee, o como lo que se lee no siempre se entiende, existe la tendencia a emplear

números como sustitutos del pensar. Pero estos números, que resumen un supuesto *impacto* de alguien, no solventan el problema de un tribunal de evaluación, de cómo valorar la producción de una institución, un departamento o un científico. El problema es de difícil solución, dado que nadie puede ofrecernos un año con un mes más para leer lo que las nuevas técnicas nos permiten encontrar. Quisiera llamar también la atención sobre la costumbre de construir mitos, como el del *Institute for Scientific Informacion* (ISI, Inc.), que estos días está terminando de elaborar sus páginas en el *World Wide Web* y en las que, entre otros productos, presenta el ranking de los *top cited scientists*, los *hottest papers* en distintas especialidades, etc. En una de esas listas más actuales el autor que ocupa el primer lugar es Robert Gallo, por ejemplo. ISI procesa aproximadamente 12 millones de citas al año y, aunque creo que nadie ha evaluado a fondo la metodología que utiliza, es obvio para cualquiera que conozca una determinada especialidad que su base de datos contiene múltiples errores. Nunca han explicitado, por ejemplo, cómo escogen las revistas fuente o *source*. Quisiera pues propugnar la necesidad de una valoración («desde el interior» de las disciplinas) de las metodologías y criterios que aplica el ISI «desde el exterior». Aunque de una manera global se puede considerar que la labor de dicha empresa es positiva, y que intervenciones como la de Espelt y la de Guardiola son de gran interés, todavía no logro saber exactamente cómo vamos a hacer frente al enorme reto que representa esta explosión de información, a veces de calidad indefinida.

C. ESPELT: Evidentemente el reto de la calidad es un reto importante y con una orientación diferente si se considera desde el punto de vista del usuario final o desde la óptica de los que actuamos de puente entre los recursos y la manera de ponerlos a disposición de este usuario. Aunque la reflexión que ha planteado sobre la necesidad de leer los artículos y el problema del tiempo puede ser muy útil, cada uno debe elegir sus prioridades y su forma de trabajar, lo que condicionará el nivel de calidad de sus estudios. Desde el mundo bibliotecario y documental siempre se ha tenido muy claro que ISI era un hecho aparte. Se trata de una institución que en su momento aportó unos avances muy importantes y, autores como Garfield, son citados ahora prácticamente como referencia histórica. Hoy día el avance de la bibliometría se dirige en otra dirección y hacia otros aspectos de la valoración de la información. Este instrumento tiene aplicaciones muy específicas y en general puede ser útil en la actualidad de forma conjunta con otras bases de datos. Además de las limitaciones del ISI sobre la selección de las fuentes, existen problemas derivados de la compleja estructuración de sus bases de datos. El tema de la calidad es muy importante pero de difícil solución. Nuestro objetivo es disponer de la información que satisfaga las necesidades del usuario y esta satisfacción tiene un número concreto. Aun en el supuesto de disponer de más tiempo para leer, es preciso seleccionar un número determinado de referencias, que no siempre serán exactamente las mejores. A veces, un artículo de mala calidad puede ser aprovechable en ciertos aspectos para una determinada investigación.

S. ERILL: Entre los interrogadores de *Medline* existentes, me ha llamado la atención alguna de las ventajas de CD-Plus que se aproximaría a los sistemas que usted ha comentado. Por ejemplo, con la opción *focus* se seleccionan automáticamente un determinado número de publicaciones, posiblemente aquellas donde la palabra buscada aparece en el resumen en más de una ocasión. También automáticamente, con la opción *contexto* se obtienen una serie de palabras, que sin relación alfabética con la expresión original, pueden ser útiles por aproximarse al tema seleccionado. En mi opinión se trata de alternativas extraordinariamente útiles.

C. ESPELT: *Focus* es un sistema que trabaja aprovechando las ventajas que ofrece el lenguaje controlado. La vertiente *contexto* es un instrumento que se está utilizando desde hace ya casi 10 años y que permite orientar al usuario. Ofrece la posibilidad de hallar los términos que aparecen más a menudo juntamente con la palabra que previamente se había seleccionado. Aunque siempre se debe identificar la relación que existe entre estos términos y lo que se está buscando, evidentemente forma un instrumento de ayuda importante.