CALIDAD DE LA INDIZACIÓN E INCIDENCIA DE ERRORES EN LA BASE DE DATOS ECOSOC

Ana Extremeño*

Resumen: Se presentan los resultados del análisis y de la evaluación de la base de datos ECOSOC, producida y distribuida por el CINDOC, relativos a los registros sobre Ciencia Política. Se aplican indicadores de calidad para la evaluación del proceso de indización de los términos empleados en la descripción temática de la base de datos, así como indicadores formales con el fin de analizar la estructura de los registros bibliográficos e identificar el índice de errores contenidos en cada uno de los campos que componen dichos registros.

Palabras clave: bases de datos bibliográficas; control de calidad; indicadores de calidad; indización; detección de errores.

Abstract: Results of the analysis and evaluation of ECOSOC database are presented. Quality indicators are used in order to evaluate the indexing process. Formal indicators are used in the analysis of the records structure and in the process of error detection.

Key words: bibliographic databases; quality control; quality indicators; indexing; error detection.

1 Introducción

El control de calidad aplicado a la gestión y tratamiento de las bases de datos documentales se ha convertido en una cuestión prioritaria para los sectores implicados en la industria de la información (1-7). Ahora bien, gran parte de los estudios dedicados a este tema se limitan a evaluaciones que enfatizan el aspecto formal de las bases de datos. Sin embargo, la evaluación de la indización es de suma relevancia, ya que este proceso está en íntima relación con el proceso de recuperación de la información contenida en aquéllas.

También es un aspecto fundamental de la calidad de una base de datos documental la estructura de los registros que la componen, ya que ésta incide de forma directa en las posibilidades de recuperación de documentos. Como también inciden los errores en la cobertura temporal.

Por último, los estudios encaminados a detectar el índice de errores vienen siendo objeto de preocupación por parte de los expertos desde hace ya más de 25 años, período en el cual se ha analizado cuidadosamente el lenguaje con la ayuda de los ordenadores, lo que ha dado lugar a la aparición de una nueva disciplina denominada Lingüística Computacional (8). La mayoría de las investigaciones en esta área se han centrado en la lengua inglesa y apenas existen trabajos al respecto sobre el idioma español. Estos últimos se han limitado a estudios de frecuencias de términos o palabras y a la aplicación de la

* Facultad de Documentación. Universidad de Alcalá. Correo electrónico: ana.extre@uah.es.
Recibido: 16-12-98; 2.ª versión 12-4-99.

ley de Zipf, según la cual la frecuencia de aparición de los términos en un texto es inversamente proporcional al número de términos que tienen una frecuencia dada (9).

ECOSOC es la base de datos española bibliográfica referencial más representativa en Economía, Sociología, Antropología y Ciencia Política, tanto en su cobertura temática como geográfica y temporal. En el momento de realizarse este análisis contaba con un total de 85.650 registros, extraídos de un número aproximado de 300 publicaciones periódicas, además de algunas actas de congresos, monografías, informes, etc., de los cuales 8.493 corresponden a Ciencia Política. Estos últimos han sido extraídos de aproximadamente 70 títulos diferentes de revistas, aunque el número que se corresponde con la totalidad de revistas, desde el inicio de la base de datos hasta la actualidad, es 363. El crecimiento anual de esta base de datos es de, aproximadamente, 5.000 referencias, de las cuales unas 500 corresponden a Ciencia Política. La entrada de los datos es permanente. Los datos están estructurados en un número de campos que oscila entre 20 y 25, según el área temática.

Está producida y distribuida por el Centro de Información y Documentación Científica (CINDOC) del Consejo Superior de Investigaciones Científicas. ECOSOC forma parte de la base de datos ISOC, que recoge la producción científica publicada principalmente en revistas de Ciencias Sociales y Humanidades, editadas en España, desde el año 1975 a la actualidad. La base de datos ISOC se divide en sub-bases de datos sobre cada disciplina científica concreta dentro de las áreas mencionadas. Todas ellas se pueden consultar conjuntamente o por separado, tanto en línea como en CD-ROM.

El propósito de este trabajo de investigación es evaluar la base de datos bibliográfica ECOSOC a través del análisis del proceso de indización, así como del índice de errores y el estudio de la estructura de los registros bibliográficos. De esta forma, se procede, en primer lugar, a un análisis cualitativo de la cobertura temática de esta base de datos, a partir de los términos empleados para describir el contenido de los documentos referenciados. En segundo lugar, se lleva a cabo un análisis formal que abarca, por un lado, los diferentes aspectos relativos a la estructura de los registros, y, por otro, un estudio de detección de errores, con el fin de conocer su incidencia en función de una óptima recuperación de la información. El análisis efectuado en este trabajo se ha centrado en los registros que se refieren a Ciencia Política en todos sus aspectos y ámbitos geográficos.

1.1 La calidad de la indización

El incremento de la distribución de la información electrónica ha supuesto la demanda de un mayor estándar de calidad del que se esperaba en los materiales impresos tradicionales. Esta demanda ha significado una preocupación significativamente superior en cuanto a la fiabilidad de la información. El proceso de indización es la tarea central de todo sistema documental, ya que los términos que describen los conceptos tratados en los documentos deberán ser los utilizados a la hora de interrogar al sistema mediante un lenguaje controlado. De esta forma, la calidad de la indización es fundamental para lograr tanto la pertinencia como la exhaustividad deseadas en toda búsqueda documental.

Los descriptores utilizados en el proceso de indización tienen que respetar las normas de construcción de los mismos, que, en líneas generales, son las siguientes:

han de representar un solo concepto —los sintagmáticos o términos compuestos no emplearán conjunciones—; cuando se trate de unitérminos deben ser un solo sustantivo, nunca verbos; si hay opción se prefiere el masculino, el singular y la forma desarrollada frente al acrónimo (excepto en palabras comúnmente utilizadas y admitidas); deben evitarse los modismos, así como respetarse el orden normal de la estructura sintáctica.

Los tres principios fundamentales que rigen la evaluación de la calidad de la indización de una base de datos documental son el grado de consistencia o coherencia, la relevancia y la exhaustividad (10).

El principio de consistencia en la indización establece que un mismo concepto debe expresarse siempre con el mismo descriptor y la misma 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.

La relevancia es la exactitud con la que un descriptor representa un concepto. Su evaluación se ha efectuado a partir de una serie de descriptores en los que se han excluido tanto los más específicos como los más genéricos, puesto que ambos tipos distorsionarían los resultados.

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. A menudo se utiliza, para medir la exhaustividad, el porcentaje de descriptores de un registro, siendo conscientes que ello solo indica el número de descriptores por documento y no contempla otros aspectos también dignos de consideración, tales como el grado de pre-coordinación que utiliza la base de datos, las correcciones de la indización y la propia política de asignación de descriptores del centro. Es preciso tener en cuenta que el número de descriptores no puede limitarse de una forma arbitraria, sino que debe establecerse en función de la materia de la base de datos, el volumen de registros, así como el tipo de usuarios y la propia estructura del lenguaje de indización. En bases de datos bibliográficas se recomienda una media entre ocho y doce descriptores.

2 Metodología

Para lograr el alcance de los objetivos descritos se han utilizado métodos estadísticos que nos han permitido elegir las muestras objeto de análisis, establecer el nivel de calidad aceptable, examinar dichas muestras y decidir la aprobación o el rechazo al nivel de calidad aceptable.

La evaluación de la calidad de la indización se ha realizado atendiendo a los tres principios fundamentales de la misma: principio de consistencia, principio de relevancia y principio de exhaustividad.

El análisis de la consistencia 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[1] (11). Para la selec-

[1] La elección de esta cifra se debe a que la bibliografía existente al respecto estima apropiada la selección de un número entre 3 y 8 documentos por racimo.

ción de los documentos nos hemos servido de los campos de título y resumen, nunca del de descriptores, pues precisamente éstos son el objeto de evaluación. Una vez realizados los racimos se han identificado los descriptores utilizados en dos o más documentos, anotando la frecuencia de aparición del término en el racimo correspondiente. Cuando un descriptor está asignado a la mitad o más de los documentos de un racimo, se considera que se ajusta a la medida de la consistencia. Es decir, que la similitud del contenido temático de los documentos de cada racimo está garantizada.

Con el fin de conocer el grado de relevancia de un término, nos hemos servido de los denominados «valores de discriminación», cuya fórmula se debe a Ju y Achiruque (12). Para calcular la discriminación de un término se ha dividido el número de registros asociados a un descriptor por el número total de registros de la base de datos. Una buena discriminación ronda el 0,05, ya que se corresponde con el 5 % de los documentos de cualquier base de datos.

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.

Con el fin de determinar la cantidad de información a la que tienen acceso los usuarios de ECOSOC se ha analizado la estructura de los registros correspondientes a Ciencia Política, para lo cual se ha examinado cada uno de los campos que los componen.

El estudio de la detección de errores se ha realizado sobre un muestreo aleatorio de treinta registros, con objeto de obtener información preliminar de la posible magnitud de los mismos. Posteriormente, se observaron los estadísticos necesarios para un muestreo completo, en función del número de palabras contenidas en los campos analizados, número de errores y *ratio* entre estos dos conceptos medido directamente y luego multiplicado por 100 para establecer porcentajes. A continuación, se diseñó un muestreo aleatorio simple para optimizar el ajuste de las medias de los porcentajes de errores y conservar la proporcionalidad de la varianza en la base de datos estudiada, teniendo en cuenta que el nivel de confianza con el que se ha trabajado es del 95 %. Así, se ha analizado una muestra de 478 registros (sobre los 8.493 registros de Ciencia Política contenidos en ECOSOC), siendo la desviación *a priori* 0,00593, el error *ratio* 0,0000516, y el porcentaje de error del ±0,005.

El análisis se ha centrado en la detección de campos vacíos, datos mal situados, referencias duplicadas, normalización de las reglas de escritura (puntos, abreviaturas, etc.) y en la detección de errores de ortografía y tipografía (13-15). No se ha hecho distinción entre estos dos últimos tipos de errores (16). Se ha aplicado como baremo un porcentaje de errores admisibles de aproximadamente un 0,3 % del total de palabras, puesto que ésta parece ser la cifra considerada como aceptable por los estudiosos del tema (17).

La detección de duplicados se ha centrado en aquellos registros que, aunque difieren en los valores de los campos, se corresponden con un mismo documento. Para detectarlos se obtuvo un listado de registros del campo ISSN o ISBN —según el tipo de documentos—.

Los errores de cobertura temporal se verificaron a través del campo «Datos fuente de la publicación», que contiene la información relativa al año/s en que fue editada la publicación. Se trata de un campo obligatorio de la base de datos, por tanto, no

puede existir ningún registro que no contenga esta información, lo que no impide que existan errores humanos o documentos en los que no se especifique la fecha de edición. Se ha trabajado con la totalidad de los registros para llevar a cabo el análisis de la cobertura temporal.

3 Resultados del análisis de la calidad de la indización

La política de indización seguida en el análisis de contenido de los documentos referidos a Ciencia Política en la base de datos ECOSOC ha estado siempre presidida por la necesidad de normalizar los términos empleados en todos sus aspectos. Sin embargo, el paso del tiempo, la diferencia de criterios e, incluso, la propia evolución del lenguaje hace que a veces la normalización no sea tan sencilla de llevar a la práctica y ello puede derivar en dificultades a la hora de controlar los términos en su totalidad. Basándonos en la metodología explicada, vamos a proceder ahora a la presentación de los resultados del análisis de la calidad de la indización, siguiendo los tres principios básicos de consistencia, relevancia y exhaustividad.

3.1 Análisis de la consistencia

Los *clusters* elegidos para llevar a cabo este análisis, junto con los descriptores utilizados para la indización de cada uno de los cinco documentos que forman parte del racimo, se detallan a continuación. Aquellos descriptores que se encuentran en dos o más documentos aparecen en cursiva.

Movimientos nacionalistas en la España actual

1. Elecciones autonómicas / campaña electoral / discurso político / programas electorales / propaganda electoral / prensa / análisis de textos / partidos políticos / ideología política / partidos nacionalistas / *nacionalismo* / autonomismo / violencia / terrorismo.
2. *Nacionalismo* / cambio social / relaciones étnicas / etnocentrismo / relaciones de dominación / minorías / racismo / xenofobia / conflictividad social / violencia de estado / relaciones norte-sur / pluralismo cultural / democratización.
3. *Identidad nacional* / *nacionalismo* / análisis marxista / *derecho de autodeterminación* / relaciones de producción / estructura de clases / proletariado / conciencia de clase / *conciencia nacional* / movimiento obrero / izquierda política.
4. *Identidad nacional* / identidad de grupo / *comunidad* / *conciencia nacional* / grupos étnicos / rasgos culturales / lengua / sociedad multiétnica / *nación* / *poder* / modelo de estado / *nacionalismo* / *derecho de autodeterminación*.
5. *Nacionalismo* / *nación* / *poder* / identidad social / fenómenos sociales / lengua / raza / *comunidad* / ideólogos políticos / líderes políticos / actitudes políticas / órganos de expresión / prensa política / nacionalidad / estado / soberanía / autodeterminación / participación política / independentismo / lucha política / análisis histórico/ conceptos teóricos.

Monarquía constitucional en España

1. Monarquía parlamentaria / jefatura del estado / régimen constitucional / relaciones institucionales / *poder ejecutivo* / relaciones internacionales / responsabilidad política/ *poder judicial* / *poder legislativo* / fuerzas armadas / absolutismo / refrendo / función legislativa / *rey* / electorado / sucesión.
2. Transición política / democratización / legitimidad / opinión pública / imagen pública / aceptación / satisfacción / constitución / partidos políticos / líderes políticos / organización territorial del estado / congreso / senado / libertades públicas / libertad de expresión / reforma constitucional / representación política / *monarquía* / pena de muerte / encuestas de opinión / cuestionarios / análisis de resultados.
3. Sistema político / democracia / *monarquía.*
4. Derecho constitucional / cortes generales / bicameralismo / *rey* / relación entre los poderes / congreso de los diputados / competencias legislativas / poderes / elección directa / tramitación legislativa / control parlamentario / interpelación parlamentaria / diputados / incompatibilidad parlamentaria / inmunidad parlamentaria / responsabilidad / elegibilidad.
5. Pensamiento político / constitucionalismo / *poder ejecutivo* / reyes / *monarquía* / soberanía nacional / división de poderes / poder político / competencias / cortes constituyentes / *poder legislativo* / *poder judicial* / constitución de 1812.

El movimiento feminista

1. *Feminismo* / *movimiento feminista* / desarrollo histórico / etnocentrismo / movimientos sociales / sistema patriarcal / represión política / derechos humanos / maternidad.
2. *Feminismo* / *movimiento feminista* / organizaciones feministas / *condición de la mujer* / empleo femenino / mujeres / *rol social* / *rol familiar* / sociedad tradicional / cambio social / división sexual del trabajo / trabajo doméstico / explotación sexual / prostitución.
3. *Mujeres* / *condición de la mujer* / *rol social* / cambio político / cambio económico / *feminismo* / iglesia católica/ influencia social / aborto.
4. Países socialistas / *movimiento feminista* / *movimiento comunista* / mujeres / rol social / rol familiar / feminismo / análisis histórico.
5. *Movimiento comunista* / *feminismo* / lucha obrera / mujeres trabajadoras / integración europea.

El proceso de la transición política de la dictadura a la democracia en España

1. Cultura política / intolerancia / individualismo / caciquismo / guerra civil / franquismo / *transición política* / *democracia* / pasividad.
2. *Transición política* / *proceso político* / política gubernamental / problemas sociales / política social / estado del bienestar / crisis económica / crisis social / corrupción política / terrorismo de estado / independencia judicial / izquierda política / derecha política / actitudes políticas.

3. *Transición política* / democratización / *proceso político* / proceso constituyente / fuerzas políticas / fuerzas armadas / función política / análisis comparativo.

4. Parlamento / *democracia* / consolidación democrática / *transición política* / control político / régimen parlamentario / reforma institucional.

5. *Transición política* / sistema educativo / reforma educativa / ciencias sociales / enseñanza / estudio de casos.

Los Sindicatos en la Europa moderna

1. Relaciones sindicales / *sindicatos* / pluralismo sindical / *afiliación sindical* / participación sindical / elecciones sindicales / *estructura del empleo* / población activa/ población ocupada / trabajo precario / cooperativismo / parados / empleo femenino / sectores económicos / dimensión de la empresa / *negociación colectiva* / análisis comparativo / distribución por sexo.

2. Movimiento obrero / sindicalismo / historia social / bibliografía.

3. Transición política / *sindicatos* / tasa de sindicación / poder sindical / fuerzas políticas / *acción sindical* / movilización de masas / relaciones partido-sindicato / conflictividad laboral / huelga / negociación colectiva / tasa de desempleo / protección social.

4. *Sindicatos* / estructura sindical / *estrategia sindical* / *afiliación sindical* / *relaciones laborales* / *acción sindical* / mercado de trabajo / *estructura del empleo* / política de empleo / pacto social / *negociación colectiva.*

5. Integración europea / *relaciones laborales* / *negociación colectiva* / concertación social / *sindicatos* / *estrategia sindical* / cohesión económica / salarios / empleo / problemas sociales / protección social / tratado de la Unión Europea.

Terrorismo en la Europa del siglo XX

1. *Violencia* / medio social / *coyuntura política* / conflictividad social / *nacionalismo* / terrorismo / imagen pública / consenso político.

2. *Violencia* / *terrorismo* / violencia política / *represión política* / ideología / *grupos terroristas* / lucha armada / coyuntura política / *sistema político* / medio social / clima social.

3. Organizaciones armadas / movimientos sociales / *grupos terroristas* / *violencia* / *terrorismo* / ideología política / estructura organizativa / reclutamiento / origen social / identidad de grupo / conciencia colectiva / *nacionalismo* / *sistema político* / mitología / cultura tradicional / interpretación histórica / trayectoria política / escisión / asesinatos / secuestros.

4. Proceso de Montjuic / *terrorismo* / *atentados* / lucha obrera / anarquismo / *represión política* / opinión internacional / restauración.

5. *Terrorismo* / *grupos terroristas* / *atentados* / víctimas.

Las tablas de la I a la VI presentan la frecuencia de descriptores en cada uno de los racimos objeto de análisis.

Tabla I

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

Descriptores	*Frecuencia*
Nacionalismo	5
Identidad nacional	2
Derecho de autodeterminación	2
Comunidad	2
Nación	2
Poder	2
Conciencia nacional	2

Tabla II

Frecuencia de descriptores del racimo *Monarquía constitucional en España*

Descriptores	*Frecuencia*
Monarquía	3
Poder ejecutivo	2
Poder legislativo	2
Poder judicial	2
Rey	2

Tabla III

Frecuencia de descriptores del racimo *Movimiento feminista*

Descriptores	*Frecuencia*
Feminismo	5
Movimiento feminista	3
Rol social	3
Rol familiar	2
Mujeres	2
Movimiento comunista	2
Condición de la mujer	2

Tabla IV

Frecuencia de descriptores del racimo *Proceso de la transición política de la dictadura a la democracia en España*

Descriptores	*Frecuencia*
Transición política	5
Democracia	2
Proceso político	2

Tabla V

Frecuencia de descriptores del racimo *Sindicatos en la Europa Moderna*

Descriptores	*Frecuencia*
Sindicatos	4
Negociación colectiva	3
Afiliación sindical	2
Estructura del empleo	2
Acción sindical	2
Relaciones laborales	2
Estrategia sindical	2

Tabla VI

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

Descriptores	*Frecuencia*
Terrorismo	5
Grupos terroristas	3
Violencia	3
Nacionalismo	2
Sistema político	2
Atentados	2
Coyuntura política	2
Represión política	2

El descripor *Nacionalismo* es el único que aparece con el grado de consistencia deseable, entre los descriptores que se corresponden con el primer racimo. Esto demuestra su coherencia con el tema central del grupo temático analizado. Este descriptor agrupa el 100 % de los documentos. El resto de descriptores son términos muy relacionados con el primero, aunque no alcanzan un índice de consistencia recomendado.

Con respecto al segundo racimo analizado, se observa que varios términos se repiten en los distintos documentos pero únicamente uno, *Monarquía,* alcanza el grado de consistencia necesario. Esto significa que para poder recuperar la totalidad de documentos del racimo es necesario emplear varios descriptores, lo que pasa necesariamente por un conocimiento exhaustivo del lenguaje de indización utilizado, así como por la utilización de las herramientas apropiadas.

En cuanto al racimo *Movimiento feminista* cabe señalar, en primer lugar, que existe un numeroso grupo de términos que se repiten, así como que tres de ellos alcanzan el grado de consistencia requerido, con lo que la homologación de la indización de este grupo está garantizada con gran precisión. Si observamos todos los descriptores del racimo, se aprecia una frecuente utilización de sinónimos, lo cual puede dificultar la estrategia de búsqueda pero asegura una recuperación pertinente y exhaustiva de documentos.

En el cuarto racimo el único término que aparece con el nivel de consistencia deseable, puesto que se repite en todos los documentos del grupo, es un término compuesto y específico, *Transición política,* totalmente coherente con la temática central del racimo.

El tema central del grupo *Sindicatos en la Europa Moderna* está representado por el descriptor *Sindicatos* que guarda un alto nivel de consistencia ya que se repite en 4 de los 5 documentos. Sorprende, sin embargo, que no se haya utilizado en todos ellos dada la especificidad del tema. Hay un segundo descriptor con el índice de consistencia garantizado y otros 5 que, aunque no lo alcanzan, al menos se repiten dos veces. Esto quiere decir que la uniformidad temática está garantizada y que la indización está totalmente homologada. Todos ellos son términos específicos referidos al más genérico, que es el más repetido y que esa es la forma más correcta de indizar un documento para garantizar su plena recuperación por varios términos.

En el sexto grupo ocurre algo similar al caso anterior. Hay un término más genérico, *Terrorismo,* que se repite en los 5 documentos, por lo que su nivel de consistencia es pleno. Otros dos descriptores, *Violencia* y *Grupos terroristas,* alcanzan el grado de consistencia necesario, puesto que se repiten en la mitad o más de documentos. Si a ello añadimos que existen cinco términos más, que aunque no alcanzan la consistencia, sí se repiten en más de un registro, llegamos a la conclusión de que el racimo está indizado con gran consistencia y homologación.

3.2 Análisis de la relevancia

Como ya hemos explicado, la relevancia se refiere a la exactitud con la que un concepto que aparece en un documento está representado por un término de indización. Este principio se ha analizado utilizando los valores de discriminación de los descriptores. En este caso se ha trabajado con los registros que conformaban la totalidad de los referidos a Ciencia Política. El resultado del análisis de la relevancia se presenta en las tablas VII y VIII.

A la vista de los datos proporcionados en la tabla VII, se concluye que prácticamente la mitad de los términos, 13 de 29, aparecen en la base de datos con frecuencias del orden de las centésimas y el resto de las milésimas. Es decir, esos 13 descriptores tienen un valor de discriminación de entre 0,01 y 0,06. Se considera que un buen valor de discriminación es aquél que se aproxima a valores cercanos al 0,05 puesto que ello implica que ese término es capaz de recuperar el 5 % de los documentos de la base de datos. Incluso existe un término que supera todos los valores

Tabla VII

Valores de Discriminación de los descriptores. Total registros base de datos 8.493

Descriptores	*Frecuencia*	*Porcentaje*
Terrorismo	149	0,01
Libertades	73	0,008
Derechos	405	0,04
Interés público	10	0,001
Orden público	25	0,003
Comunidades europeas	15	0,001
OTAN	213	0,02
Tratados internacionales	39	0,004
Conflictos internacionales	94	0,01
Diplomacia	34	0,004
Emigración	18	0,002
Relaciones económicas	20	0,002
Transición política	537	0,06
Guerra civil	249	0,02
Reforma política	199	0,02
Proceso electoral	5	0,0005
Intentos involucionistas	61	0,007
Patronal	6	0,0005
Sindicatos	235	0,02
Partidos políticos	549	0,06
Fuerzas armadas	249	0,02
Iglesia	198	0,02
Monarquía	76	0,008
Gobierno	291	0,03
Estado	1.109	0,13
Administración pública	67	0,003
Poder legislativo	27	0,003
Poder ejecutivo	37	0,004
Poder judicial	44	0,005

Tabla VIII

Porcentaje de documentos recuperables por cada descriptor

Descriptores	*Valor de discriminación*	*Documentos que recuperan (%)*
Terrorismo	0,01	1
Derechos	0,04	4
OTAN	0,02	2
Conflictos internacionales	0,01	1
Transición política	0,06	6
Guerra civil	0,02	2
Reformas políticas	0,02	2
Sindicatos	0,02	2
Partidos políticos	0,06	6
Fuerzas armadas	0,02	2
Iglesia	0,02	2
Gobierno	0,03	3
Estado	0,13	1

restantes con un poder de discriminación del 0,13. Por consiguiente, el 45% de los descriptores analizados se ajustan al principio de relevancia; el resto tienen un valor menor.

En la tabla VIII se muestra el valor de discriminación aceptable y el porcentaje de documentos recuperables por cada uno de los descriptores.

3.3 Análisis de la exhaustividad

El análisis de la exhaustividad se ha realizado calculando la media de descriptores empleados para describir un documento en una muestra de 50 registros elegidos al azar. El resultado obtenido es el que se presenta en la tabla IX.

Tabla IX

Descriptores por documento

Doc.	*N.º descriptores*	*Doc.*	*N.º descriptores*
1	9	26	4
2	9	27	5
3	9	28	12
4	7	29	11
5	5	30	11
6	7	31	12
7	7	32	7
8	6	33	9
9	7	34	10
10	9	35	6
11	8	36	7
12	12	37	7
13	8	38	11
14	5	39	6
15	5	40	5
16	12	41	10
17	10	42	8
18	8	43	6
19	9	44	8
20	7	45	6
21	11	46	4
22	10	47	3
23	4	48	4
24	7	49	6
25	6	50	8

La media de descriptores utilizados es 7, lo que significa una exhaustividad bastante ajustada a la medida considerada como recomendable, que oscila entre 8 y 12 descriptores por documento. No obstante, hay que tener en cuenta que esas cifras dependen de la capacidad que tengan los términos de representar conceptos. En esta base de datos se observa que no existen descriptores muy extensos, ya que la tendencia es a términos simples o compuestos de dos o tres palabras. Sin embargo, aunque el promedio de descriptores se ajusta a una correcta evaluación de la exhaustivi-

dad, debemos tener en cuenta que se observa poca homogeneidad a la hora de establecer criterios al respecto, puesto que el número de descriptores empleados por documento es muy irregular. Los hay con 2 ó 3 descriptores, en cambio otros llegan a tener incluso más de 10.

4 Resultados del análisis de la estructura de los registros

Los registros de la base de datos ECOSOC se estructuran en los campos que se describen a continuación. Todos ellos, excepto el de notas, son susceptibles de recuperarse.

— Número de registro: número que identifica a cada registro.
— Autor/es: personas físicas o jurídicas que firman el documento.
— Título: del documento en castellano.
— Título en inglés: título original en inglés o versión inglesa de los títulos en otro idioma.
— Título en otro idioma: título original en idiomas distintos del español o del inglés.
— Título colectivo: título de libros de autoría colectiva, en que se publica el documento.
— Lugar de trabajo: organismo público o privado al que pertenece/n el autor/es del documento.
— Título de la revista: nombre de la publicación periódica en la que aparecen los documentos.
— Serie: nombre y número de la publicación periódica, cuando no es una revista.
— Datos fuente de la publicación: año/s en que fue editada la publicación.
— Notas: informaciones complementarias, en letra más reducida.
— Editor.
— ISSN: (International Standard Serial Number).
— ISBN: (International Standard Book Number).
— Tipo de documento: código de la tipología documental empleada.
— Modalidad documental: dentro de la tipología documental, se expresa el enfoque o perspectiva que caracteriza al documento.
— Lengua: idioma original del documento.
— Localización del documento: siglas indicativas de la institución en cuyos fondos se encuentra el documento.
— Nombre del Congreso: número de orden, fecha y lugar, cuando se trata de ponencias o comunicaciones.
— Clasificación temática: código numérico de 6 dígitos, de acuerdo a una clasificación elaborada en el propio centro productor. Los dos primeros dígitos se refieren al área temática, los dos siguientes a la disciplina y los otros dos a la subdisciplina, por lo tanto las búsquedas pueden realizarse descendiendo al nivel de subdisciplina. Un documento puede estar comprendido hasta en cuatro clasificaciones distintas; la primera de ellas se refiere al tema principal del documento.

— Descriptores: se tratan de palabras-clave, unitérminos o términos compuestos, extraídas del contenido del trabajo y que sirven de punto de acceso para su localización y recuperación.
— Identificadores: se recogen en este campo los nombres relativos a personas, instituciones, asociaciones, títulos de obras, etc.
— Topónimos: nombres de lugares relevantes en el texto.
— Resumen: se incluye en la referencia siempre y cuando haya sido elaborado por el autor de la publicación y conste en los documentos originales.

Con respecto a la cantidad de información proporcionada en cada campo cabe señalar que, si bien el principio de brevedad debe regir todo lo relativo a la presentación de información en forma electrónica, el detalle y la exhaustividad son deseables siempre que no supongan una redundancia injustificada de la información. De la misma forma, la existencia de campos repetitivos o innecesarios ralentiza el proceso de recuperación, al tiempo que dificulta la asimilación eficaz de la información.

Así, en lo que se refiere a la exhaustividad o cantidad de información de cada uno de los campos, ésta es la deseable, ya que permite a los usuarios de esta base de datos obtener un conocimiento completo del documento al que se refiere el registro, tanto en el aspecto formal como en el del contenido. Ahora bien, se quebranta el tantas veces mencionado principio de brevedad, ya que la existencia de algunos de los campos apenas aporta información relevante al usuario, pudiendo generar, incluso, cierto tipo de confusiones. Tal es el caso del campo *Modo de documento*, cuya diferencia con respecto al campo *Tipo de documento* es difícilmente perceptible para un usuario cualquiera. En cuanto a los distintos campos referidos al título del documento, parecería quizás más lógico incluir, en uno solo, el título original y, entre paréntesis, la traducción del mismo al español. Por otra parte, la denominación *Título en otro idioma* da lugar a confusiones, ya que no se sabe cuál de los dos títulos es el original. Con respecto al campo *Topónimos*, nos parece que la información que éste proporciona podría incluirse en el campo *Identificadores*.

En el caso de las Ciencias Sociales, donde suelen predominar las búsquedas temáticas, son los campos que proporcionan información de contenido los más apreciados. En este sentido, ECOSOC presenta el índice de calidad requerido, ya que sus registros cuentan, por un lado, con varios campos de valor añadido muy apreciables a la hora de una recuperación temática, tales como *Clasificación temática, Descriptores, Identificadores, Notas y Resumen*. Por otra parte, los campos descriptivos sirven, fundamentalmente, para delimitar las búsquedas, para resolver cuestiones puntuales y para realizar estudios estadísticos y bibliométricos. La exhaustividad de ECOSOC a este respecto permite a sus usuarios obtener una información complementaria de gran utilidad, al tiempo que delimitar las búsquedas con precisión. Resaltamos la existencia del campo *Lugar de trabajo*, ausente en otras bases de datos de la materia, ya que es muy importante para llevar a cabo análisis de la procedencia institucional o geográfica de la producción científica de un país o de un área científica determinada.

Concluimos que, pese a no tratarse de una base de datos de texto completo, los datos proporcionados por los registros de ECOSOC permiten a los usuarios obtener una noción completa, precisa y exhaustiva sobre los documentos primarios referenciados, especialmente desde que, en el año 1990 se incluyó el campo *Resumen*.

5 Resultados del análisis de la detección de errores

El análisis de detección de errores se ha efectuado en todos los campos, sobre una muestra de 478 documentos. Los resultados se recogen en la tabla X.

Tabla X

Detección de errores

Campos	*Ortográf/tipográf.*		*Datos mal situados (%)*	*Normalización (%)*
	Palabras	*%*		
Autor	1.997	0,2	0,4	0,6
Título	4.140	0,04	0	0
Lugar de trabajo	2.432	0,8	1	6,4
Título Revista	1.918	0,1	0	0
Descriptores	2.915	0,2	0	4,4
Identificadores	2.100	0,1	0	0
Topónimos	857	0	0	0,4
Resumen	54.204	0,1	—	—

Cabe señalar, como un claro indicador de calidad de esta base de datos, la inexistencia de registros duplicados. Existe un total de 360 registros en toda la base de datos que no proporcionan información sobre los autores de los documentos, ni tampoco se constatan los datos relativos al lugar de trabajo en 7.046 registros. En 5.503 y 4.842 casos, los campos de *Identificadores* y *Topónimos* carecen de información, respectivamente.

Con respecto al análisis de la detección de errores ortográficos y tipográficos, cabe señalar que, en general, el índice se encuentra por debajo de lo que se considera admisible, pues a excepción del campo *Lugar de trabajo*, en el resto no llega al 0,3 %. En uno de los campos, el de *Topónimos*, no se ha detectado ningún error de este tipo.

Señalamos que la base de datos cuenta con dos índices alfabéticos asociados al campo *Autor*, que permiten recuperar estos datos por palabras y por frases, respectivamente. Esto hace que el índice de errores repercuta en menor medida a la hora de la recuperación.

En cuanto al campo de *Título*, nos parece importante comentar que, en la muestra analizada, no se ha encontrado ningún registro con el campo *Título en inglés*, pero, sin embargo, sí existen 130 documentos con el campo *Título en otro idioma*, de los cuales, 10 son en inglés y 1 en castellano. Esto corrobora la idea ya expresada anteriormente de la redundancia de campos relativos al título del documento.

El índice de errores relacionados con datos mal situados es insignificante, ya que el mayor, que se corresponde con el campo *Lugar de trabajo*, es del 1 %.

A la vista de los resultados, es el campo *Lugar de trabajo* el que más errores de todo tipo presenta. Se trata, a nuestro juicio, de un campo importante a la hora de delimitar las búsquedas, así como de conocer la producción científica de las distintas instituciones y áreas geográficas. Nos parece, sin embargo, que el hecho de que el nombre de los centros aparezca mediante abreviaturas dificulta el proceso de búsque-

da, pues aunque existen unas tablas en el menú de ayuda donde se desarrollan y se explica el modo de buscar correctamente el lugar, sería preferible dar la opción al usuario de buscar en este campo por los términos completos.

En este campo se ha observado, también, un porcentaje considerable de errores (6,4%) debidos a una falta de normalización. No existe coherencia a la hora de presentar un mismo término, bien de forma abreviada, bien desarrollada, o los acrónimos y la forma completa con la que se corresponden. El lugar geográfico de la institución aparece referido, bien a la ciudad, bien a la provincia o Comunidad Autónoma, por ejemplo. Los signos de puntuación tampoco siguen una política de consistencia, ya que comas y puntos, por ejemplo, se alternan sin atenerse a un criterio específico.

En cuanto al campo *Datos fuente*, sólo un 0,2 % de los registros analizados presentaba un error de puntuación. Error que no supone, sin embargo, silencio en la recuperación, ya que este campo sólo puede ser recuperado por año de publicación.

Todos los datos del campo *Editor* aparecen constatados correctamente (30 documentos). De los 68 registros que hacen referencia a un congreso, existen 2 con errores que no permiten recuperar el documento por este campo y 4 que no mantienen el orden correcto. En cuanto al campo donde se señala la *Clasificación temática*, cabe decir que, pese a no tratarse de un campo obligatorio, en la muestra analizada no existe ningún registro con este campo vacío. Todos los documentos se encuentran clasificados siguiendo la codificación establecida.

De los errores detectados en el campo *Descriptores*, un 0,2 % se debe a errores ortográficos y tipográficos, y un 4,4 % a los derivados de un control en su normalización. En general, los descriptores respetan las normas relativas a su estructura y construcción.

El proceso de detección de errores en el campo de *Identificadores*, cuyos términos siguen la misma estructura que la de los descriptores, presenta resultados muy satisfactorios, ya que sólo se ha detectado un 0,1 % de errores ortográficos y ninguno de otro tipo.

En cuanto al campo *Topónimos*, que sigue las recomendaciones de la UNESCO para los nombres geográficos, cabe señalar que se trata de un campo cuyo control es complejo, ya que, a menudo, las denominaciones de los lugares geográficos sufren variaciones. Esta base de datos cuenta con un tesauro específico de topónimos, desde 1993, que garantiza la normalización de los términos empleados. Esto se refleja en los resultados, ya que, el 0,4 % de errores de normalización detectados en este campo se corresponden con referencias anteriores a 1993.

Por último, con respecto al campo *Resumen*, es importante señalar que éste sólo se incluye cuando el documento cuenta con un resumen de autor, razón por la cual no hemos entrado a analizar la calidad de los resúmenes, ni tampoco hemos estudiado si éstos se rigen por una política consistente en cuanto a su tipología y forma. El índice de errores ortográficos se encuentra por debajo del límite considerado aceptable.

6 Conclusiones

La primera conclusión que se extrae del análisis efectuado es que los descriptores utilizados en la indización de los documentos de la base de datos ECOSOC abar-

can todo el espectro de la ciencia política en general, así como lo relativo a nuestro entorno geográfico y cultural en particular.

Asimismo, la calidad de la indización de la base de datos objeto de estudio se ajusta a los parámetros marcados en lo que respecta a la consistencia, la relevancia y la exhaustividad de los términos empleados, lo que garantiza una correcta representación del contenido de los documentos, y, por consiguiente, la posibilidad de obtener resultados precisos y/o exhaustivos, según el objeto de la demanda, en las búsquedas temáticas.

El número de campos en que se estructuran los registros de la base de datos analizada asegura la exhaustividad informativa de sus registros, tanto formal como de contenido. Sin embargo, existen campos innecesarios y que apenas aportan información. Es digno de destacar el campo *Lugar de trabajo* de los autores, que nos permite conocer el origen geográfico e institucional de los mismos y que, sin embargo, no es habitual en otras bases de datos sobre la materia.

En lo que respecta al índice de errores, éste se encuentra por debajo de lo que se considera admisible. Por otra parte, estos errores no se mantienen en el tiempo, por lo que se constata la existencia de una depuración periódica de los mismos, lo que es indicativo de una buena política de mantenimiento de la base de datos.

Así pues, y según los resultados obtenidos en el análisis de evaluación llevado a cabo, la base de datos ECOSOC tiene un nivel de calidad superior a los parámetros considerados como medios, en lo que respecta al control terminológico, la estructura de sus registros y el índice de errores.

7 Bibliografía

1. EXTREMEÑO, A.; MOSCOSO, P. El control de la calidad en bases de datos de Ciencias Sociales. *Boletín de la ANABAD* 1998, vol. XVLIII, n.º 1, pp. 231-253.
2. HARRY, V.; OPPENHEIM, C. Evaluations of electronic databases. *Online & CDROM Review* 1993, vol. 17, n.º 4, pp. 211-222, y n.º 6 p.p 339-351.
3. JALKANEN, T. et al. A metric evaluation system for database quality. *Proceedings of the 3rd International Society for Knowledge Organization (ISKO) Conference,* 1994, junio, 20-24, Copenhaguen, Denmark.
4. MAY, N. A. A methodology for the measurement of quality of electronic databases. *Proceedings of the 3rd International Society for Knowledge Organization (ISKO) Conference,* 1994, junio, 20-24, Copenhaguen, Denmark.
5. MEDAWARD, K. Database quality: a literature review of the past and a plan for the future. *Program,* 1995, vol. 29, n.º 3, pp. 257-272.
6. O'NEILL, E. T.; VIZINE-GOETZ, D. Quality control in Online databases. *Annual Review of Information Science and Technology,* 1988, vol. 23, p. 125-156.
7. RODRIGUEZ YUNTA, L. Evaluación e indicadores de calidad en bases de datos. *Revista Española de Documentación Científica,* 1998, vol. 21, n.º 1, pp. 9-23.
8. *Campo de estudio consistente en la aplicación de la ciencia de la computación a la estructura y significado del lenguaje,* dirigido sobre todo por Noam Chomsky.
9. ZIPF, G. K. *Selected studies of the principle of relative frecuency in language,* 1992, Cambridge, Mass.: Harvard University Press.
10. WULFF BARREIRO, E. Calidad de la indización en bases de datos de tamaños cualesquiera. *Boletín de la Asociación Andaluza de Bibliotecarios,* 1993, vol. 9, n.º 31, pp. 41-45.

11. WHITE, H. D.; GRIFITH, B. C. Quality of indexing in online databases. *Information Processing and Management,* 1987, vol. 23 , pp. 211-214.
12. JU, C. M.; ACHIRUQUE, I. Quality of indexing library and information science database. *Online Review,* 1989, vol. 13, n.º 1, pp. 11-35.
13. ZAMORA, A. Automatic detection and correction of spelling errors in a large data base. *Journal of the American Society for Information Science,* 1980, vol. 30, n.º 1, pp. 51-57.
14. POLLOCK, J.J.; ZAMORA, A. Collection and characterization of spelling errors in scientific and scholarly texts. *Journal of the American Society for Information Science,* 1983, vol. 34, n.º 1, pp. 51-58.
15. O'NEILL, E. T. VIZINE,-GOETZ, D. The impact of spelling errors on databases and indexes. *Proceedings National Online Meeting,* 1987, May, 9-11, New York, pp. 313-320.
16. DAMEREAU definió 4 tipos de errores tipográficos en la década de los 60, que han permanecido hasta hoy.
17. SPINAK, E. Errores ortográficos en el ingreso en bases de datos. *Revista Española de Documentación Científica,* 1995, vol. 18, n.º 3, pp. 307-319.