LA CALIDAD EN LA INDIZACIÓN DE DOCUMENTOS: ELEMENTO INDISPENSABLE PARA OPTIMIZAR LA RECUPERACIÓN DE INFORMACIÓN

Ana Extremeño Placer[*]

RESUMEN:

Se establece una metodología a seguir para lograr un control de la calidad en los procesos de indización de los documentos de una base de datos bibliográfica, ya que la pertinencia e idoneidad a la hora de escoger los términos más adecuados para representar el contenido de los trabajos son considerados elementos clave a la hora de lograr una mayor y mejor recuperación informativa. La evaluación de la calidad de la indización se dirigirá a los tres principios elementales de la misma: la consistencia, la relevancia y la exhaustividad.

PALABRAS CLAVE: Indización, control de calidad, bases de datos bibliográficas, recuperación de información

ABSTRACT:

A methodology to evaluate the indexing process in a bibliographic database is applied. Indexing quality is analysed taken into consideration the three principal indicators of consistency, relevancy and exhaustivity.

KEY WORDS: Indexing, quality control, bibliographic databases, infomation retrieval.

1. Introducción:

Una de las principales premisas a la hora de diseñar una base de datos es que sea capaz de satisfacer las necesidades informativas del usuario de la misma, las cuales han de resolverse con la mayor exhaustividad y pertinencia posible.

El proceso de indización consiste en escoger los términos más apropiados para representar el contenido de un documento, por tanto, la pertinencia e idoneidad de los mismos son elementos claves en la evaluación de la calidad de una base de datos. La calidad de la indización es fundamental para lograr la eficacia en la recuperación de documentos pertinentes. Así, nuestro propósito es analizar la calidad del proceso de indización, con el objeto de conocer su incidencia en una recuperación de información idónea para el usuario, y establecer una serie de parámetros para evaluar dicha calidad [1]. La metodología establecida se aplica a una base de datos concreta en materia de ciencias sociales, al entender que éstas, junto con las ciencias humanas, utilizan un lenguaje menos preciso y normalizado y más ambiguo e inexacto que las ciencias experimentales y la tecnología lo que hace especialmente importante en esas áreas un control terminológico adecuado [2].

La evaluación de la calidad de la indización se rige por tres principios fundamentales: consistencia, relevancia y exhaustividad [3]. El principio de consistencia establece que cada concepto debe expresarse siempre con el mismo descriptor e idéntica morfología. La relación entre concepto y descriptor ha de ser biunívoca: a cada concepto le debe corresponder un descriptor y a cada descriptor un concepto. El principio de relevancia mide la exactitud con la que un concepto está representado por un término de indización. La exhaustividad está relacionada con el número de nociones que caracteriza el contenido íntegro del documento y el número de descriptores empleados para describir esos conceptos.

El análisis de la consistencia se realiza mediante la elección de grupos de documentos o racimos -*clusters* - de contenido temático similar, para lo cual se recurre a los campos de título y resumen, nunca al de descriptores, ya que son éstos los que se han de evaluar. Conviene que la mitad de los racimos abarque temas generales y la otra mitad específicos. Una vez seleccionados los racimos, se identifican los descriptores presentes en más de un

[*] Facultad de Documentación.Universidad de Alcalá

registro, con el fin de seleccionar los que se repiten en la mitad o más de los registros que componen el racimo. Estos últimos serán los que se atengan al principio de consistencia [4]

El grado de relevancia de un término se mide utilizando los llamados "valores de discriminación", descritos por Ju y Achiferuque [5], y que consisten en dividir el número de registros asociados a un descriptor por el número total de registros de la base de datos. Se estima que un valor aceptable de discriminación es el cercano al 0.05, ya que se corresponde con más del 5% de los documentos de cualquier base de datos. Para calcular la relevancia de la indización no es conveniente utilizar descriptores muy específicos, que aparecen en pocos registros, ni tampoco los muy generales, que se encuentran en muchos.

Por último, la evaluación de la exhaustividad está directamente relacionada con el número de términos que describen los diferentes conceptos del documento. Se recomienda una media de entre 8 y 12 descriptores.

El análisis de estos tres principios se ha realizado en una base de datos concreta sobre ciencias sociales, con tamaños de las muestras prefijados a efectos de demostración, pero habría que aplicar un tratamiento estadístico completo que nos determinase los tamaños necesarios para trabajar con una determinada cota de error. La base de datos elegida para ser analizada ha sido *Political Science Abstracts (PSA).* Se trata de una base de datos referencial, bibliográfica, que contiene alrededor de 178.000 registros desde 1975 a la actualidad, y cuyo contenido informativo se refiere a publicaciones periódicas y libros, todos ellos sobre Política y Análisis Político, con una cobertura mundial. El productor de la base de datos es el IFI/Plenum Data Corporation, organismo privado de Estados Unidos.

2. Metodología:

El análisis de la consistencia se ha efectuado identificando grupos de documentos, (*clusters)*, con un contenido temático similar. De esta manera, se han formado seis racimos, (tres de temática general y tres referidos a temas más específicos), con cinco documentos cada uno. Una vez construidos los racimos se han identificado los descriptores utilizados en dos o más documentos, anotando la frecuencia de aparición del término en el racimo correspondiente.

El grado de relevancia se ha calculado a través de los *valores de discriminación* de un grupo de veinticinco descriptores elegidos al azar.

El análisis de la exhaustividad de la indización se ha llevado a cabo midiendo el número medio de términos empleados para describir un documento en una muestra de cincuenta registros.

3. Resultados:

3.1. Consistencia de la calidad

A continuación se muestran los 6 racimos temáticos analizados, junto con los descriptores empleados. Se subrayan aquellos términos que aparecen repetidos en uno o más documentos.

1) Movimientos nacionalistas en la España actual:

1. Army all nations/ basques/ dictator/ nationalization/ secrecy/ social movements
2. Judge/ nationalism/ political opposition/ socialism/ Spain
3. Basques/ western-europe/ ideology/ international relations/ nationalism/ political party
4. Democratic process and institutions/ economic domination of any underdevelopment country by a development-one/ nationalism/
5. Ethnicity and ethnic groups/ nationalism/ political philosophy/ social movements/ social structure

2) Monarquía constitucional en España:

1. legitimacy/ monarch-monarchy-king-queen/ Spain
2. Western-Europe/ political-development-or-political-degeneration/ political-integration-o-different-areas/ political-systems-as-a-whole/Spain
3. Constitutions/ foreign-and-crossnational-political-institutions-and-behaviour/ leadership/Spain/ stability-or-inestability/ monarch-monarchy-king-queen/
4. Coercion-violence/ democratic-process-and-institutions/ armed-forces-and-policy/ sutudy-of-history-as-subject-matter/ military-industrial-complex/ modeling-models/ policy/ Portugal/ public-relations/ revolution/ society-as-a-whole/ Spain
5. Coup d'etat/ decision-making-and-game-theory/ democratic-process-and-institution/ Martin-Luther-King-Jr./ political-culture-Spain

3) El movimiento feminista:

1. Female-sex/ feminism-feminist
2. Female-sex/ feminism-feminist/ Iran-islamic/ worl-and-religion/revolution/socialism
3. Female-sex/ globalization-on-a-global-scale/ political-development-or-political-degeneration/ political-evaluation/ sex/ social-sciences
4. Feminism-feminist/ muslim-people-and-religion/ nationalism/ political-movement
5. Africa/ class-consciousness/ feminine-politics/ female-sex/ feminism-feminist/ Madagascar

4) El proceso de la transición política de la dictadura a la democracia en España.

1. Democratic-process-and-institutions/ contemporary-Europe/ modern-Greece/ Italy/ Portugal/ Spain
2. Choice-in-any-context/ corporatism/ democracy-changes-in-for-specific-countries-and-conditions/ power-dominant-grouping-of-a-society/ regimes/ totalitarianism
3. Democracy-changes-in-for-specific-countries-and-conditions/ elections/ political-evaluation/ political-party/ reform-reformers/ Spain
4. Democracy-changes-in-for-specific-countries-and-conditions/ Democratic-process-and-institutions/ contemporary-Europe/ modeling-models/ political-development-or-political-degeneration/ Spain
5. Democracy-changes-in-for-specific-countries-and-conditions/ international-relations/ political-analisys/ Portugal/ Republic-of-South-Africa/ Spain

5) Los Sindicatos en la Europa moderna:

1. All or most industry/ Italy/ labor unions but not guilds/ political economy/ trade unions
2. Capitalism/ economics strategy/ theory building or theoretical approach/ trade unions/ workers laborers and working conditions
3. Bulgaria/ communism/ democracy changes in for specific countries and conditions/ Eastern Europe/ political party/ trade unions
4. Economics/ labor unions but not guild/ pluralism/ power participation in decision making/ contemporary Russia/ workers laborers and working conditions
5. Economics/ Great Britain/ law and legal systems/ marketing system/ reform reformers/ trade unions

6) Terrorismo en la Europa del siglo XX.

1. Basques/ Spain/ terrorism
2. Power-dominant-groupings-of-a-society/ Irish-republican-army/ peace-peace-movement-peace-groups/ terrorism/ violence
3. Anti-semitism/ Canada/ conservatism-and-conservatives/ fascism/ nazi-movement-all-nations/ radical-radicalism/ terrorism/ violence
4. Terrorism/ violence
5. Public-image-image-construction/ international-relations/ mass-media-newspapers-magazines-tv-radio- etc./ terrorism/ violence

Tabla I. Frecuencia de descriptores del racimo *Movimientos nacionalistas en la España actual*

DESCRIPTORES	FRECUENCIA
Basques	2
Social movements	2
Nationalism	4

Únicamente un término aparece en más de la mitad de documentos, por tanto, su consistencia y capacidad de recuperación sobre el tema específico del grupo, son elevadas.

Tabla II. Frecuencia de descriptores del racimo *Monarquía constitucional en España*

DESCRIPTORES	FRECUENCIA
Monarch-monarchy-king-queen	2
Democratic-process-and-institution	2
Spain	4

En este grupo temático no se ha aplicado un criterio homogéneo a la hora de la indización, y únicamente el topónimo alcanza el grado de consistencia. Sorprende que los otros dos descriptores no alcancen la consistencia puesto que dan un sentido muy exacto y preciso al tema.

Tabla III. Frecuencia de descriptores del racimo *Movimiento feminista*

DESCRIPTORES	FRECUENCIA
Female-sex	4
Feminism-feminist	4

La consistencia de los dos términos más empleados está plenamente garantizada. Grupo muy homogéneo y una indización muy bien homologada.

Tabla IV. Frecuencia de descriptores racimo *Proceso de la transición política de la Dictadura a la Democracia en España*

DESCRIPTORES	FRECUENCIA
Democratic-process-and-institutions	2
Democracy-changes-in-for-specific-countries-and-conditions	4
Spain	4

Existen dos descriptores con elevado grado de consistencia. Uno, compuesto de varios términos, caso frecuente en esta base de datos y que podríamos denominarlos "multidescriptores". El otro, es el topónimo correspondiente a nuestro país.

Tabla V. Frecuencia de descriptores del racimo *Sindicatos en la Europa Moderna*

DESCRIPTORES	FRECUENCIA
Labor-unions-but-not-guilds	2
Trade-unions	4
Workers-laborers-and-working-conditions	2
Economics	2

El único descriptor que alcanza el nivel de consistencia es un término específico y que, claramente expresa el tema central del racimo. Sorprende que no se utilice en el cuarto documento.

Tabla VI. Frecuencia de descriptores del racimo *Terrorismo en la Europa del siglo XX*

DESCRIPTORES	FRECUENCIA
Terrorism	5
Violence	4

Existe un descriptor presente en todos los documentos elegidos. También en esta ocasión el término es muy específico y con una clara evidencia de lograr la plena consistencia. Hay un segundo descriptor, que también alcanza el grado de consistencia, y que es un término simple con una carga semántica totalmente vinculada al primero. Entre ambos la conectividad de documentos temáticamente afines está garantizada.

La conclusión del análisis de la consistencia en la indización de la base de datos analizada es que ésta es elevada. De los seis racimos estudiados cinco tienen descriptores que la alcanzan, lo que asegura una correcta correspondencia entre los conceptos y los términos que los expresan.

3.2. Relevancia de la indización

Se ha puesto en primer lugar el término en español y a continuación su transcripción al inglés, tal y como aparece en la base de datos, con el objeto de facilitar la búsqueda a cualquier usuario, seguido del número de referencias que contienen el término.

Tabla VII. Valores de discriminación de los descriptores

Término español	Término inglés	Frecuencia	Total registros	Porcentaje
Terrorismo	terrorism	1.508	178.000	0,008

Término español	Término inglés	Frecuencia	Total registros	Porcentaje
Libertades	freedom	1.532	178.000	0,008
Derechos	right	4.630	178.000	0,02
Interés Público	public interest	275	178.000	0,001
Orden Público	public policy	21.764	178.000	0,1
Estado	State	7.384	178.000	0,04
OTAN	north atlantic treaty org.	2.256	178.000	0,01
Tratados Internacionales	international trade	6.493	178.000	0,03
Diplomacia	diplomacy	3.338	178.000	0,01
Emigración	emigration	1.879	178.000	0,01
Relaciones Económicas	economics	20.003	178.000	0,1
Gobierno	Government	14.184	178.000	0,07
Guerra Civil	civil war	865	178.000	0,004
Proceso Electoral	elections	4.331	178.000	0,02
Intentos involucionistas	coup d´etat	822	178.000	0,004
Patronal	labor unions	3.449	178.000	0,01
Sindicatos	trade unions	818	178.000	0,004
Partidos Políticos	political party	7.202	178.000	0,04
Fuerzas Armadas	armed forces	5.270	178.000	0,02
Iglesia	Church	1.232	178.000	0,006
Monarquía	monarchy	108	178.000	0,0006
Administración Pública	public administration	7.285	178.000	0,04
Poder Legislativo	legislative bodies	1.554	178.000	0,008
Poder Ejecutivo	executive	3.219	178.000	0,01
Poder Judicial	judiciary	2.977	178.000	0,01

El grado de relevancia alcanza un promedio de 0,02, lo que significa que la mayoría de los descriptores analizados están dentro de la considerada como idónea, puesto que ésta se mueve en valores cercanos al 0,05, lo que supone una recuperación de documentos del 5%.

Se observa tendencia a utilizar términos simples, más fáciles de aproximarse a un genérico que a un específico, por lo que, de cara a la recuperación de información especializada, es aconsejable no limitar la búsqueda a descriptores sino ayudarse de los campos de texto libre.

Hay un segundo tipo de descriptores, también frecuentes, y que ya han sido denominados como "multidescriptores", que serían más indicados para realizar las búsquedas específicas.

A continuación se exponen los términos que entran dentro de un grado de relevancia aceptable, con el porcentaje de documentos que son capaces de recuperar cada uno de ellos:

Tabla VIII. Porcentaje de documentos recuperables por cada descriptor

Descriptor	Valor de discriminación	% Documentos que recuperan
Rights	0,02	2
State	0,04	4
NorthAtlant.Treaty O.	0,01	1
International Trade	0,03	3
Diplomacy	0,01	1
Emigration	0,01	1
Government	0,07	7
Elections	0,02	2
Labor Unions	0,01	1
Political Party	0,04	4
Armed Forces	0,02	2
Public Administration	0,04	4
Executive	0,01	1
Judiciary	0,01	1

El resto de descriptores tiene una relevancia menor y, para recuperar documentos, se mueven en cifras de centésimas o milésimas. Frecuentemente la razón de la menor relevancia es

síntoma de mayor especificidad en los términos, los cuales han de ser recuperados utilizando sinónimos, lo que es habitual en las búsquedas en bases de datos. Se ha estimado que, en *PSA*, el número de sinónimos necesarios para recuperar un porcentaje exhaustivo de documentos oscila entre dos y tres para cada término, dependiendo de su grado de especificidad.

3.3. Exhaustividad de la indización

Tabla IX. Descriptores por documento

Doc.	Nº descript.	Doc.	Nº descript.
1	6	26	2
2	6	27	4
3	5	28	5
4	3	29	5
5	6	30	2
6	6	31	5
7	6	32	5
8	6	33	6
9	3	34	6
10	7	35	6
11	9	36	6
12	4	37	6
13	6	38	5
14	2	39	6
15	10	40	6
16	6	41	6
17	6	42	6
18	6	43	7
19	3	44	9
20	3	45	7
21	5	46	6
22	7	47	6
23	6	48	7
24	10	49	6
25	7	50	6

La media de descriptores utilizados por cada uno de los documentos es de **5,9**, lo que implica una exhaustividad no muy alta, ya que la recomendación más extendida es la de utilizar un número de descriptores entre 8 y 12 por documento. Sin embargo, se observa que este número es muy homogéneo para todos los documentos, existiendo muy pocas diferencias entre uno y otro, lo cual probablemente sea síntoma de homologación en este aspecto.

4. Conclusiones:

4.1. La base de datos analizada verifica un elevado control terminológico que se demuestra en el alto grado de consistencia alcanzado en la indización. Así, *Political Science Abstracts* respeta el principio de biunivocidad en cuanto a la relación entre concepto y descriptor.

4.2. Los descriptores analizados son capaces de recuperar un promedio del 2% de documentos en la base de datos objeto de análisis. Por tanto, el grado de relevancia en la indización es elevado. La pertinencia en la recuperación de documentos está garantizada.

4.3. El grado de exhaustividad en el número de descriptores empleados por documento está por debajo de la media. Sin embargo, hay que tener en cuenta la peculiaridad de los mismos, puesto que se trata de descriptores que incluyen una cadena de términos de genéricos a específicos, por lo que no es necesario emplear gran número de ellos para describir el contenido temático.

4.4. Dadas las mayores dificultades en cuanto a homologación terminológica de las Ciencias Sociales, la conclusión general es que la indización de la base de datos *Political Science Abstracts,* teniendo en cuenta sus características de postcoordinación y control terminológico en todos sus aspectos, asegura una recuperación documental con un elevado grado de pertinencia, relevancia y exhaustividad.

5. BIBLIOGRAFIA

[1] May, N.A. A methodology for the measurement of quality of electronic databases. *Proceedings of the 3td International Society for Knowledge Organization, (ISKO), Conference.* 1994, junio, 20-24, Copenhaguen, Denmark

[2] White, H.D.; Grifith, B.C. Quality of indexing in online databases*. Information Processing and Management*, 23, 1987, 211-214

[3] Extremeño, A.; Moscoso, P. El control de la calidad en bases de datos de Ciencias Sociales. *Boletín de la ANABAD*, XVLIII (1), 1998, 231-253.

[4] H.R. Tibbo. Indexing for humanities. *Library and Information Science Abstracts,* 45 (8), 1994, 607-19

[5] Palma Villalón, M.V. Técnicas y métodos para mejorar la calidad de la indización y su recuperación en bases de datos documentales de Ciencias Sociales y Humanidades. *V Jornades Catalanes de Documentació*, 1995, Barcelona, 223-239

[6] D. Soergel. Indexing and retrival performance: the logical evidence. *Library and Information Science Abstracts*, 45 (8), 1994, 589-99

[7] White, H.D.; Grifith, B.C. Quality of indexing in online databases*. Information Processing and Management*, 23, 1987, 211-214

[8] Ju, C. M. y Achiferuque, I. Quality of indexing library and information science database. *Online Review,* 13(1), 1989, 11-35.