Evaluación del uso de los Encabezamientos de materia en el catálogo colectivo CIRBIC frente al uso de palabras de cualquier campo para la recuperación de la información

Elvira González Sereno
Inocencia Soria González

Unidad de Coordinación de Bibliotecas del CSIC

C/ Jorge Manrique, 27 28006 Madrid
Tlf: (91) 585 49 83
Fax: (91) 564 42 02

1. POSIBILIDADES DE BÚSQUEDA EN EL SISTEMA ALEPH

El catálogo colectivo CIRBIC-Libros, contiene actualmente algo más de medio millón de registros bibliográficos pertenecientes a las más de 80 bibliotecas especializadas en todas las ramas del saber que componen la red de bibliotecas del CSIC.
El sistema ALEPH, utilizado para la automatización de estas bibliotecas es un sistema integrado con módulos de catalogación, búsqueda, préstamo, adquisiciones, mantenimiento, control de suscripciones y distintas utilidades.
El módulo de búsqueda permite la consulta de los catálogos colectivos CIRBIC de forma guiada, libre o acudiendo a los índices de los ficheros de acceso. ALEPH ofrece la posibilidad de utilizar mecanismos de truncado, adyacencia, etc. y operadores booleanos para afinar y mejorar los resultados de una búsqueda.

Los ficheros actualmente accesibles en el catálogo CIRBIC-Libros son de tres tipos:

a) Ficheros de autoridades: son los ficheros de Autores (AU), Títulos (TL) y Materias (MT). Estos ficheros se generan a partir de los encabezamientos asignados por las distintas bibliotecas de la red a sus documentos, y son la base de las entradas de autoridad creadas y mantenidas por la Unidad de Coordinación de Bibliotecas del CSIC. Cada una de éstas poseen relaciones y notas de uso o alcance que orientan tanto al bibliotecario como al lector en sus consultas.
El fichero de acceso de encabezamientos de materia, construido con el sistema de relaciones y notas propio de los tesauros está formado actualmente por cerca de 40.000 entradas, de las cuales 21.000 son términos admitidos y 17.200 son referencias que conducen de un encabezamiento no admitido a otro admitido.

b) Ficheros índices, que permiten el acceso directo a los documentos ligados a ellos, son los de la CDU (CDU), ISBN (BN), y Depósito legal (DP).

c) Ficheros de palabras, generados automáticamente a partir de los distintos campos de los registros bibliográficos: fichero de palabras de cualquier campo del registro (PA), fichero de palabras de los encabezamientos de autor (PAU), fichero de palabras de los títulos y series (PTL), y fichero de palabras de los encabezamientos de materia (PMT).

2. OBJETIVOS Y METODOLOGIA

El objeto de nuestro estudio es evaluar la recuperación temática de documentos en el catálogo CIRBIC-Libros partiendo de la búsqueda en los ficheros de palabras PA (no controlado por generarse de cualquier campo bibliográfico), y PTL y PMT (controlados ya que provienen de encabezamientos normalizados: títulos y series y encabezamientos de materia).

Para partir de peticiones de información reales se envió a 12 bibliotecas de la red (6 de Humanidades y Ciencias Sociales y 6 de Ciencia y Tecnología) un pequeño formulario en el que se pedía a los bibliotecarios que recogieran durante 4 días los temas objeto de demanda de información.

Una vez recogidos éstos, procedimos a efectuar las búsquedas y descartamos en una primera aproximación aquellas que proporcionaban o muchos o poquísimos documentos y resultaban por tanto o poco manejables o no muy significativas para la evaluación. Finalmente fueron seleccionadas 25 búsquedas, 16 de ellas pertenecientes a las áreas de ciencia y tecnología y 9 a las de humanidades y ciencias sociales.
Para consultar tanto en el fichero de palabras pertenecientes a cualquier campo del registro (PA) como en el de palabras del título (PTL) hay que tener lógicamente en cuenta la lengua en que están escritos los documentos, por el contrario al efectuar las búsquedas en el fichero de palabras generado por los encabezamientos de materia, sólo es necesario utilizar el español, lengua del catálogo.
Considerando que, según un estudio estadístico realizado en enero del 96, el 75% de las monografías que

forman parte del catálogo CIRBIC-Libros están escritas en español o en inglés, hemos introducido para la consulta en el fichero general de palabras y en el de palabras del título los términos truncados en ambas lenguas cuando lo hemos considerado necesario. En los dos ficheros hemos utilizado idénticas secuencias de palabras y utilizado los operadores booleanos AND, OR y NOT cuando el caso lo ha requerido.
Las búsquedas se han repetido en el fichero de palabras pertenecientes a los encabezamientos de materia únicamente en español. En algunos casos los términos utilizados en este fichero han sido idénticos a los de los ficheros de palabras pertenecientes a todos los campos y ficheros de palabras

	Palabras usadas en la búsqueda	
PREGUNTAS FORMULADAS	**En ficheros de palabras y palabras del título**	**En fichero de palabras de materias**
Absorción del sonido	**absor?, sonido?,sound?**	**absorción, sonido**
Angeles	**angel? not pau=angel?**	**angeles**
Animación cultural	**animacion?, animation?, cultural?**	**animacion socio-cultural**
Arquitectura barroca	**arquitect?, architect?, barroc?, baroqu?**	**arquitect?, barroc?**
Arseniuro de galio	**arseniuro?, arsenide?, galio, gallium**	**arseniuro, galio**
Ascensores	**lift?, lifts?, ascensor?, elevator?**	**ascensor?**
Cerezos	**cerez? not pau=cerez?, cherry not pau=cherry**	**cerez?**
Contaminación Mediterráneo	**mediterran?, contamina?, pollu?**	**mediterraneo, contaminacion**
Control del ruido	**ruido, noise?, control?**	**ruido?, control?**

Corriente alterna	**corrient?, current? altern?**	**corriente?, alterna?**
Corrosion del acero	**acer? steel?corros?**	**acero, corrosion**
Detectores	**detector?, sensorsensors?, sensores**	**detectores, sensores**
Educacion de adultos	**educa?, enseñan?, formac?, teach?, format?, adult?**	**educacion, enseñanza?, adulto?**
Fundición a presión	**fundici?, presion?die casting?**	**fundicion?, presion?**
Hidalguía	**hidalg? not pau=hidalg?, not ptl=quijote**	**hidalg?, hidalgos**
Mamíferos de Africa	**mamifer?, mammal?, africa?**	**mamiferos, africa**
Misticismo español	**mistic?, spain?, spani?, españ?**	**mistic?, españ?**
Mujeres de/en el Islam	**mujer?, wom!n, mahometan?, islam?, muslim?, musulman?, arab?**	**mujeres en el islam, mujeres musulmanas,mujeres-derecho islamico, mujeres arabes**
Orientacion escolar	**orientacion?, pedagogic?, escolar?, educativa?**	**orientacion, pedagogica**
Películas delgadas	**pelicul?, film?delgada?, thin?**	**pelicula?, delgada?**
Poliuretanos	**poliuret?, polyureth?**	**poliuret?**

Sordos	**sordo?, deaf?**	**sordo?**
Suelos de España	**suelo?, soil?, españ?, spain, spani?**	**suelo?, españa**
Transporte aéreo	**transport?, aereo?aeria?, air**	**transport?, aereo**
Vibración	**vibraci?, not espectr?vibrati? not spectr?**	**vibracion not espectro**

del título, en otros se han usado las palabras sin truncar conociendo previamente que formaban parte de un encabezamiento admitido en el tesauro. En los cuadros adjuntos aparecen reflejados los términos tal y como se consideraron usar en cada búsqueda concreta.

3. ANÁLISIS DE LOS RESULTADOS

Se han agrupado los encabezamientos elegidos, sin considerar su área temática, en cuatro clases dependiendo de su fórmula de expresión.

1. Un concepto expresado por una sola palabra (Angeles, Ascensores, Cerezos, Detectores, Hidalguía, Poliuretanos, Sordos y Vibración)

2. Un concepto expresado por más de una palabra (Animación cultural, Arquitectura barroca, Arseniuro de galio, Corriente alterna, Fundición a presión, Orientación escolar, Películas delgadas y Transporte aéreo).

3. Más de un concepto y por tanto expresados por más de una palabra (Absorción del sonido, Control del ruido, Corrosión del acero y Educación de adultos).

4. Conceptos limitados geográficamente (Contaminación del Mediterráneo, Misticismo español, Mamíferos de Africa, Mujeres en/de el Islam, Suelos de España)

1. Un concepto expresado por una sola palabra

El 87'72% del total de documentos recuperados incluían la palabra pedida en el título, el 34'51% la incluían en materias. El 10'99% de los registros tenían la palabra requerida únicamente en el campo de materias sin duplicar en el título.
La pertinencia media por palabras de cualquier campo en este primer grupo fue la más baja de los 4 grupos analizados: un 59'60%, debido a que en la mitad de los casos (Angeles, Cerezos, Hidalguía y Sordos) se produjeron homonimias con palabras de nombres de autores o palabras de títulos de obras literarias, pese a haber utilizado el operador "NOT PAU=palabra homónima" en el primer paso de la búsqueda para eludir el ruido tremendo que provocaban los nombres personales.
En el caso de "Hidalguía" se hizo necesario además añadir un segundo operador NOT antepuesto a la palabra "quijote" para que el sistema rechazase las más de 700 obras que de o sobre la obra de Cervantes aparecen en nuestro catálogo y que resultan claramente no pertinentes.
Aunque la búsqueda del mismo término "hidalg?" por palabras de materia (PMT) mejoraba los resultados, la pertinencia seguía siendo baja (42'8%). Encabezamientos del tipo "Dolores Hidalgo (Méjico)" o "Hidalgo, Miguel" enturbiaban los resultados.
Únicamente cuando se empleó en la pregunta el término completo sin truncar admitido como encabezamiento

de materia: "Hidalgos" se obtuvo la respuesta deseada.
"Angeles", con la escasísima pertinencia del 2'86% después de la limitación hecha por "NOT PAU=angel?", resultó el caso más llamativo.
Incluso cuando se analizan los registros recuperados por la palabra completa admitida como encabezamiento de materia la pertinencia sigue siendo baja (30'7%) ya que esta misma palabra forma parte de otros encabezamientos de materia correspondientes a la ciudad de Los Angeles, otros geográficos y algún nombre personal como materia.
La mejor forma y más rápida en este caso fue acudir al índice de encabezamientos de materia que proporcionó una relevancia máxima.
Las respuestas sobre Ascensores, Poliuretanos, Detectores, y Vibración resultaron muy pertinentes tanto por palabras como por materias. En el caso de "Vibración", al ser un tema muy general, se supuso de entrada que el usuario no buscaba los espectros de vibración que se eliminaron con el operador booleano NOT.
Los casos de "Poliuretanos" y "Ascensores" al ser palabras muy concretas con pocas posibilidades de confusión no presentaron dificultad alguna. En "Vibración" y "Detectores" el hecho de acudir al índice de encabezamientos de materia supuso una mejora considerable de la expresión de búsqueda ya que las relaciones semánticas establecidas bajo "Detectores" orientan sobre sus distintas denominaciones (sensores, contadores, biodetectores, etc.), y ayudan a situar "Vibración" en distintos contextos.

2. Un concepto expresado por más de una palabra

Del total de los documentos recuperados, un 64'67% tenían las dos palabras que se ponían como condición en el campo de título, un 60'05% del total cumplían esa condición en el campo de materia y un 29'23% la cumplían únicamente en el campo de materia y no en el de título.
El 85'67% de los registros recuperados por palabras de cualquier campo fueron pertinentes, de ellos el 18'93% no tenían completado el campo de materia.
El hecho de utilizar más de una palabra en la búsqueda cuando se expresa un único concepto no plantea más dificultad que cuando ese concepto viene expresado por una sola palabra ya que son palabras que suelen aparecer unidas en cualquier contexto.
Destaca en "Animación cultural" la baja recuperación que se consigue por palabras del título y que contrasta con una altísima pertinencia cuando la consulta se realiza por palabras de cualquier campo o por palabras de materia.
En el caso de "Arquitectura barroca", aunque hemos considerado pertinentes un 95% de los documentos recuperados por palabras de cualquier campo, nos cabe la duda de su relevancia con respecto al número total de documentos que relacionados con este tema pueda haber en el catálogo. En algunos casos las materias en las que se puede especificar el estilo artístico o utilizar una subdivisión cronológica en la práctica constituyen cuasisinónimos. Hemos comprobado que existe una tendencia en los catalogadores a asignar a este tipo de documentos un encabezamiento más genérico y otro más específico o a utilizar el término con subencabezamiento cronológico. Para una búsqueda más completa hubiera sido necesario buscar por el término más genérico Arquitectura-16..
En los conceptos del área de ciencia y tecnología las pertinencias menores se detectaron en "Corriente alterna" y "Transporte aéreo". En el primero se recuperaron tambien documentos relativos a motores y maquinaria de corriente alterna y en el segundo se recuperaron un 30'77% de registros no pertinentes relacionados con temas medioambientales que no respondían a la demanda.
Por último en "Arseniuro de galio", el bajo porcentaje encontrado por PTL, extraño al ser palabras tan concretas e inconfundibles, se debe al uso en muchos de ellos la abreviatura inglesa con que se denomina a este compuesto (Gaas). En materias se debe al uso de encabezamientos más genéricos que dificultan su recuperación.

3. Más de un concepto expresado por más de una palabra

Del total de documentos recuperados, un 40'18% tenían las palabras que se ponían como condición en el campo de título, un 66'39% tenían las dos palabras en el campo de materia y un 52'75% cumplían esa condición sólo en el campo de materia y no en el de título.
El resultado de la búsqueda por palabras de cualquier campo fue el más pertinente de los cuatro grupos analizados, el 95'37%. El 19'38% de los registros considerados pertinentes no tenían asignados encabezamientos de materia.
Este tipo de pregunta en la que ya es necesario introducir dos o más condiciones a la búsqueda, aumenta la dificultad ya que además no son necesariamente palabras que deban ir unidas para poder expresar el concepto que buscamos. Suelen ser además condiciones generales (Corrosión, Absorción, Educación, etc.) que se pueden aplicar a muchos otros conceptos y que generalmente en la redacción de los encabezamientos de materia van a dar lugar a estructuras de Encabezamiento-Subencabezamiento.

En "Educación de adultos" es donde se pone de manifiesto la necesidad de acudir al índice de encabezamientos de materia para descubrir que se usa por Educación popular, forma que aparece en muchos de los títulos.

4. Conceptos limitados geográficamente

El 49'46% de los documentos cumplían las condiciones exigidas en el título, el 60'36% las cumplían en el campo de materia y el 34'00% en el campo de materia pero no en el de título.
De los registros recuperados por palabras de cualquier campo resultaron pertinentes el 85'10%. Sin completar el campo de materias había un 15'67%.
El caso "Mamíferos de Africa" plantea el problema de delimitar si sólo buscamos obras generales sobre éstos o nos interesa tambien todos los animales de los distintos grupos que están bajo el gran grupo Mamíferos.
En "Suelos de España", cuando interrogamos por "suelo?", recuperamos tambien todo lo que hay sobre los suelos desde distintos puntos de vista del uso urbano, agrícola, etc. que no respondían a la demanda del usuario que se limitaba al punto de vista edafológico. Por el contrario se recuperaron numerosos documentos por PA y no por materias sobre Suelos de todas las zonas de España, por aparecer en todos los registros España. Ministerio de Agricultura, responsable de la obra.

5. Consideraciones globales

Del total de documentos recuperados, el 60'50% tenían los términos que se ponían como condición en el campo de título, el 55'32% en el campo de materia y el 31'74% cumplían esa condición en el campo de materia pero no en el de título.
Los resultados obtenidos en la búsqueda por palabras pertenecientes a todos los campos del registro muestran una pertinencia media del 81'43%.
De estos registros considerados como pertinentes el 22'20% del total eran registros incompletos que carecían del campo de materias.
Como puede observarse en los cuadros adjuntos, 15 de las consultas planteadas (60% del total) alcanzan una pertinencia en la búsqueda por palabras pertenecientes a cualquier campo del registro por encima del 80%; 7 de las restantes (28% del total) sitúan su pertinencia en más de un 60%; una (4% del total) en el 44% y 2 (8% del total) por debajo del 10%.
Como cabía esperar la mayor pertinencia se da en aquellos conceptos más concretos o expresados por palabras menos ambiguas que difícilmente se pueden preguntar de otra forma y que, sobre todo en el campo de las ciencias y tecnología forman parte del título.

4. CONCLUSIONES

1. Los ficheros de materia cuando tienen un mantenimiento regular y el usuario conoce bien su manejo y cómo moverse por el entramado de sus relaciones, constituyen una herramienta de gran utilidad en la recuperación bibliográfica, de hecho más de la cuarta parte de los documentos recuperados en estas búsquedas no habrían aparecido si hubieran carecido del campo de materias.
Es de destacar lo poco significativos que resultan los títulos en el campo de las humanidades en general y en casos concretos del área de ciencias cuando se trata de conceptos que pueden ser expresados por varios términos (Ej. Contaminación del Mediterráneo)

2. La recuperación por título se dificulta cuando se pide como condición que aparezcan en el campo de título dos palabras. En estos casos el número de registros recuperados por este campo se reduce considerablemente (51'43% frente a 87'7% cuando se exige sólo una palabra), en especial cuando una de las palabras que se ponen como condición es un término que se presta a la ambigüedad o que puede expresarse con otros sinónimos.

3. Algunos títulos de serie que contienen las palabras exigidas como condición crean bastante ruido en determinados casos (Ej. Serie Hidalguía, Serie Orientación escolar) mientras que en otros ayudan a la recuperación (Ej. Películas delgadas, que consigue agrupar documentos relativos a películas ferromagnéticas, metálicas, etc. considerados también pertinentes).

4. El sistema de relaciones que acompaña a los encabezamientos normalizados llega a ser fundamental en casos complejos. En la gran mayoría de los casos que constituyen esta muestra para responder a una consulta fue suficiente recurrir a un solo encabezamiento de materia, en otros tales como "Misticismo español" y "Mujeres en/de el Islam" podían encontrarse documentos pertinentes indizados con distintos encabezamientos: Mujeres en el Islam, Mujeres musulmanas, Mujeres árabes, Mujeres-Derecho islámico, Conducta sexual-Aspectos religiosos-Islam, Feminismo-Países musulmanes y algunos más.
En estos casos resulta particularmente útil consultar, dentro del fichero de autoridades de materia, el sistema de relaciones semánticas que vienen expresadas bajo cada encabezamiento y que le unen con otros y las notas que delimitan el uso y alcance de algunos de ellos.

5. Los nombres geográficos encierran algunas dificultades de recuperación especiales: En muchos títulos no aparece el

nombre de lugar, de modo que al incluirlo como condición de búsqueda se limita mucho la recuperación por este campo. Si a esto añadimos la tendencia que hemos detectados en algunos catalogadores a omitir los subencabezamientos de lugar, especialmente si ese lugar es España, la dificultad se vuelve a presentar en el campo de materias.

Cuando una de las condiciones exigidas en la consulta es un nombre de lugar con una extensión territorial amplia (país, continente...) cabe la posibilidad de que documentos que tratan del mismo tema pero limitados a una extensión territorial menor a la formulada en la búsqueda puedan ser también de interés para el lector.