EVALUACION DEL RENDIMIENTO DE TESAUROS ESPAÑOLES EN SISTEMAS DE RECUPERACION DE INFORMACION

B. Gil Urdiciain*

Resumen: El artículo describe un estudio experimental desarrollado en bases de datos españolas en el que se establece la comparación del rendimiento entre diferentes tesauros y el lenguaje natural, en el proceso de recuperación de la información. El test se llevó a cabo mediante la combinación de los métodos analítico y de muestreo y validación de datos. Las consultas se realizaron de modo interactivo, evaluándose los registros y modificándose las estrategias a la vista de los resultados. Las referencias recuperadas en cada modalidad se valoraron en base a las tasas de precisión y exhaustividad. Los resultados muestran que en lenguaje libre se consiguió una precisión del 63,4 % y una exhaustividad del 59,5 %; con la ayuda del tesauro ambos índices mejoraron: en precisión se alcanzó un 86,8 % y en exhaustividad un 61,6 %. Se concluye que el lenguaje controlado consigue más bajos niveles de ruido que el libre al tiempo que puede llegar a aportar tan altos índices de exhaustividad como aquél; igualmente, a la vista de los resultados de la comparación se puede concluir que el lenguaje controlado neutraliza las deficiencias del libre y viceversa y, por tanto, ambos son complementarios.

Palabras clave: recuperación de información; tesauros; evaluación de tesauros; lenguaje libre vs. lenguaje controlado.

Abstract: The article describes an experimental test developed in Spanish databases. The test compares the performance between different thesauri and the natural language, in the process of information retrieval. The test was carried out with the combination of the analytic and sampling methods. The searches were made interactively, the records were evaluated, and the search strategies modified based on the parcial results. The references obtained with each kind of language were evaluated on the basis of the precision and recall ratios. The results point out that free text obtained a precision ratio of 63.4 % and a recall of 59.5 %; with the help of the thesauri both ratios improved: the precision reached a 86.8 % and the recall a 61.6 %. The conclusion is that the controlled vocabulary experiences lower levels of noise than the free text, at the same time it can develop high ratios of recall as the controlled one. Also, it is possible to conclude that the controlled vocabulary neutralizes the deficiences of the free text and vice versa, so both are complementary.

Key words: information retrieval; thesauri; performance of thesauri; free text vs. controlled vocabulary.

* Universidad Complutense. Madrid.
Recibido: 15-1-98.

1 Introducción

En teoría, si buscáramos exhaustivamente en toda la colección de documentos que compone una base de datos, llegaríamos a encontrar determinada información solicitada, pero con un desmesurado gasto de tiempo y esfuerzo. Para que la inversión de tiempo en dicha tarea se minimice, la información se indiza con objeto de señalar la existencia o no de documentos relevantes acerca de distintas materias, y se almacena siguiendo unos criterios de clasificación. Durante el proceso de indización se seleccionan los términos apropiados de un lenguaje documental para la representación del contenido del documento y se le asignan de acuerdo con las reglas establecidas. En el momento de la recuperación se siguen esos mismos criterios y, si las materias utilizadas para indizar y alimentar el sistema son coincidentes con las empleadas en el proceso de búsqueda, los documentos recuperados serán relevantes. De lo dicho se deduce la importancia que tienen en todo el proceso la calidad del lenguaje de clasificación o indización empleado, la calidad de la indización y, por último, la capacidad de seleccionar las materias adecuadas para formular una ecuación de búsqueda. Lógicamente, la persona que va a interrogar al sistema necesita conocer lo mejor posible la naturaleza del lenguaje documental utilizado para el proceso de normalización de la información, así como el grado de detalle con que se analizaron los documentos. En base a su conocimiento podrá formular preguntas con un nivel de profundidad o generalidad adecuados, aumentando la probabilidad de recuperar información relevante. Aún conociendo estas variables, raramente coinciden los descriptores, palabras clave, etc., utilizados en la indización de un documento y en una pregunta de recuperación, dándose una mayor tasa de coincidencia, como es natural, cuando se trata de cuestiones genéricas.

Como alternativa al lenguaje documental se puede recurrir al vocabulario empleado en el propio documento para realizar dicha representación; en tal caso, se hablará de indización y recuperación en lenguaje libre. Entre las ventajas que, a priori, se pueden reconocer al uso de un lenguaje controlado, está su capacidad para adaptarse con facilidad a necesidades de información más o menos específicas del usuario potencial. Con el test que describimos en este trabajo tratamos de contrastar sus ventajas y desventajas, por otra parte bien conocidas, en comparación con el uso del lenguaje no controlado, en los Sistemas de Recuperación de Información (SRI) españoles, tratando con necesidades de información reales.

Como decíamos anteriormente, existen muchos elementos que influyen en la efectividad del proceso de recuperación de información, pero compartimos con Blair la idea de que *el más importante aspecto [...] concierne a cómo están representados los documentos en un sistema. El más rápido ordenador o las más sofisticadas técnicas de búsqueda no pueden superar los problemas derivados de una pobre representación [de naturaleza lingüística] de los documentos* (1).

2 Antecedentes

Desde los años cincuenta se han venido realizando evaluaciones de los lenguajes utilizados en los SRI para la indización y la recuperación de documentos. Las características de las mismas, su nivel de profundidad y aspectos analizados, varían de unas

a otras, pero todas, en una u otra forma, han alimentado el continuo debate entre los profesionales de la documentación acerca de la conveniencia y eficacia de los lenguajes controlados y no controlados para realizar dichas operaciones.

Existe abundante literatura sobre la evaluación de estos sistemas. El primer test conocido fue el ASTIA-Uniterm test, realizado en 1953. La investigación se centraba básicamente en el análisis de los resultados, aunque no se tiene constancia del desarrollo del proyecto, a excepción de un breve trabajo publicado por GULL (2). Para dicho trabajo se indizaron 15.000 documentos por dos grupos diferentes de especialistas en la materia, que provenían de la Armed Services Technical Information Agency (ASTIA) y de Documentation, Inc. El primer grupo utilizó la lista de encabezamientos de materia creada por ASTIA, el segundo, el sistema Uniterm. Se seleccionaron 98 demandas de entre las que había recibido ASTIA en el curso de sus actividades. Del análisis de los resultados derivó una mayor capacidad y calidad de recuperación con el procedimiento de unitérminos que con el de encabezamientos de materia.

En 1954, Cleverdon y Thorne (3) realizaron un test sobre el sistema Uniterm, cuyos resultados fueron inconclusos, pero permitieron sentar las bases metodológicas de futuras investigaciones.

Algunos estudios realizados sobre evaluación de SRI a lo largo de los últimos años evaluaron solamente los lenguajes utilizados para indizar y recuperar los documentos; otros se centraron en el proceso mismo de indización, etc. La gran mayoría se trataba de estudios de laboratorio, minuciosamente preparados con objeto de controlar cualquier posible variable, pero a partir de los años 80 se manifestó un creciente interés por analizar necesidades de usuarios reales. Los experimentos de Blair y Maron (1984) (4), así como los de Saracevic, Kantor, Chamis y Trivison (1988) (5) son indicativos de esa tendencia. No es posible mencionar aquí todos los tests llevados a cabo hasta la fecha, pero nos parecen de obligada referencia los proyectos Cranfield I y Cranfield II, cuya metodología y resultados han servido de guía a investigaciones posteriores, incluida la que realizamos en este trabajo.

El desarrollo del proyecto *Cranfield I,* a cargo de Cleverdon (6), tuvo lugar en Gran Bretaña entre los años 1957 y 1962, auspiciado por la Association of Special Libraries and Information Bureaux (ASLIB), con la colaboración de la National Science Fundation (USA).

Se comparó el rendimiento que ofrecían cuatro sistemas de indización: se comprobaron la Clasificación Decimal Universal, un índice alfabético de materias, un sistema de clasificación facetado y el sistema Uniterm. Los resultados se midieron sobre la base de dos parámetros: *exhaustividad* y *precisión*, que habían sido expresados algunos años antes por Fairthorne en las frases: *ABNO: All-But-Not-Only* (alta exhaustividad) y *OBNA: Only-But-Not-All* (alta precisión) (7) que, a su vez, habían sido sugeridos por Perry, Kent y Berry (8) dos años antes. Fue Cleverdon, sin embargo, el primero en poner en práctica la evaluación de los lenguajes documentales utilizados para el tratamiento y recuperación en SRI, basándose fundamentalmente en estos valores.

La *exhaustividad* mide la capacidad del sistema para recuperar documentos útiles, mientras que la *precisión* mide la habilidad de rechazar material no relevante (9). Teniendo en cuenta estos índices, la flexibilidad de un sistema se podrá valorar en función de su capacidad para adaptarse a las necesidades de exhaustividad y precisión de los usuarios.

La *precisión* se halla dividiendo el número de documentos relevantes recuperados entre el número de documentos recuperados. La *exhaustividad* se calcula dividiendo el número de documentos relevantes recuperados entre el número de documentos relevantes existentes en la colección.

Los resultados demostraron que los cuatro lenguajes conseguían prácticamente los mismos niveles de rendimiento a la hora de recuperar los documentos fuente. El porcentaje de recuperación fue el siguiente:

Unitérminos	82%
Encabezamientos de materia	81,5%
Clasificación Decimal Universal	75,6%
Clasificación facetada	73,8%

La diferencia es poca, como muestran las cifras, con excepción del sistema facetado, que dio más baja tasa de exhaustividad que los otros tres. El bajo rendimiento de este lenguaje se atribuyó a la rigidez del orden fijo de combinación de términos que le caracteriza. En otras palabras, Cranfield no descubrió diferencias significativas en la eficacia de los lenguajes de indización estudiados, un resultado que ha sido reiteradamente contrastado en estudios comparativos posteriores.

En cualquier caso, la importancia del primer proyecto Cranfield no estriba en los resultados obtenidos, tanto como en el método utilizado, el primero en su género, que sirvió de modelo a los tests realizados posteriormente.

El segundo proyecto Cranfield (1963-1966), tenía como principal objetivo investigar los componentes de los lenguajes de indización y los efectos que esos diversos componentes tenían sobre el rendimiento de los SRI.

El experimento se desarrolló de muy distinta forma al primero. Para llevarlo a cabo se construyeron 33 tipos diferentes de lenguajes de indización variando terminologías y estructuras. Cada lenguaje variaba igualmente tanto en el uso de términos simples y compuestos como en la incorporación de jerarquías. En todos ellos se controlaron sinónimos y homógrafos. Asimismo, se tuvieron en cuenta los índices de exhaustividad y precisión, tratando de averiguar qué lenguaje de indización aumentaba o disminuía esos parámetros. Los resultados del proyecto se pueden resumir en un mejor rendimiento de los lenguajes formados por términos simples, de los que se habían eliminado los casos de sinonimia y todos aquellos términos que generaban ambigüedad. Otra conclusión que pudo extraerse del test fue que la simple coordinación de términos resultó ser el recurso más efectivo para aumentar la precisión, independientemente de que dichos términos formaran parte o no de un lenguaje controlado. Este resultado, que parecía echar por tierra la necesidad de utilizar los lenguajes documentales, fue ampliamente rebatido por el hecho de haberse dado en una experimentación artificial, en la que se mantenían bajo control variables que, en necesidades de información reales, son difícilmente controlables.

A la vista de los resultados se pudo mantener la hipótesis de que existía una relación inversa entre exhaustividad (número de documentos que encontramos cuando realizamos una búsqueda) y precisión (probabilidad de que correspondan a una necesidad de información). Dicha hipótesis explica que, a medida que se amplía el campo de búsqueda, se encuentra más cantidad de información.

La confirmación de esa proporción inversa implicaría que es inútil cualquier in-

tento de maximizar ambos índices en un mismo sistema de información. Foskett sostiene, sin embargo, que *se pueden dar casos en los que, al examinar los resultados de búsquedas individuales, encontremos que se ha conseguido un 100 % de exhaustividad con un 100 % de precisión o, por el contrario, puede haber casos en los que ambos dan como resultado cero* (10). Van Slype también considera posible que ambos criterios mejoren simultáneamente en una misma consulta, siempre que *se pongan en funcionamiento nuevos medios (cualificación de los documentalistas, calidad del lenguaje documental y sofisticación del programa informático de búsqueda)* (11).

3 Metodología

En el presente trabajo experimental, a diferencia de los mencionados proyectos, no se analizan los resultados de la indización para introducir datos en el sistema con uno u otro lenguaje, sino que se establece la comparación entre la capacidad de recuperación de dos diferentes lenguajes: libre, basado fundamentalmente en términos simples o unitérminos, y los tesauros utilizados en los seis SRI consultados.

Teníamos que elegir entre realizar un experimento de evaluación artificial, controlando a priori todas las posibles variables: niveles de indización, tipología de los documentos, cantidad y calidad de los documentos, recursos automáticos de búsqueda y lenguajes documentales, o bien, basar nuestra evaluación en necesidades de información reales, para lo cual contábamos con serias limitaciones. La primera, y fundamental, la constituía el hecho de no contar con un equipo de personas para realizar el análisis de un número representativo de documentos, dado que se trataba de un trabajo individual. Optamos por analizar el comportamiento de seis tesauros frente al lenguaje libre, en la recuperación de información en bases de datos reales. Las bases elegidas para este propósito también fueron seis, de las cuales tres son del área de las ciencias sociales y otras tres de ciencia y tecnología. Así pues, el objetivo de la investigación se concreta en el examen del rendimiento de tales tesauros, desde la perspectiva del usuario, en el contexto de las bases de datos en las que se utilizan para el tratamiento documental.

El escaso número de bases de datos con tesauro propio, unido a dificultades derivadas de la ausencia de automatización y a traslados de sede de algunos centros, en el momento de realizar este trabajo, hicieron difícil la elección de las bases que precisábamos para hacer el test. Tras comprobar las prestaciones de unas y otras, optamos por las seis, que citamos a continuación, las cuales, si bien no son totalmente homogéneas, reúnen las condiciones necesarias, es decir, todas ellas son bases de datos bibliográficas que cuentan con tesauro propio y son de acceso directo. El volumen de documentos que registran difiere sustancialmente:

— Base de datos ICYT de Biología animal	11.820
— Base de datos PIE (Programa de Investigación y Desarrollo Electrotécnico)	1.360
— Base de datos Biblioma, del Ministerio de Medio Ambiente	26.362
— Base de datos del Centro de Documentación de la Mujer	12.600
— Base de datos PSEDISOC del CSIC	18.089
— Base de datos de la Biblioteca del Instituto de Migraciones y S. Sociales	21.392

Pero consideramos que esta particularidad, en vez de ser un inconveniente, constituye un valor añadido para la experimentación, ya que permite valorar el rendimiento de tesauros destinados a tratar fondos documentales de diverso tamaño.

Para realizar el análisis comparativo preparamos, en principio, un cuestionario de posibles preguntas a formular en cada base de datos, para lo cual fue imprescindible la lectura de obras y revistas especializadas en las materias a consultar, dada su heterogeneidad y grado de especialización.

Teniendo en cuenta que las preguntas debía realizarlas una sola persona, se plantearon desde un primer momento a un nivel de amplitud suficiente como para poder conseguir resultados válidos para la investigación y, al propio tiempo, lo más específicos que fuera posible, puesto que para poder validar los resultados, iba a ser necesario leer cada una de las referencias derivadas de las búsquedas. En algunos casos, incluso, se hizo necesario consultar los documentos completos al no ser posible hacer una valoración de su relevancia con el simple análisis de título y descriptores (en aquellas bases en las que el registro no contiene campo de resumen). El conocimiento previo del nivel de indización media que realizan las distintas bases también nos sirvió para plantear estrategias adecuadas a nuestros fines.

Se procedió, en primer lugar, a formular una ecuación de búsqueda en lenguaje natural, realizándose una segunda estrategia en lenguaje controlado. Inmediatamente después, se evaluó la relevancia de cada documento recuperado con uno y otro tipo de lenguaje.

Cuando se recuperó en lenguaje libre, el sistema trató de localizar la información en todos los campos del registro (salvo los que se componen de palabras vacías o forman parte de claves), por lo que tomó también los unitérminos que componen los descriptores. La búsqueda en esta modalidad se realizó, pues, consultando los campos de título, título original, descriptores principales y secundarios, topónimos e identificadores, resumen e índices, comunes todos ellos a la gran mayoría de las bases de datos. Las palabras clave compuestas se hallaron con un operador de adyacencia (para lo cual tenían que estar unidas en el texto). No todos los centros admitían el uso de palabras clave compuestas para la preparación de las ecuaciones, de modo que se hizo necesario recurrir a la postcoordinación en algunas de ellas. Lógicamente las palabras clave utilizadas para recuperar con lenguaje natural no siempre fueron coincidentes con los descriptores componentes de los tesauros y, en ocasiones, lo fueron sólo parcialmente. En cuanto al campo de título, las palabras clave que conformaban las ecuaciones de búsqueda aparecían a veces en el mismo, si bien la búsqueda en este campo presenta el problema de que se pierde la información existente en títulos de idioma extranjero, por lo que su eficacia es relativa y está condicionada a la lengua en la que está escrito el documento. El campo de resumen permite una mayor efectividad en la búsqueda en lenguaje libre, lo que, a su vez, produce, como indicábamos anteriormente, un detrimento en lo que a precisión informativa se refiere.

La recuperación en lenguaje controlado se realizó consultando los campos de descriptores. Los tesauros que cubren las bases sugieren términos relevantes para la búsqueda, de modo que la preparación de las ecuaciones con este procedimiento resultó más sencilla; no hacía necesario pensar en posibles términos sinónimos o polisémicos para delimitar los conceptos. La última fase del proceso consistió en imprimir o en exportar a disquete cada una de las referencias resultantes para proceder a su posterior análisis.

El procedimiento seguido para la valoración de los resultados fue el propuesto por Cleverdon, teniendo en cuenta los índices de precisión y exhaustividad. En respuesta a cada cuestión se recuperaron documentos con un índice de relevancia diverso; algunos fueron totalmente relevantes, otros respondían de forma dudosa a la pregunta formulada, y otros tenían una correspondencia marginal con dicha pregunta. Para el análisis se tomaron en cuenta sólo aquellos totalmente relevantes. Como nuestro grado de conocimiento de las diferentes materias no era homogéneo, solicitamos el consejo de especialistas en el contenido temático de las distintas bases de datos cuando nos surgieron dudas al establecer los criterios de relevancia, de tal manera que al nuestro se sumara el del potencial usuario conocedor de la materia. En cualquier caso, tratamos de valorar los documentos con la mayor objetividad posible, conocedores de que la medida de la relevancia es una estimación subjetiva que depende del nivel de conocimientos de cada individuo. *El problema* —escribe Foskett al respecto— *reside en el hecho de que los lectores buscan información partiendo de su nivel de conocimiento —su marco de referencia— con el mínimo esfuerzo, mientras que los autores presentan la información bajo su propio marco de referencia; cada uno de nosotros tiene su marco propio, por lo que la coincidencia nunca puede ser exacta. Podemos diseñar nuestros sistemas de recuperación de información para optimizar la probabilidad de coincidencia entre las preguntas de los usuarios con las respuestas que consiguen, pero aceptando de hecho que nunca será perfecta* (12).

4 Desarrollo de las búsquedas

Antes de comenzar cada consulta fue necesario aprender el lenguaje específico de cada sistema. Si bien es cierto que a pesar de las diferencias entre unos y otros, tienen todos rasgos comunes que permiten adaptarse fácilmente a su manejo, ello implica un esfuerzo más a realizar en el proceso. A este respecto escribe Amat: *Un inconveniente en relación con la figura del distribuidor de bases de datos radica en el hecho de que cada uno tiene que desarrollar su propio programa de recuperación de información o lenguaje de interrogación. Este hecho provoca que no se pueda trabajar con un único lenguaje de interrogación, sino que se han de conocer los diferentes lenguajes de los distribuidores con los que se trabaja* (13).

El manejo de las bases en CD-ROM fue muy sencillo, ofreciendo la posibilidad de trabajar con autonomía. En los SRI en los que la consulta se hizo en línea, fue necesaria la ayuda del documentalista o especialista en materia de recuperación; no obstante, estuvimos presentes durante el desarrollo de todas las búsquedas, en una u otra modalidad, puesto que, con frecuencia, se hacía necesaria la modificación de las estrategias. A veces, la combinación de conceptos daba como resultado una ausencia total de referencias, obligándonos a renunciar a los aspectos menos importantes de la cuestión, o a ampliar el alcance de algún concepto.

Para la preparación de las ecuaciones se analizó el área temática de las preguntas en detalle, de forma que para construir cada ecuación se tuvieran en cuenta los conceptos y relaciones que pudieran sugerir alternativas flexibles para preparar la búsqueda utilizando la capacidad de interacción de los SRI. No sólo se desarrollaron las estrategias iniciales sino que se fueron evaluando las referencias encontradas. Seguimos con este procedimiento la técnica sugerida por Van Rijsbergen (14) y Croft (15) con-

sistente en partir de un documento considerado a priori relevante en relación con la pregunta formulada y tratar de localizar otros similares basándonos en él.

Nuestro desconocimiento de las materias tratadas en algunas bases de datos también nos hizo recurrir a la alternativa denominada *pearl growing* para preparar estrategias, que consiste en analizar un documento (completo o título y resumen) conocido como representativo o de importancia (relevancia) en el área de conocimiento deseada y, en base al vocabulario que emplea, preparar la estrategia (16). Estas alternativas fueron posibles gracias a la naturaleza heurística de la recuperación en línea, que permitió ir modificando las estrategias a la vista de los resultados que íbamos obteniendo. En muchos casos, el examen de las referencias recuperadas mostró nuevos términos que pudimos incluir en nuevas ecuaciones.

A pesar de que utilizamos estos recursos, que nos fueron de gran utilidad, al tratar con algunos temas era difícil darnos cuenta de cualquier variación posible en la denominación de los conceptos presentes en las preguntas, de modo que para resolver este inconveniente solicitamos la ayuda de expertos en las materias. En el campo de las ciencias sociales fueron más frecuentes las dudas que en el área de ciencia y tecnología. En aquel campo se dieron en ocasiones falsas interpretaciones de las preguntas y surgieron dificultades para combinar términos; este último inconveniente se resolvió, a menudo, con el uso de palabras clave o descriptores precoordinados. En algunos supuestos, la inclusión de los conceptos de la pregunta conectados con el operador [*Y*] dio un resultado preciso de la información requerida. En ciertas consultas se amplió la búsqueda simplemente omitiendo uno de los conceptos de partida.

Nos queda, por último, señalar que en ningún caso se limitaron los registros a determinado tipo de documentos, sino que se tuvieron en consideración las referencias derivadas de monografías, artículos, actas de congresos, etc. Tampoco se tuvo en cuenta la fecha de publicación (antigüedad) de los documentos, dada la heterogeneidad de las bases. Se trató de encontrar, en todo momento, la totalidad de la información disponible sobre la materia, no sólo las referencias más pertinentes: para ello se formularon las cuestiones, por lo común, partiendo de lo general a lo particular, hasta llegar al más reducido o específico nivel posible sin correr riesgos de perder información relevante.

4.1 Aspectos y consideraciones para la preparación de las estrategias de búsqueda

A continuación describiremos aquellos aspectos que tuvimos que considerar en la preparación de las estrategias de búsqueda. Estos elementos hacen referencia al tipo de términos utilizados, niveles de concreción, problemas de sinonimia y polisemia, operadores utilizados, problemática en el uso del lenguaje libre y otras variables que hubimos de tener en cuenta a fin de diseñar una óptima estrategia de búsqueda. Dichos aspectos pueden ser concretados en los siguientes puntos:

1) Se analizaron los temas generales concernientes a cada pregunta en sus facetas constituyentes, así como los términos sinónimos y cuasisinónimos de cada materia. Se ordenaron y agruparon en conjuntos y mediante el álgebra de Boole —utilizada en todas las bases—, se precisaron los temas hasta conseguir las referencias que respondían al perfil de búsqueda solicitado. Se visua-

lizaron en pantalla las muestras obtenidas a fin de comprobar que los documentos se centraban en la materia deseada. Según la respuesta obtenida, se ampliaba o reducía la búsqueda mediante diversos recursos.

2) Con objeto de conseguir en la búsqueda inicial un alto nivel de exhaustividad, se incluyó más de una palabra para describir cada concepto. Si el número de registros resultante indicaba que alguno de los términos empleados resultaba demasiado general, se procedía a realizar una nueva combinación de términos, omitiendo alguno de los más generales, obteniendo así un nivel adecuado de referencias.
3) Algunas bases de datos ofrecían la posibilidad de utilizar términos genéricos incluidos en los tesauros que facilitaban la amplitud de las búsquedas. Su empleo supuso un significativo ahorro de tiempo y esfuerzo, puesto que hizo innecesaria la localización de terminología precisa, al tiempo que simplificó la construcción de la estrategia. Al elegir estos descriptores genéricos se evitó el uso de palabras que describían conceptos demasiado imprecisos que podían generar un alto grado de ambigüedad, tales como *sistema, medida, problema,* etc.
4) A fin de alcanzar un alto nivel de precisión, se recurrió al uso de términos específicos y, en su caso, a la combinación de los mismos. Como es lógico, los términos específicos tienen la particularidad de concentrarse en unos pocos documentos de la colección, lo que facilita su identificación. Así, por ejemplo, en la base de datos de Electrotecnia, el alto nivel de precisión concretó los términos utilizados en las estrategias en unos pocos documentos de la colección. Si tenemos en cuenta el limitado número de documentos que la componen, es razonable que en algunos casos las tasas de precisión y exhaustividad resultantes se aproximasen y que, incluso, fueran coincidentes.
5) En algunas consultas ocurrió que las preguntas incluían una gran variedad de conceptos diferentes, por lo que se trató de representarlos por medio de términos unidos por el operador [*Y*]. En tales casos se conseguía un alto nivel de precisión y, a veces, dio como resultado el silencio.
6) El operador de negación [*NO*], que permite combinar expresiones alternativas y agrupar términos representativos de un mismo concepto, restringiendo así el volumen de información recuperada, fue un recurso empleado en contadas ocasiones, ya que, aunque facilita las búsquedas en algunos casos, su uso implica el riesgo de perder información relevante. En su lugar, optamos por recurrir a los términos genéricos y relacionados de los tesauros o por utilizar expresiones alternativas en lenguaje libre.
7) Como consecuencia del uso de términos ambiguos, imprecisos o inapropiados, en ocasiones se dieron falsas combinaciones, dando como resultado la recuperación de referencias que no eran lo bastante precisas. Para eliminar estos inconvenientes se consultaron los términos de indización de alguna referencia relevante recuperada con anterioridad, lo que nos permitió contextualizar los términos de la pregunta y replantear la estrategia.
8) En algún caso se dio la circunstancia de que el lenguaje controlado no disponía de términos adecuados para describir los conceptos requeridos. Dado que nuestro estudio consistía en establecer una comparación entre el lenguaje libre y el controlado, optamos por incorporar términos más genéricos para que

el tesauro pudiera dar una respuesta satisfactoria. Sin embargo, y como cabía esperar, el resultado fue impreciso, ya que se alejaba de las necesidades de información que se precisaban en la pregunta.

9) Por último, cabe decir que cuando se utilizó el lenguaje libre se tuvieron en cuenta todas las posibles variaciones de un mismo concepto, así como el truncado de prefijo y sufijo. Asimismo, en su caso, se incluyeron las variaciones en singular y plural de cada palabra.

5 Criterios de evaluación

La evaluación de la cantidad y calidad de la información recuperada, se hizo teniendo en cuenta las tasas de efectividad, es decir, *precisión* y *exhaustividad.* Los resultados se presentan como porcentajes de dichos valores. Se consideraron relevantes las respuestas que correspondían adecuadamente al objetivo de la búsqueda.

En la práctica es difícil medir estos parámetros, debido a que la estimación de la relevancia es subjetiva. Hallar el índice de precisión genera pocos problemas, excepto cuando en una determinada búsqueda no se recupera ningún documento. La exhaustividad presenta muchas complicaciones, incluso cuando se trata de pequeñas colecciones, ya que hallar este índice, que es inevitablemente un valor relativo, requiere que cada documento de la colección sea contrastado en relación con cada consulta sobre un tema determinado, esto es, exige el conocimiento del número total de documentos relevantes de la colección con respecto a la pregunta. Ante la imposibilidad de determinar el volumen de documentos que tiene una base de datos sobre un tema en particular, diferentes especialistas han propuesto métodos alternativos que tratan de salvar la dificultad para calcular este valor.

Lancaster (17) utiliza en dos ocasiones un método de muestreo para estimar el índice de exhaustividad de una gran base de datos, es decir, midiendo la relevancia de un conjunto de documentos de la colección. Salton propone el mismo procedimiento y sostiene que la exhaustividad no es un valor exacto, sino una estimación del número total de documentos relevantes de la colección: *La valoración de la relevancia se hace en base a un subconjunto de documentos de la colección. Alternativamente, una consulta dada puede ser procesada por una variedad de diferentes métodos de búsqueda y recuperación, dando por supuesto que todos los documentos relevantes van a ser recuperados por medio de dichas búsquedas. Los resultados se combinan entonces en una única lista de resultados. La lista de documentos relevantes se obtiene mediante la valoración de la relevancia de esa lista de resultados* (18). Van Slype considera dos posibilidades para determinar la exhaustividad: *ya sea sistemáticamente, examinando las referencias una a una (lo que tiene el riesgo de durar mucho); o bien interrogando de nuevo el fondo con una serie de ecuaciones muy amplias (con pocos o ningún Y, con muchos O), incluso basándose en la clasificación* (19).

En el test que realizamos, hallamos las tasas de exhaustividad combinando las técnicas de muestreo, las de clasificación y la interrogación al sistema por los términos más genéricos de cada pregunta. Entre los documentos recuperados por medio de esta última alternativa, se encontraron muchos no relevantes, lógicamente, es

decir, se dio un altísimo índice de ruido documental. Se imprimieron, por ejemplo, todos los documentos en los que figurara el descriptor o palabra clave *insectos*, en la base de datos de Biología animal. En la estrategia realizada para hallar la tasa de precisión, el número de documentos que respondían a la cuestión: *Fisiología de la reproducción en los insectos*, fue de tan sólo 4, pero, al buscar por el término general *insectos*, tuvimos que analizar un total de 1.558 documentos. La proporción en este caso es exagerada, pero da una idea del propósito que perseguíamos y que se ha conseguido: no excluir del análisis ninguna información existente en la colección sobre cada tema.

El volumen de documentos extraídos mediante el último método mencionado nos hizo desistir, en un principio, de la tarea —en total se recuperaron 11.906 referencias—. Decidimos, finalmente, seguir adelante convencidos de que, aunque minucioso, era el único procedimiento que garantizaba resultados fiables. Tras el largísimo proceso de análisis teníamos la certeza de haber dado con toda la información relevante de las bases en relación con los temas propuestos.

Para hallar la exhaustividad mediante técnicas de muestreo, se procedió de la siguiente forma: se seleccionó en primer lugar la muestra generando un listado de números aleatorios por medio del generador de números aleatorios del compilador Pascal, cambiando la semilla para cada base de datos. Se hizo, a continuación, una estimación del tamaño de la muestra:

Para estimar el porcentaje de documentos relevantes definimos *Ai*, variable dicotómica, de forma que:

$$Ai < \begin{cases} 0 \text{ si el registro } i \text{ no es relevante} \\ 1 \text{ si el registro } i \text{ es relevante} \end{cases}$$

Por lo que el estimador del porcentaje será:

$$p = \frac{\Sigma Ai}{n}$$

Σ = suma.
p = porcentaje estimado de registros relevantes.
n = número de registros de la muestra.

Como se iba a realizar un muestreo aleatorio simple, sin reposición, la varianza del estimador de [p] sería:

$$V(p) = \frac{N-n}{N-1} \cdot p9$$

N = tamaño de la población
n = tamaño de la muestra
p = probabilidad de que Ai valga 1
q = probabilidad de que Ai valga 0

Por lo tanto el error estimado será:

$$E = k' \sqrt{\frac{N \cdot npq}{N \cdot 1n}}$$

Según Kish, *la definición básica del error estándar de la media es igual para cualquier diseño de muestra: es el error estándar de la distribución de muestreo para ese determinado diseño de muestra* (20).

Para obtener un tamaño de muestra significativa al 95 % y con un error no superior al 0,02 (diferencia entre el valor estimado y el valor real), despejamos [*n*]:

$$n = \frac{K'Npq}{E_1(N-1) + K_1pq}$$

Como desconocemos la varianza la sustituimos por 0,5 (1 – 0,5) = 0,25, que es el valor máximo que podría tomar.

El tamaño de la muestra de las cinco bases sometidas a este procedimiento fue el siguiente:

Biología animal	1.995 + 100 = 2.095
Medio ambiente	2.200 + 100 = 2.300
Centro Doc. Mujer	2.016 + 100 = 2.116
Psicología	2.119 + 100 = 2.219
Servicios sociales	2.158 + 100 = 2.258

La base de datos de Electrotecnia se analizó en su totalidad, dado el reducido volumen de sus fondos, por lo que no fue necesario proceder mediante este método para conocer la colección.

En todas las bases, como puede observarse, se incluyeron 100 documentos de reserva, en previsión de posibles bajas que hubieran sufrido las colecciones, o de cualquier otra circunstancia que alterara el volumen de la muestra.

6 Resultados obtenidos

Al proceder a la comparación de resultados se detectaron accidentes lingüísticos tales como polisemias, sinonimias, etc. Estas cuestiones lingüísticas explican, por sí solas, las diferencias encontradas entre ambos procedimientos de búsqueda.

En total, las 60 búsquedas en lenguaje libre recuperaron 1.321 documentos, de los cuales 838 fueron relevantes. El mismo número de búsquedas utilizando lenguaje controlado dio como resultado un total de 983 referencias, 848 de las cuales fueron relevantes. Los conjuntos de documentos recuperados variaron en función del nivel de especificidad de la información requerida (véase tabla I).

Estos datos confirman el alto porcentaje de ruido documental que genera el lenguaje natural en comparación con el controlado. La tasa de silencio, sin embargo, fue muy similar con ambos procedimientos de búsqueda; a lo largo del test, el total de

Tabla I
Resultados de las estrategias de búsqueda

	Documentos totales recuperados	*Documentos relevantes*	*%*	*Documentos no relevantes*	*%*
Lenguaje libre	1.321	838	63,4	483	36,6
Lenguaje controlado	983	848	86,2	135	13,7

documentos relevantes no recuperados con tesauro fue de 568; en libre la cifra ascendió a 573. La diferencia, de tan sólo 5 documentos, no es estadísticamente significativa. El silencio documental generado en la búsqueda con lenguaje controlado tuvo su origen en la ausencia de descriptores apropiados en casos muy concretos. Este fenómeno se dio, por ejemplo, en una de las consultas hechas en las la base de datos del Centro de Documentación de la Mujer, motivado por la falta de un término representativo de un concepto que ha cobrado significación en el entorno social en fecha posterior a la edición del tesauro. En la cuestión se solicitaba documentación acerca del *Empleo de mujeres inmigrantes*, tema sobre el que no se halló ninguna referencia con la ayuda del tesauro, pero que, con la introducción de la palabra clave *dominicanas*, dio como resultado un total de 9 referencias.

Las tasas de precisión y exhaustividad conseguidas mediante lenguaje libre y con tesauro se pueden comparar en la tabla II.

Tabla II
Tasas de precisión y exhaustividad obtenidas

	Lenguaje libre			*Tesauro*		
	Documentos relevantes recuperados	*Precisión*	*Exhaustividad*	*Documentos relevantes recuperados*	*Precisión*	*Exhaustividad*
Biología animal	142	87,6%	54,20%	145	96%	63,20%
Electrotecnia	59	44,90%	65,50%	63	78,60%	61,30%
Medio ambiente	133	64,40%	69,20%	121	93,80%	62,30%
Mujer	287	47,70%	55,40%	317	82,40%	58,40%
Psicología	70	63,70%	54,80%	85	81,70%	65,60%
Servicios sociales	147	72,50%	58,20%	117	88,70%	59,30%
Totales	838	63,40%	59,50%	848	86,80%	61,60%

Si hallamos la media de los seis sistemas de recuperación los resultados arrojan las siguientes cifras: en lenguaje libre se consiguió una precisión del 63,4 % y una exhaustividad del 59,5 %; con la ayuda del tesauro ambos índices mejoraron: la precisión alcanzó un 86,8 % y la exhaustividad un 61,6 %.

Las diferencias son del 23,4 % en precisión, favorables a la recuperación por medio de tesauro, y del 2,1 % en exhaustividad, también con la balanza a favor del lenguaje controlado. Como puede observarse, uno de los resultados más significativos fue que los dos lenguajes proporcionan prácticamente el mismo nivel de rendimiento en lo que a exhaustividad se refiere, siendo, sin embargo, las

diferencias registradas en las tasas de precisión estadísticamente significativas. Al comparar los valores de dichas tasas conseguidos en las tres bases de datos de ciencia y técnica con las de ciencias sociales hallamos diferencias destacables en lenguaje libre. Los valores de precisión en los dos tipos de bases alcanzan una diferencia del 13 %, y de un 20,5 % en exhaustividad, ambos favorables al campo de la técnica. Con el uso del lenguaje controlado la diferencia en lo que a precisión se refiere, en los dos tipos de bases de datos, alcanza el 15,6 %, siendo en exhaustividad del 3,5 %.

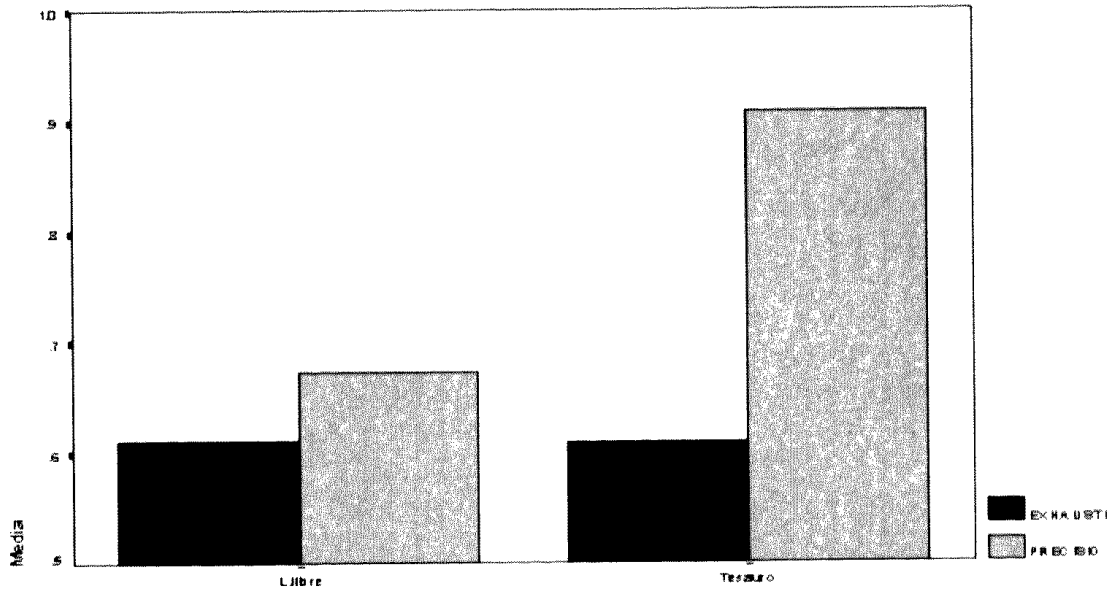
Los errores más comunes derivaron de las siguientes razones: a) inadecuada selección de términos de búsqueda; b) establecimiento de relaciones incorrectas entre los términos utilizados para preparar las ecuaciones; c) no haber tenido en cuenta la existencia de polisemia en algunas materias para nosotros del todo desconocidas; d) inclusión en las estrategias de algún término incorrecto; e) representar algún concepto de forma demasiado amplia.

Como consecuencia, se dieron casos de irrelevancia, debido a que se utilizaron términos de búsqueda que representaban conceptos que no aparecían reflejados en las referencias recuperadas o que lo estaban de forma marginal. Los fallos fueron menos al emplear tesauros en las estrategias, gracias a que el nivel de precoordinación de todos ellos fue adecuado haciendo innecesario el uso de recursos sintácticos extra para unir los términos precisos para la recuperación. Afortunadamente, fueron varios los casos en los que el cuidadoso repaso de las referencias permitió advertir fallos que no resultaban evidentes y que podrían haber modificado en alguna medida los resultados.

Mediante un diagrama de barras podemos observar de forma resumida y gráfica los valores conseguidos con ambos lenguajes.

Figura 1

Promedio de los índices de precisión y exhaustividad resultantes



FUENTE

Las columnas representan el promedio de cada índice en un tipo de lenguaje en particular, pudiendo observarse los siguientes aspectos:

Son patentes las diferencias existentes entre los valores recogidos de precisión y exhaustividad, siendo este último claramente inferior en ambos tipos de lenguajes. Concretamente, en libre la diferencia entre ambos valores es del 3,9 % y en controlado del 25,2 %. Aunque las cifras no lo reflejan, también es notablemente significativo que el índice de exhaustividad sea superior con el uso del tesauro, lo cual contradice la creencia generalizada de que en lenguaje libre esta tasa es siempre mayor a la conseguida bajo control del vocabulario. Si bien a lo largo de la experimentación hubo casos en los que el uso del lenguaje libre dio un mayor índice de exhaustividad que el tesauro, fueron tan pocos que sus resultados no son estadísticamente significativos. En resumen, el test tiende a mostrar que se obtienen mejores resultados en lenguaje controlado en lo que respecta a ambos índices.

7 Conclusiones

1. La baja tasa de equivalencia de algunos tesauros —Electrotecnia (0,1) y Medio ambiente (0,26)— pudo haber influido en las tasas de exhaustividad, ya que son las dos únicas bases en las que el lenguaje libre superó al controlado en lo que a dicho valor se refiere.
2. El uso de un vocabulario inadecuadamente específico produce resultados con bajo nivel de relevancia y afecta a la exhaustividad en aquellos casos en los que no hay términos para describir conceptos significativos. De hecho, la ausencia de descriptores genéricos en algunos tesauros parece haber sido la causa por la cual el índice de exhaustividad utilizando estos lenguajes fue inferior al conseguido con lenguaje libre.
3. Los resultados de la comparación de la efectividad de seis tesauros y el lenguaje natural en el que se presentan los documentos de seis bases de datos bibliográficas demuestran que el control del vocabulario es factor determinante del éxito en el proceso de recuperación de información para lograr altos niveles de precisión; el aporte de sinónimos y otros términos relacionados con los conceptos de las preguntas mejoraron este índice en un 23,4 % respecto al conseguido mediante lenguaje natural.
4. El tesauro es el lenguaje de recuperación por excelencia en bases de datos especializadas, superando al libre incluso en lo que a índices de exhaustividad se refiere. Este hecho se da de forma particular en bases de ciencia y tecnología, caracterizadas por una terminología muy consolidada.

Si el lenguaje documental es completo, es decir, si incluye todos los posibles términos referidos a un mismo concepto, se consigue un elevado nivel de exhaustividad, en cuyo caso hay poca propensión por parte del usuario a utilizar en la recuperación términos no recogidos en su corpus. También puede influir en el mencionado resultado el nivel de consistencia en la indización (la utilización de los mismos términos siempre en el proceso), más probable en bases que emplean descriptores genéricos. La variable indización no ha sido, sin embargo, tenida en cuenta en este estudio, de modo que sólo la apuntamos como hipotética causa de influencia en los resultados.

Así pues, en contra de la creencia generalizada de que el empleo del lenguaje libre aporta más altos niveles de exhaustividad que el de los tesauros en la recupera-

ción, con este lenguaje documental se consiguieron mejores tasas con el simple recurso de incorporar descriptores genéricos de partida y todos los específicos que comprenden. Paralelamente, en libre, la exhaustividad se incrementó usando todos los posibles sinónimos de los términos de búsqueda.

5. El lenguaje controlado consigue más bajos niveles de ruido que el libre, particularmente cuando los títulos de los documentos no son representativos de sus contenidos.

6. La tasa de silencio es muy similar con los procedimientos de búsqueda libre y controlada. El silencio documental generado en la búsqueda con tesauro tuvo su origen en la ausencia de descriptores apropiados en casos muy concretos.

7. El control de la sinonimia en búsquedas en lenguaje libre obliga al buscador a realizar malabarismos lingüísticos para encontrar todas las posibles acepciones de un mismo concepto. Esta situación no se da con el uso de un tesauro, en cuyo índice sistemático se encuentran todos los descriptores organizados por categoría semántica.

8. El uso del tesauro permite contextualizar los términos de búsqueda evitando falsas combinaciones que surgen en lenguaje libre, particularmente cuando se han de emplear términos polisémicos en el proceso.

9. La preparación de ecuaciones se simplifica —cuando se utiliza un tesauro— por su capacidad de incorporar a la búsqueda términos alternativos que concretan los temas propuestos en las cuestiones. Esa capacidad de inducción beneficia además la recuperación de documentos relevantes evitando un exceso de referencias inútiles cuyo examen lleva tiempo. Con lenguaje libre es el propio usuario quien ha de controlar el vocabulario cada vez que quiere recuperar información. De la experiencia aquí desarrollada se infiere que este hecho se da especialmente en bases de datos de ciencias sociales, en las que surgieron más casos de polisemia que en las bases técnicas.

10. El lenguaje libre ofrece la posibilidad de recuperar información muy especializada y actualizada, cuya terminología o no está incluida en el vocabulario controlado o no está representada de forma suficientemente específica.

11. Cuando las palabras del lenguaje natural coinciden con los descriptores o con los términos componentes de descriptores sintagmáticos, la recuperación con ambos procedimientos es equivalente, al igual que los resultados.

12. El análisis comparativo de uno y otro permite deducir que el lenguaje controlado neutraliza las deficiencias del lenguaje libre y viceversa, por lo que los dos sistemas no sólo no son antagonistas sino que se complementan uno al otro. La lista finita de descriptores de un tesauro no puede abarcar todos los conceptos de que se compone una disciplina y la inclusión de un campo adicional de palabras clave podría permitir la identificación de conceptos que no están recogidos en el corpus del tesauro. En muchas consultas no coinciden los documentos recuperados por medio de ambos lenguajes, siendo, sin embargo, en ambos casos documentos relevantes los conseguidos, de ahí la conveniencia de utilizar los dos lenguajes en combinación para no perder información. Ello pone de manifiesto la inconsistencia en la indización en que, en mayor o menor grado, caen los analistas y productores de datos.

13. Los índices de precisión resultantes con ambos lenguajes tienen diferencias, en las cuales no ha influido el tipo de base de datos consultada, es decir, no se ha observado variación sustancial en lo que a este valor se refiere, si se comparan los resultados obtenidos en bases de datos de ciencia y tecnología o bases de ciencias sociales y humanidades.

8 Referencias

1. BlAIR, D. C. *Language and representation in information retrieval.* Amsterdam, etc.: Elsevier, 1990, p. 155.
2. GULL, D. Seven years of work on the organisation of materials in the special library. *American Documentation,* 7 (1956), p. 320-329.
3. CLEVERDON, C. W.; THORNE, R.G. *A Brief Experiment with the Uniterm System of Coordinate Indexing for the Cataloging of Structural Data.* RAE Library Memorandum n. 7, AD 35004. Farnborough, England: Royal Aircraft Establishment, 1954.
4. BLAIR, D.; MARON, M. E. An evaluation of retrieval effectiveness for a full-text document retrieval system. *Commun. ACM,* 28 (3) (1984), p. 281-299.
5. SARACEVIC, T. et al. A study of information seeking and retrieving. I. Background and methodology; II. Users, questions and effectiveness; III. Searchers, searches and overlap. *Journal of the American Society for Information Science,* 39 (3) (1988), p. 161-216.
6. CLEVERDON, C. W. *Aslib Cranfield Research Project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems.* Cranfield: College of Aeronautics, 1962.
7. FAIRTHORNE, R. Automatic retrieval of recorded information. *Computer journal* (1958), p. 36-41.
8. PERRY, J.; KENT, A.; BERRY, M. *Machine literature searching.* New York, etc.: Interscience Publishers, 1956.
9. Un documento se juzga relevante cuando responde a una solicitud de información. Relevancia se emplea, a menudo, como término sinónimo de pertinencia, y como tal es considerado por algunos autores. Sin embargo, son mayoría los especialistas que establecen diferencias entre ellos. Lancaster, Foskett y Kemp, por ejemplo, hablan de relevancia cuando se refieren a la valoración que realizan una o varias personas en relación con una determinada solicitud de información, encontrando coincidencia entre pregunta y respuesta. Pertinencia sería la valoración que hace el usuario de una respuesta dada por un SRI a una necesidad concreta de información formulada por él mismo. Puede, como es natural, existir coincidencia en cuanto a las estimaciones de los dos índices, pero existe un sutil matiz diferenciador entre ambos: un documento considerado pertinente, no necesariamente puede ser valorado como útil o relevante por una persona ajena a la formulación de la cuestión.
10. FOSKETT, A. C. *The subject approach to information.* 5.ª ed. London: Clive Bingley, 1996, p. 85.
11. SLYPE, G. van. *Los lenguajes de indización: Concepción, construcción y utilización en los sistemas documentales.* Madrid: Fundación Germán Sánchez Ruipérez, 1991, p. 194.
12. FOSKETT, A. C. Op. cit., p. 15.
13. AMAT, N. *La documentación y sus tecnologías.* Madrid: Pirámide, 1994, p. 139.
14. RIJSBERGEN, C. J. van. *Information retrieval.* 2.ª ed. London: Butterword & Co., 1979.
15. CROFT, W. B. A model of cluster searching based on classification. *Information systems,* 1980, 5, p. 189-195.
16. HARTLEY, R. J. *Online searching: principles & practice.* London (etc.): Bowker-Saur, 1990, p. 171.
17. LANCASTER, F. W. *Evaluation of the Medlars demand search service.* Bethesda, Md.: National Library of Medicine, 1968; e *Information retrieval systems: characteristics, testing and evaluation.* 2.ª ed. New York: Willey, 1979.
18. SALTON, G.; McGILL, M. J. *Introduction to modern information retrieval.* New York (etc.): McGraw-Hill Publishing Company, 1983, p.166-167.
19. SLYPE, G. Van. Op. cit., p. 193.
20. KISH, L. Selección de la muestra. En *Los métodos de investigación en las Ciencias Sociales.* Barcelona: Paidós, 1953, reimp. 1987, p. 177.