ANÁLISIS DEL LENGUAJE CONTROLADO EN TRES BASES DE DATOS INTERNACIONALES

PURIFICACIÓN MOSCOSO [1]
ANA EXTREMEÑO [2]

Resumen: Se analiza el lenguaje controlado utilizado para la indización de tres bases de datos internacionales de Ciencias Sociales, utilizando los parámetros fundamentales que deben regir todo proceso de indización: la relevancia, la consistencia y la exhaustividad. Asimismo, se comparan los resultados obtenidos en el análisis de los lenguajes controlados post-coordinados y pre-coordinados de cara a la recuperación de información por materias. Estos resultados constatan la mayor idoneidad del uso de descriptores frente a encabezamientos de materia en los entornos automatizados.

Palabras clave: Bases de datos; Indización; Relevancia; Consistencia; Exhaustividad; Ciencias Sociales

Abstract: This paper analyses the indexing of three Social Sciencies International databases taken into consideration relevance, consistency and exhaustivity of the vocabulary used in the controlled languajes. Post-coordinated languaje and pre-coordinated languaje are compared, and results identify the coherence of the use of descriptors versus subject headings.

Keywords: Databases; Indexing; Relevance; Consistency; Exhaustivity; Social Sciences.

[1] Decana de la Facultad de Documentación. Universidad de Alcalá. C/ San Cirilo s/n. 28801 Alcalá de Henares (Madrid, España)
correo electrónico: p.moscoso@uah.es
[2] Profesora de la Facultad de Documentación. Universidad de Alcalá . C/ San Cirilo s/n. 28801 Alcalá de Henares (Madrid, España)
Correo electrónico: ana.extre@uah.es
Artículo recibido: 31-05-99. Aceptado: 26-08-99.
INFORMACIÓN, CULTURA Y SOCIEDAD. No. 2 (2000) p. 45-64

1. Introducción

En el proceso de recuperación de información electrónica son varios los aspectos que inciden de forma directa en los resultados obtenidos. Por un lado, la cobertura de la base de datos, la exhaustividad de los datos de los registros, el índice de errores, así como la indización de los documentos se convierten en parámetros fundamentales. Por otro, el sistema de búsqueda determina la capacidad de recuperar a través de lenguaje libre y/o controlado, la posibilidad de delimitar las estrategias de búsqueda o de combinar conjuntos mediante diferentes operadores, por ejemplo. Por último, entra también en juego el conocimiento del usuario de las técnicas de recuperación de información y de la materia específica de la base de datos, así como de la estructura y organización de la misma.

Cada uno de estos aspectos cobra mayor o menor relevancia en función de múltiples variables, por lo que no es posible identificar la primacía de ninguno de ellos sin delimitar previamente el contexto de actuación. Ahora bien, si admitimos que una de las formas de optimizar la precisión de los resultados de una búsqueda es recurrir al uso de lenguajes controlados, tenemos necesariamente que admitir que la indización desempeña un papel fundamental para lograr tanto la pertinencia como la exhaustividad deseadas en toda búsqueda documental.

Sin embargo, no todas las disciplinas cuentan con tesauros específicos e idóneos para la recuperación de información en forma electrónica, ni tampoco con una práctica habitual por parte de los usuarios que haga de los descriptores o encabezamientos términos utilizados familiar y adecuadamente. Así, por ejemplo, la terminología empleada en la mayor parte de las disciplinas que conforman las ciencias sociales es ambigua, imprecisa, provisional e inestable, por lo que carece de la exactitud que rige la de otras ciencias. Igualmente, el significado de los términos cobra acepciones diferentes según la materia, la época, el contexto social en el que se utilizan o el enfoque de las distintas escuelas. Y precisamente esta inestabilidad y falta de normalización terminológica plantea serias dificultades tanto en la elaboración de tesauros como en la indización de los documentos (Sahli, 1981; Rigss, 1979).

Todo ello es consecuencia directa de las características intrínsecas de las materias pertenecientes a las ciencias sociales, como, por ejemplo, la subjetividad que preside gran parte de los trabajos de este campo, o la fragmentación que ha dado lugar a múltiples disciplinas y subdisciplinas. Las ciencias sociales carecen del carácter universal del que gozan otras ciencias, por lo que es difícil consensuar una terminología válida con independencia del contexto social, cultural, político o económico para el que se utilice. Todos estos problemas terminológicos afectan, sin duda, al proceso de indización y, por consiguiente, al de recuperación, ya que los términos que describen los conceptos tratados en los documentos deben ser los mismos a los utilizados al interrogar al sistema mediante un lenguaje controlado.

A todo esto hay que añadir que todavía hoy, una gran parte de las bases de datos de ciencias sociales siguen utilizando encabezamientos de materia para representar el contenido de los trabajos en ellas referidos. La progresiva tesaurización de las listas de encabezamientos de materia ocurrida a raíz de la aparición de las bases de datos como los equivalentes electrónicos de los tradicionales repertorios impresos ha tenido una incidencia mucho menor en el campo de las ciencias sociales y las humanidades que en disciplinas del ámbito de la tecnología, de las ciencias experimentales o de las ciencias de la salud.

Los problemas que presenta el uso de encabezamientos para la recuperación de información por materias han quedado demostrados a lo largo de numerosos estudios (Cochrane y Markey, 1983; Larson y Graham, 1983; Cochrane, 1985; Lipetz y Paulson, 1987; Kaske, 1988; Hunter, 1991; Larson, 1991). Los encabezamientos de materia se pensaron para que el usuario los reconociera a través de una búsqueda alfabética y lineal, pero no para que los formulara, ya que su lengua no es la corriente; muchas veces son demasiado generales o demasiado específicos; los subencabezamientos no siempre ayudan; y tanto el orden de la cadena como las inversiones carecen de sentido para la lógica de un usuario no profesional. Así, los encabezamientos de materia, pensados para un entorno manual en el que la búsqueda se basa en el reconocimiento, no funcionan en un entorno automatizado, donde la búsqueda se basa, fundamentalmente, en la formulación.

Sin embargo, la implementación de la versión electrónica de los índices y repertorios de resúmenes no ha servido, en el caso de las ciencias sociales, para que los productores y distribuidores de estas bases de datos se replantearan este importante problema del acceso por materias, por lo que sus usuarios se encuentran con graves dificultades para obtener resultados satisfactorios en este tipo de búsquedas.

El objetivo principal de este trabajo es analizar el lenguaje de indización utilizado en tres bases de datos internacionales de ciencias sociales. Para ello, en primer lugar, se analizará la relevancia y la consistencia de las formas aceptadas. En segundo lugar, se evaluarán las diferencias existentes entre el uso de un lenguaje controlado post-coordinado y un lenguaje pre-coordinado de cara a la recuperación de información por materias. Por último, se estudiará también la exhaustividad de la indización.

1.1. Descripción de las bases de datos objeto de estudio

Las tres bases de datos objeto de estudio son *PAIS International*, *IBSS Extra* y *Political Science Abstracts*. Se trata de tres bases de datos referenciales bibliográficas que abarcan el campo de las Ciencias Sociales.

PAIS International está producida por el organismo público estadounidense Public Affairs Information Service (PAIS), y tiene un crecimiento anual de

unos 15.000 registros. Esta base de datos es el equivalente electrónico de tres publicaciones del mencionado organismo, que son: *PAIS Bulletin* (1976-1990), *PAIS Foreign Language Index* (1972-1990) y *PAIS Internation in print* (1991-hasta la actualidad). Las materias específicas que abarca son, principalmente: Economía, Negocios, Legislación, Ciencia Política, Relaciones Internacionales, Administración Pública, Finanzas, Educación, Demografía, Estadística, y Sociología, entre otras. *PAIS International* hace uso de un lenguaje de indización de precoordinación de encabezamientos, idéntico, en su forma y estructura, al utilizado en los tres repertorios impresos que dan lugar a esta base de datos. Se trata de extensas cadenas jerárquicas que siguen la estructura tradicional de los encabezamientos de materia. No obstante, la rigidez que impone el acceso por este lenguaje controlado se palía de alguna forma ya que el sistema de recuperación permite acceder a cada uno de los términos que componen la cadena. El usuario tiene acceso en línea a la relación alfabética de encabezamientos.

El productor de *IBSS Extra* (International Building Science and Structured Abstract) es la British Library of Political and Economic Science, perteneciente a la Economics and Political School londinense. Su crecimiento es de, aproximadamente, 100.000 registros al año que se refieren a trabajos del ámbito de la Economía, la Ciencia Política, la Sociología y la Antropología, publicados en una amplia selección de publicaciones internacionales (aproximadamente el 30% de los registros hacen referencia a trabajos en idiomas distintos al inglés). Su cobertura temporal abarca desde 1980 hasta la actualidad, y los datos comprendidos entre los años 1980 y 1986 fueron recogidos por el International Committee for Social Science Information and Documentation. Esta base de datos utiliza un lenguaje de indización basado en la postcoordinación de descriptores. El usuario no tiene acceso a un tesauro en línea, ni tampoco a una relación exclusiva de los descriptores utilizados en el campo de materias. Sí puede, sin embargo, acceder a una lista de todos los términos de los campos de autor, título, fuente, resumen y materias por los cuales se puede acceder a los registros.

La base de datos *Political Science Abstracts* (PSA) está producida por el organismo privado estadounidense IFI/Plenum Data Corporation. Su crecimiento anual es de aproximadamente 10.000 referencias que se concentran, fundamentalmente, en trabajos del área de la Ciencia Política. Su cobertura temporal abarca publicaciones desde el año 1975 hasta la actualidad, y aunque predominan publicaciones del ámbito norteamericano (aproximadamente un 22,2% del total) tiene una cobertura geográfica internacional. El lenguaje de indización de esta base de datos es un claro reflejo de la paulatina tesaurización de las listas de encabezamientos, ya que conviven descriptores y encabezamientos de materia, si bien se observa una clara tendencia al uso de los primeros frente a los segundos.

2. Criterios de análisis de la indización

La calidad de la indización ha sido objeto de estudio por numerosos expertos, especialmente desde la segunda mitad de la década pasada (White y Griffith, 1987; Soergel, 1994; Extremeño y Moscoso, 1998). A partir de este momento empiezan a enfatizarse las cuestiones de índole conceptual en el análisis y el control de la calidad de los recursos electrónicos, en detrimento de los estudios meramente cuantitativos (Medaward, 1995). El interés por la calidad de la indización da lugar a toda una serie de trabajos dedicados a diseñar y aplicar metodologías válidas para evaluar este parámetro (Sievert y Verbeck, 1987; UNE, 1991; May, 1994; Palma Villalón, 1995).

Como principio general, el uso de descriptores en el proceso de indización debe regirse por las normas de construcción de los mismos. Así, han de representar un solo concepto, los sintagmáticos o términos compuestos no deben emplear conjunciones; los unitérminos deben ser un solo substantivo, no deben utilizarse verbos; si hay opción se prefiere el masculino frente al femenino, el singular frente al plural y la forma desarrollada frente a los acrónimos, excepto en aquellos casos en los que éstos se utilicen comúnmente; deben evitarse los modismos, así como respetarse el orden normal de la estructura sintáctica de la lengua en cuestión.

Además de estas pautas de índole general, existen, fundamentalmente, tres indicadores principales de la calidad de la indización de los documentos. Éstos son: la relevancia, la consistencia y la exhaustividad.

El principio de relevancia alude a la necesidad de utilizar descriptores que reflejen fielmente el contenido del documento. Según esto, deben rechazarse los descriptores demasiado generales y demasiado particulares con respecto a los conceptos expresados en los documentos. Hay que tener en cuenta, sin embargo, que la relevancia de un documento está directamente relacionada con la carga entrópica de la forma de indización elegida; es decir, con la aportación de información unitaria que comporta. Por ello, el principio de relevancia de los conceptos no carece de cierta subjetividad, tanto por parte del indizador como del propio usuario, pues, como ya afirmaba Perreault "una característica propia de la relevancia es que cualquier indicio sobre ella ya es de por sí valorativo y por tanto, no objetivo" (Perreault, 1966).

El principio de consistencia en la indización establece que un mismo concepto debe expresarse siempre con el mismo descriptor y la misma morfología. Así, la relación entre concepto y descriptor ha de ser biunívoca, de forma que a cada concepto le corresponda un descriptor y a cada descriptor un concepto.

La exhaustividad está relacionada con el número de nociones que caracterizan el contenido íntegro del documento y el número de descriptores empleados para describir los conceptos. La adecuación al principio de exhaustividad implica que todos los temas, objetos y conceptos tratados en el documentos estén bien

determinados en la indización. Es preciso tener en cuenta que esta cifra no debe limitarse de una forma arbitraria, sino que debe establecerse en función de la materia de la base de datos, el volumen de registros, la tipología de usuarios para quien va dirigida la base de datos y la propia estructura del lenguaje de indización. Como principio general se recomienda una media entre ocho y doce descriptores para los documentos referidos en una base de datos bibliográfica.

Además de estos tres principios fundamentales, existen otros dos criterios que sirven para medir la calidad de la indización: la selectividad y la uniformidad. El primero se refiere a la necesidad de que sólo estén representados aquellos conceptos que sean de interés para los usuarios de la base de datos en cuestión. La adecuación al principio de uniformidad consiste en que tanto indizadores como usuarios deben describir el contenido de los documentos de un mismo tema de la misma manera.

La consistencia y la relevancia son los principios que más directamente están relacionados con la exhaustividad y la pertinencia de los resultados de una búsqueda, entendidas ambas como las principales medidas de eficacia de la recuperación de información. Si entendemos la tasa de acierto o de exhaustividad como la proporción de documentos pertinentes encontrados en relación al conjunto de documentos pertinentes que posee la base de datos, y la tasa de precisión o pertinencia como la proporción de documentos pertinentes en relación al conjunto de documentos resultado de una búsqueda documental, es obvio que la exactitud de los descriptores utilizados en la indización, así como la necesidad de que cada concepto esté expresado con el mismo descriptor influirá decisivamente en el resultado de las búsquedas.

3. Metodología

Para lograr el alcance del objetivo principal de este trabajo se han elegido tres bases de datos cuyo lenguaje de indización refleja las tres tendencias ya explicadas. El análisis de la indización se ha basado en los principios de relevancia y consistencia de los términos empleados y su relación con las medidas de eficacia de la recuperación, así como en la exhaustividad de la descripción del contenido. El acceso a estas bases de datos se ha efectuado utilizando la versión en CD-ROM del distribuidor Silver Platter.

El grado de relevancia se ha analizado calculando valor de discriminación de las formas aceptadas, que indica el porcentaje de registros asociados a un descriptor o encabezamiento, y, por consiguiente, recuperables haciendo uso de él. Para ello, se ha calculado el cociente entre el número de registros asociados a la forma en cuestión y el total de la base de datos (Ju y y Achiruque, 1989). Se considera que una buena discriminación se aproxima al 0,05, lo que significa que el descriptor o encabezamiento del que se trate aparece, al menos, en el 5% de los

documentos de la base de datos. Para llevar a cabo el análisis de la relevancia se han escogido veintiocho conceptos relacionados con los acontecimientos políticos de los países de habla hispana durante las últimas décadas, y para identificar el descriptor o encabezamiento que mejor representa cada uno de estos conceptos se ha accedido a los índices de las bases de datos, así como al campo de materias de los registros.

En cuanto al análisis de la consistencia, éste se ha llevado a cabo identificando grupos o racimos de documentos, también denominados *clusters*, sobre la base de un contenido temático similar. Así, se han formado seis racimos (tres de temática amplia y tres referidos a temas muy específicos) con cinco documentos cada uno. Para la selección de los documentos se ha recurrido a los campos de título y resumen, nunca al de materias, ya que precisamente son los términos contenidos en éste el objeto del análisis. Una vez obtenidos los racimos, se han identificado los descriptores utilizados en dos o más documentos, con el fin de conocer la frecuencia de utilización de los mismos. Se considera que un descriptor o encabezamiento es consistente cuando está asignado a la mitad o más de los documentos de un racimo.

El análisis de la exhaustividad se ha realizado a partir de una muestra de cincuenta registros, obtenida de forma aleatoria, con un tamaño prefijado a efectos de demostración, sin tratamiento estadístico.

4. Análisis e interpretación de los resultados

4.1. Relevancia de la indización

Las tablas I, II y III presentan el valor de discriminación de los términos del lenguaje controlado, así como la discriminación de los mismos si el acceso se efectúa por lenguaje libre en el campo de título, en cada una de las tres bases de datos objeto de estudio. La tabla IV compara el porcentaje de registros recuperables por lenguaje controlado y por lenguaje libre en las tres fuentes analizadas.

Como principio general cabe decir que cuanto más se aleja hacia abajo el valor de discriminación de 0,05, menor es el número de registros capaz de recuperar el sistema, y, por consiguiente, mayor es la probabilidad de silencio en el resultado de la búsqueda. Por el contrario, cuanto más se aleje hacia arriba del valor indicado, mayor es la cantidad de registros susceptibles de recuperarse, por lo que la probabilidad de ruido aumenta.

En lo que respecta a *PAIS International*, los valores se han calculado sobre 375.000 registros, que equivale al total de la base de datos en el momento de realizar el estudio.

Tabla I

Análisis de la relevancia de la indización en *PAIS International*

Concepto	Término	Campo de título		Campo de materias	
		Frecuencia	V.Discriminac	Frecuencia	V.Discriminac
Terrorismo	terrorism	683	0,001	1.224	0,003
Libertades	freedom	494	0,001	407	0,001
Derechos	right	1.178	0,003	2.207	0,005
Interés Público	public interest	205	0,0005	269	0,0007
Orden Público	public policy	919	0,002	826	0,002
Comunidades Europeas	European Communities	1.068	0,002	9.673	0,02
OTAN	NATO	692	0,001	1.496	0,003
Tratados Internacionales	international trade	970	0,002	5.314	0,01
Conflictos Internacionales	international problems	134	0,0003	2.028	0,005
Diplomacia	diplomacy	705	0,001	586	0,001
Emigración	emigration	1.084	0,002	2.752	0,007
Relaciones Económicas	economics	323	0,0008	5.416	0,01
Transición Política	political transition	20	0,00005	571	0,001
Guerra Civil	civil war	108	0,0002	336	0,0008
Reformas Políticas	political reforms	83	0,0002	586	0,001
Proceso Electoral	elections	1890	0,005	2.895	0,007
Intentos involucionistas	coup d´etat	29	0,00007	295	0,0007
Patronal	labor union	9	0,00002	104	0,0002
Sindicatos	trade unions	496	0,001	4.134	0,01
Partidos Políticos	political party	271	0,0007	3.245	0,008
Fuerzas Armadas	armed forces	577	0,001	2.456	0,006
Iglesia	Church	478	0,001	1.756	0,004
Monarquía	monarchy	31	0,00008	70	0,0001
Gobierno	government	252	0,00006	1004	0,002
Administración Pública	public administration	161	0,0004	2.107	0,005
Poder Legislativo	legislative bodies	4	0,00001	88	0,0002
Poder Ejecutivo	executive	17	0,00004	140	0,0003
Poder Judicial	judiciary	10	0,000002	235	0,0006

Si se comparan los valores de discriminación en el campo de título y en el de materias, se observa que, a excepción de tres casos en los que éstos son iguales, en el resto siempre es mayor el obtenido en el campo de materias. En *PAIS International* el valor de discriminación de los encabezamientos de materia es muy bajo, ya que únicamente cuatro se mueven entre los valores recomendados, aunque en ninguno de ellos éste alcanza el 0,05 (tabla I). La capacidad de recuperación de estos cuatro encabezamientos oscila entre el 1 y el 2% de referencias.

Ocho encabezamientos son capaces de recuperar entre el 0,01% y el 0,09% de registros; dieciséis, entre el 0,1% y el 0,9%; y cuatro, entre el 1% y el 2% del total de referencias de la base de datos. Si, por el contrario, la estrategia de búsqueda se delimita al campo de título, la exhaustividad de los resultados desciende considerablemente, ya que ninguno de los términos es capaz de recuperar un porcentaje de documentos igual o mayor que el 1%. Así, siete de los términos de la muestra recuperarían entre el 0,001% y el 0,009% de referencias; siete también entre el 0,01% y el 0,09% y los otros trece términos de la muestra entre el 0,1% y 0,9% del total de registros de la base de datos (tabla IV).

Los resultados referentes a la base de datos *IBSS Extra* se han calculado sobre 188.209 registros (total de la base de datos en el momento de llevar a cabo el estudio). Atendiendo a los datos obtenidos en el análisis, la cifra de descriptores que se mueve dentro de un rango considerado aceptable triplica a la que resulta del estudio de la base de datos *PAIS International*. Así, doce de los descriptores de la muestra presentan valores de discriminación que se mueven entre el 0,01 y el 0,06 (tabla II). La recuperación por lenguaje libre en el campo de título arroja unos valores de discriminación siempre inferiores, aunque en cuatro casos éstos oscilan entre el 0,01 y el 0,04.

Los datos recogidos en la tabla IV constatan que quince de los descriptores de la muestra son capaces de recuperar entre un 0,1% y un 0,9% de registros de la base de datos *IBSS Extra*, y doce descriptores entre un 1% y un 6%. Ahora bien, si por el contrario, la búsqueda se realiza en el campo de título, al igual que ocurría en el caso anterior, la exhaustividad de los resultados es significativamente más baja, ya que los porcentajes de referencias recuperables son ostensiblemente menores. Así, siete de los términos seleccionados recuperarían entre el 0,01% y el 0,09% de los registros contenidos en *IBSS Extra*, dieciséis entre el 0,1% y el 0,9%, y sólo cuatro entre el 1% y el 4% (tabla IV).

En cuanto a los valores obtenidos en el análisis de la relevancia de la indización de *PSA*, éstos se han calculado sobre los 178.000 registros que formaban se el total de la base de datos en el momento de efectuar el estudio. De los encabezamientos/descriptores analizados en *PSA*, diecisiete presentan un valor de discriminación medido en centésimas, con un grado de relevancia entre el 0,01 y el 0,04. En el resto de los casos el cociente resulta en milésimas.

Tabla II

Análisis de la relevancia de la indización en *IBSS*

CONCEPTO	DESCRIPTOR	CAMPO DE TÍTULO		CAMPO DE MATERIAS	
		Frecuencia	V.Discrimin.	Frecuencia	V.Discrimin.
Terrorismo	terrorism	979	0,005	1.178	0,06
Libertades	freedom	1.738	0,009	2.048	0,01
Derechos	right	2.382	0,01	3.602	0,01
Interés Público	public interest	233	0,001	248	0,001
Orden Público	public policy	76	0,0004	256	0,001
Comunidades Europeas	European Communities	1966	0,01	5.373	0,02
OTAN	NATO	598	0,003	1.096	0,005
Tratados Internacionales	international trade	158	0,0008	563	0,002
Conflictos Internacionales	international problems	416	0,002	621	0,003
Diplomacia	diplomacy	1327	0,007	1.393	0,007
Emigración	emigration	613	0,003	934	0,004
Relaciones Económicas	economics	718	0,003	2.772	0,01
Transición Política	political transition	564	0,002	4.612	0,02
Guerra Civil	civil war	431	0,002	1.176	0,006
Reformas Políticas	political reforms	146	0,0007	1.481	0,007
Proceso Electoral	elections	3.386	0,01	4.279	0,02
Intentos involucionistas	coup d'etat	68	0,0003	209	0,001
Patronal	labor union	10	0,00005	52	0,0002
Sindicatos	trade unions	1659	0,008	4.587	0,02
Partidos Políticos	political party	746	0,003	3.415	0,01
Fuerzas Armadas	armed forces	725	0,003	1.115	0,005
Iglesia	Church	1.478	0,007	2.636	0,01
Monarquía	monarchy	277	0,001	519	0,002
Gobierno	government	7.555	0,04	9.661	0,05
Administración Pública	public administration	768	0,004	2.773	0,01
Poder Legislativo	legislative bodies	27	0,0001	203	0,001
Poder Ejecutivo	executive	102	0,0005	282	0,001
Poder Judicial	judiciary	35	0,0001	255	0,001

TablaIII

Análisis de la relevancia de la indización de *PSA*

CONCEPTO	TÉRMINO	CAMPO DE TÍTULO		CAMPO DE MATERIAS	
		Frecuencia	V.Discrimin.	Frecuencia	V.Discrimin.
Terrorismo	terrorism	658	0,03	1.508	0,008
Libertades	freedom	722	0,004	2.027	0,01
Derechos	right	883	0,004	2.576	0,01
Interés Público	public interest	99	0,0005	275	0,001
Orden Público	public policy	919	0,005	21.764	0,1
Comunidades Europeas	European Communities	353	0,001	2.518	0,01
OTAN	NATO	529	0,002	2.256	0,01
Tratados Internacionales	international trade	792	0,004	1.146	0,006
Conflictos Internacionales	international problems	75	0,0004	6.670	0,03
Diplomacia	diplomacy	876	0,04	3.338	0,01
Emigración	emigration	950	0,005	1.879	0,01
Relaciones Económicas	economics	390	0,002	7.159	0,04
Transición Política	political transition	33	0,0001	1.807	0,01
Guerra Civil	civil war	159	0,0008	863	0,004
Reformas Políticas	political reforms	85	0,0004	8.118	0,04
Proceso Electoral	elections	122	0,0006	1.847	0,01
Intentos involucionistas	coup d'etat	320	0,001	822	0,004
Patronal	labor union	41	0,0001	3.449	0,01
Sindicatos	trade unions	245	0,001	818	0,004
Partidos Políticos	political party	3.189	0,01	7.619	0,04
Fuerzas Armadas	armed forces	576	0,003	5.270	0,02
Iglesia	Church	467	0,002	1.173	0,006
Monarquía	monarchy	196	0,001	108	0,0006
Gobierno	government	3.999	0,02	2.681	0,01
Administración Pública	public administration	423	0,002	7.285	0,04
Poder Legislativo	legislative bodies	602	0,003	1.554	0,008
Poder Ejecutivo	executive	396	0,002	922	0,005
Poder Judicial	judiciary	284	0,001	2.977	0,01

La comparación entre los datos obtenidos en las búsquedas por lenguaje controlado y las realizadas en el campo de título ofrece unos resultados semejantes a los expuestos para las otras dos bases de datos objeto de este trabajo, ya que el porcentaje de documentos recuperables mediante las formas del lenguaje controlado es siempre mayor al que resulta de las búsquedas en el campo de título. Así, nueve descriptores/encabezamientos son capaces de recuperar entre el 0,1% y el 0,8% de referencias y diecisiete entre el 1% y el 4%. Hay que señalar que una de las formas presenta un grado de relevancia de 0,1, lo que indica una capacidad de recuperación del 10% de los registros de la base de datos, porcentaje que entraña una alta probabilidad de ruido (tabla III). Si la búsqueda se efectúa por lenguaje libre en el campo de título, siete términos recuperarían entre el 0,01% y el 0,09%; diecisiete entre el 0,1% y el 0,9%; y cuatro entre el 1% y el 4% de los registros de *PSA*.

Tabla IV

Lenguaje controlado versus lenguaje libre.
Porcentaje de registros recuperables

PAIS International		IBSS Extra		*PSA*	
Controlado	**Libre**	**Controlado**	**Libre**	**Controlado**	**Libre**
0,01	0,0002	0,02	0,005	0,06	0,01
0,02	0,001	0,1	0,01	0,1	0,01
0,02	0,002	0,1	0,01	0,4	0,04
0,03	0,004	0,1	0,03	0,4	0,04
0,06	0,005	0,1	0,04	0,4	0,05
0,07	0,006	0,1	0,05	0,5	0,06
0,07	0,007	0,1	0,07	0,6	0,08
0,08	0,008	0,2	0,08	0,6	0,1
0,1	0,02	0,2	0,1	0,8	0,1
0,1	0,02	0,3	0,1	0,8	0,1
0,1	0,03	0,4	0,2	1	0,1
0,1	0,04	0,5	0,2	1	0,1
0,2	0,05	0,5	0,2	1	0,2
0,2	0,07	0,6	0,3	1	0,2
0,3	0,08	0,7	0,3	1	0,2
0,3	0,1	0,7	0,3	1	0,2
0,4	0,1	1	0,3	1	0,2
0,5	0,1	1	0,3	1	0,3
0,5	0,1	1	0,4	1	0,3
0,5	0,1	1	0,5	1	0,4
0,6	0,1	1	0,7	1	0,4
0,7	0,1	1	0,7	2	0,4
0,7	0,2	2	0,8	3	0,5
0,8	0,2	2	0,9	4	0,5
1	0,2	2	1	4	1
1	0,2	2	1	4	2
1	0,3	5	1	4	3
2	0,5	6	4	10	4

4.2. Consistencia de la indización

Para llevar a cabo el análisis del principio de consistencia se han formado seis racimos atendiendo a los siguientes temas: Movimientos nacionalistas en la España actual; Monarquía Constitucional en España; El movimiento feminista; El proceso de la transición política de la dictadura a la democracia en España; Los sindicatos en la Europa moderna; y El terrorismo en la Europa del siglo XX. Cada racimo se ha formado con cinco documentos.

Con respecto a la base de datos *PAIS International*, ninguno de los encabezamientos del primero, quinto y sexto racimos son consistentes, ya que no se repiten con la frecuencia estimada (tablas V y VI). En el segundo racimo dos encabezamientos son consistentes y en el tercero y el cuarto sólo uno alcanza el nivel de consistencia. El porcentaje de encabezamientos que se repiten en cada racimo oscila entre el 4,5% y el 21,4% (tabla VI). Según los criterios de análisis explicados en la metodología, los resultados constatan que los encabezamientos de materia utilizados en la indización de esta base de datos tienen un nivel de consistencia muy bajo.

Tabla V
Frecuencia de los descriptores en los racimos temáticos analizados en *PAIS International*

RACIMOS	Descriptores	*Frecuencia*
1. Movimientos nacionalistas en la España actual	Basque-provinces-Spain-nationalism	2
	Political-parties-Catalonia-Spain	2
	Voting-Spain-Catalonia-Spain	2
2. Monarquía constitucional en España	Spain-Government-and-policies	4
	Juan-Carlos-I-King-of-Spain	2
	Monarchy-Spain	4
3. Movimiento feminista	Feminism	5
	Women´s organization	2
4. Transición política de la dictadura a la democracia en España	Democracy-Spain	4
	Spain-government-and-policies	4
5. Sindicatos en la Europa moderna	Trade-unions-organizing-activities	2
6. Terrorismo en la Europa del Siglo XX	Terrorism-international-aspects	2

Tabla VI
Índice de la consistencia de la indización de *PAIS International*

Racimos	Número de descriptores	Repetidos	Consistentes
Primero	18	3 (16,6%)	Ninguno
Segundo	14	3 (21,4%)	2 (14,2%)
Tercero	29	2 (6,8%)	1 (3,4%)
Cuarto	21	2 (9,5%)	1 (4,7%)
Quinto	22	1 (4,5%)	Ninguno
Sexto	19	1 (5,2%)	Ninguno

En cuanto a la base de datos *IBSS Extra*, cinco de los seis racimos analizados cuentan con un descriptor consistente (tablas VII y VIII). El porcentaje de descriptores que se repiten en cada racimo oscila entre el 10% y el 20%. El nivel de consistencia del lenguaje de indización utilizado presenta un grado considerablemente más alto que el caso anterior, lo que se debe, sin duda, a que se trata de un lenguaje post-coordinado.

Tabla VII
Frecuencia de los descriptores en los racimos temáticos analizados en *IBSS Extra*

RACIMOS	Descriptores	*Frecuencia*
1. Movimientos nacionalistas en la España actual	Nationalism	4
	Political doctrines	2
	Self government	2
2. Monarquía constitucional en España	Constitutional monarchies	2
	Heads of State	2
3. Movimiento feminista	Women´s movements	2
	Women and politics	2
	Feminism	3
4. Transición política de la dictadura a la democracia en España	Democratization	4
	Political development	2
5. Sindicatos en la Europa moderna	Trade unionism	3
	Labor relations	3
	Trade unions	2
6. Terrorismo en la Europa del Siglo XX	Political violence	2
	Terrorism	5

Tabla VIII
Índice de consistencia de la indización de *IBSS Extra*

Racimo	Número de descriptores	Repetidos	Consistentes
Primero	21	3 (14,2%)	1 (4,7%)
Segundo	10	2 (20%)	Ninguno
Tercero	16	3 (18,7%)	1 (6,25%)
Cuarto	20	2 (10%)	1 (5%)
Quinto	21	3 (14,2%)	1 (4,7%)
Sexto	20	2 (10%)	1 (5%)

La indización de la base de datos *PSA* es la que, según los resultados, presenta el nivel de consistencia más alto. Cinco de los seis racimos analizados cuentan con algún descriptor/encabezamiento consistente, y, además, en tres racimos, los consistentes son dos. (tablas IX y X). Sin embargo, los porcentajes de los repetidos en cada racimo, entre el 6,6% y el 14,2%, son considerablemente más bajos que en *IBSS Extra*. Apuntamos que, a excepción de una, todas las formas consistentes son descriptores, lo que viene a corroborar la tesis de que este tipo de indización se adecua mejor a las características de los entornos automatizados.

Tabla IX
Frecuencia de los descriptores en los racimos temáticos analizados en *PSA*

RACIMOS	**Descriptores**	*Frecuencia*
1. Movimientos nacionalistas en la España actual	Basques	2
	Social movements	2
	Nationalism	4
2. Monarquía constitucional en España	Democratic-process-and-institution	2
	Spain	4
3. Movimiento feminista	Female-sex	4
	Feminism-feminist	4
4. Transición política de la dictadura a la democracia en España	Democratic-process-and-institutions	2
	Democracy-changes-in-for-specific-countries-and conditions	4
	Spain	4
5. Sindicatos en la Europa moderna	Labor-unions-but-not-guilds	2
	Trade unions	4
	Workers-laborers-and-working-conditions	2
	Economics	2
6. Terrorismo en la Europa del Siglo XX	Terrorism	5
	Violence	4

Tabla X
Índice de la consistencia de la indización de *PSA*

Racimo	Número de descriptores	Repetidos	Consistentes
Primero	25	3 (12%)	1 (4%)
Segundo	30	2 (6,66%)	1
Tercero	24	2 (8,3%)	2 (8,3%)
Cuarto	30	3 (10%)	2 (6,6%)
Quinto	28	4 (14,2%)	1 (3,5%)
Sexto	23	2 (8,6%)	2 (8,6%)

Por consiguiente, si nos atenemos a los porcentajes de descriptores o encabezamientos repetidos en cada racimo, es la base de datos *IBSS Extra* la que cuenta con los porcentajes mayores. Incidimos en el hecho de que es la única que utiliza un lenguaje post-coordinado para la indización de todos los registros de la base de datos. Ahora bien, cabe señalar que la delimitación geográfica de los conceptos analizados no se representa de forma consistente en esta base de datos, mientras que en *PAIS International* y en *PSA* sí. Es decir, si se realiza una estrategia de búsqueda en la que se combinan el descriptor temático y el geográfico, los resultados adolecerían de un alto grado de silencio.

4.3. Exhaustividad

El principio de exhaustividad está directamente relacionado con el número de temas que caracterizan el contenido completo de un documento y el número de descriptores empleados para describir estos conceptos. A este respecto, la media recomendada es de entre ocho y doce descriptores.

La media de encabezamientos utilizados en los registros de *PAIS International* es de 4,6, lo que, en principio, representa una medida de exhaustividad por debajo de la estimada. De la muestra analizada, el mayor número de registros, veintisiete, tienen entre dos y cuatro encabezamientos; en veinte registros el contenido del documento está representado por entre cinco y siete encabezamientos, y únicamente en tres casos el campo de materias reúne entre ocho y nueve encabezamientos (tabla XI). Los datos proporcionados indican, además, que no existe una clara política de control en este aspecto.

La media de descriptores utilizados para representar el contenido de los documentos referidos en *IBSS Extra* es 6, que aunque más alta que la de *PAIS International*, no llega a la medida recomendada. Treinta de los registros de la muestra cuentan con un número de descriptores que oscila entre cinco y siete, doce registros tienen entre dos y cuatro y en seis casos el campo de materias se constituye con entre ocho y diez descriptores. Únicamente dos documentos están indizados con más de diez (tabla XI). Así, el conjunto mayor de registros se sitúa cerca de la medida recomendada.

La uniformidad en el número de descriptores asignados a cada documento es fundamental de cara a la recuperación de información, ya que la combinación de un número de conceptos en la estrategia de búsqueda superior a la media deriva en silencio, mientras que si la cifra es inferior da lugar a ruido. Por ello, la falta de homogeneidad detectada a este respecto impide al usuario formarse un modelo predeterminado del tipo de estrategia que debe llevar a cabo.

Por último, la media de descriptores de la muestra analizada en *PSA* se acerca también a seis. Se trata de la base de datos que presenta unos resultados más homogéneos, ya que se dan las menores variaciones en cuanto al número de descriptores utilizados. Así, el conjunto mayor, el formado por treinta y siete registros, cuenta con una cifra de descriptores que oscila entre cinco y siete. Hay nueve documentos indizados con entre dos y cuatro descriptores, y cuatro con entre ocho y diez (tabla XI).

Tabla XI.
Exhaustividad de la indización

Número de descriptores	Número de registros		
	PAIS International	*IBSS Extra*	*PSA*
2-4	27	12	9
5-7	20	30	37
8-10	3	6	4
+ 10	0	2	0

5. Conclusiones

Tras el análisis del principio de relevancia de la indización en las tres bases de datos objeto de estudio se concluye que, en general, las formas utilizadas en el lenguaje controlado no alcanzan el valor de discriminación recomendado. Sin embargo, en el caso de *PSA*, que hace uso de una indización mixta de lenguaje pre-coordinado y lenguaje post-coordinado, más de la mitad de las formas analizadas alcanzan valores cercanos al recomendado, y en casi todos los casos se trata de descriptores y no encabezamientos. En el caso de *IBSS Extra*, son un tercio los que se aproximan a este valor. En cuanto a *PAIS International*, no es significativo el número de encabezamientos que se mueven cercanos al valor aceptado.

Así pues, podemos concluir que el uso de descriptores es más adecuado en la indización de las fuentes electrónicas ya que su relevancia es mayor y, por consiguiente, garantizan una exhaustividad mayor en los resultados de las búsquedas.

Asimismo, los resultados obtenidos en este estudio constatan que la utilización de un lenguaje controlado posibilita la recuperación de conjuntos de registros mayores que los que resultan del uso de un lenguaje libre.

En lo que se refiere a la consistencia de las formas utilizadas en la indización de los documentos de las tres bases de datos analizadas, cabe decir que se detectan variaciones importantes entre los distintos racimos estudiados. Los resultados relativos a la frecuencia de repetición de los encabezamientos/ descriptores permiten concluir que el nivel de consistencia no alcanza, en ningún caso, las pautas recomendadas. Asimismo, el número de repeticiones en los racimos correspondientes representa un porcentaje muy bajo con respecto al total de los términos estudiados. *IBSS Extra* es la que presenta mejores resultados, mientras que *PAIS International* es la que tiene los índices más bajos. Si a ello añadimos que en *PSA* son los descriptores los que, en general, alcanzan siempre el grado más alto de consistencia, podemos concluir que el empleo de un lenguaje post-coordinado garantiza unos niveles de consistencia más altos que los que resultan del uso de un lenguaje pre-coordinado.

En cuanto a la exhaustividad de la indización de los documentos, cabe concluir que es éste el indicador de calidad de la indización que mejor se adecua a las pautas recomendadas. Los resultados constatan, además, que sólo en *PSA* se observa una política de control que garantice la uniformidad deseada, pues tanto en *PAIS International* como en *IBSS Extra* se observan variaciones significativas en los distintos registros.

Bibliografía

Cochrane, P. A. 1985. Redesign of catalogs and Indexes for Improved Online Subject Access. Phoenix: Oryx Press.

Cochrane, P. A. y K. Markey. 1983. Catalog use studies -since the introduction of online interactive catalogs: impact on design for subject access. En *Library and Information Science Research.* Vol. 5, no. 4, 337-363.

Collantes, L. Y. 1995. Degree of agreement in naming objects and concepts for information retrieval. En *Journal of the American Society for Information Science*. Vol. 46, no. 2, 116-132.

Extremeño, A. y P. Moscoso. 1998. El control de la calidad en las bases de datos de Ciencias Sociales. En *Boletín de la ANABAD.* Vol. 48, no. 1, 231-253.

Hunter, R. N. 1991. Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis. En *RQ*. Vol. 30, no. 4, 395-402.

Ju, C. M. y I. Achiruque. 1989. Quality of indexing in library and information science databases. En *Online Review.* Vol. 13, no. 1, 11-35.

Kaske, N. K. 1988. The variability and intensity overtime of subject searching in an online public access catalog. En *Information Technology and Libraries*. Vol. 7, no. 3, 273-287.

Larson, R. 1991. The decline of subject searching: long-term trends and patterns of index use in an online catalog. En *Journal of the American Society for Information Science*. Vol. 42, no. 3, 197-215.

Larson, R. y V. Graham. 1983. Monitoring and evaluating MELVYL. En. *Information Technology and Libraries.* Vol. 2, no. 1, 93-104.

Lipetz, B. A. y P. J. Paulson. 1987. A study of the impact of introducing an online subject catalog at the New York State Library. En *Library Trends*. Vol. 35, no. 4, 597-617.

May, N. A. 1994. A methodology for the measurement of quality of electronic databases. 3rd International Society for Knowledge Organization (ISKO), Copenhaguen (Denmark).

Medaward, K. 1995. Database quality: a literature review of the past and a plan for the future. En *Program*. Vol. 29, no. 3, 23-29.

Palma Villalón, M. V. 1995. Técnicas y métodos para mejorar la calidad de la indización y su recuperación en las bases de datos de Ciencias Sociales y Humanidades. En *Jornadas Catalanas de Documentación*, p. 223-239.

Perreault, J. M. 1966. Documentary relevance and structural hierarchy. En *American Documentation.*, 136.

Rigss, F. N. 1979. A New Paradigme for Social Sciences Terminology. En *UNESCO, INTERCEPT Project.* Vol. 6, no. 3, 150-158.

Sahli, M.S. 1981. Terminology of the Social Sciences: the Term Cognitive Processes in the Thesauri of Two-Discipline Information Systems. Michigan: UMI.

Sievert, E. y A. Verbeck. 1987. The indexing of the literature of online searching: a comparison of ERIC and LISA. En *Online Review*. Vol. 11, no. 2, 95-104.

Soergel, D. 1994. Indexing and retrieval performance: the logical evidence. En *Journal of the American Society for Information Science*. Vol. 45, no. 8, 589-599.

UNE 50-121-91. 1991. Documentación: métodos para el análisis de documentos, determinación de su contenido y selección de los términos de indización. Madrid: AENOR.

White, H. D. y B. C. Griffith. 1987. Quality of indexing in online databases. En *Information Processing and Management.* no. 3, 211-214.